SLNet: A Spectrogram Learning Neural Network for Deep Wireless Sensing

Zheng Yang and Yi Zhang, *Tsinghua University;* Kun Qian, *University of California San Diego;* Chenshu Wu, *The University of Hong Kong*

https://www.usenix.org/conference/nsdi23/presentation/yang-zheng

This paper is included in the
Proceedings of the 20th USENIX Symposium on
Networked Systems Design and Implementation.

April 17–19, 2023 • Boston, MA, USA

978-1-939133-33-5

SLNet: A Spectrogram Learning Neural Network for Deep Wireless Sensing

Zheng Yang[1], Yi Zhang[1], Kun Qian[2], Chenshu Wu[3*]

[1] Tsinghua University, [2] University of California San Diego, [3] The University of Hong Kong
{hmilyyz,zhangyithss,qiank10}@gmail.com,chenshu@cs.hku.hk

ABSTRACT

Advances in wireless technologies have transformed wireless networks from a pure communication medium to a pervasive sensing platform, enabling many sensorless and contactless applications. After years of effort, wireless sensing approaches centering around conventional signal processing are approaching their limits, and meanwhile, deep learning-based methods become increasingly popular and have seen remarkable progress. In this paper, we explore an unseen opportunity to push the limit of wireless sensing by jointly employing learning-based spectrogram generation and spectrogram learning. To this end, we present SLNet, a new deep wireless sensing architecture with spectrogram analysis and deep learning co-design. SLNet employs neural networks to generate super-resolution spectrogram, which overcomes the limitation of the time-frequency uncertainty. It then utilizes a novel polarized convolutional network that modulates the phase of the spectrograms for learning both local and global features. Experiments with four applications, *i.e.*, gesture recognition, human identification, fall detection, and breathing estimation, show that SLNet achieves the highest accuracy with the smallest model and lowest computation among the state-of-the-art models. We believe the techniques in SLNet can be widely applied to fields beyond WiFi sensing.

1 INTRODUCTION

We are entering the era of Artificial Intelligence of Things (AIoT) where trillions of devices are pervasively connected and, more importantly, equipped with advanced sensing intelligence. They can sense the physical space and gain awareness of contexts such as locations, activities, motion, vital signs, etc. With advances in wireless sensing, all these could be achieved using pervasive wireless infrastructure, without dedicated sensors, wearables, or cameras. As promising as it is, existing wireless sensing is approaching its limits using conventional signal processing methods and faces performance bottlenecks in distinguishing task-relevant features from entangled irrelevant features in signals.

With its remarkable success in numerous fields, deep learning has become increasingly popular, and also seemingly effective, for wireless sensing, promising the next breakthrough for practical wireless sensing systems for AIoT. Most of the prior works perform conventional signal processing (*e.g.*, frequency transformation) in tandem with deep neural networks, such as convolutional neural networks, which are mainly designed for visual data like images and videos. RF data, most commonly Channel State Information (CSI) data, however, fundamentally differs from visual data in multiple unique aspects: 1) *Non-visual*[1]: RF data contains physical and geometric connotations in time, space, and frequency domains that are not visually intelligible; *2) Complex*: RF data is complex-valued with both amplitude and phase information; *3) High-dimensional*: While visual data are mostly 2D or 3D, RF data comes with multiple dimensions of time, subcarriers, antennas, and/or transceivers. In addition, it is generally more difficult to build a large RF dataset for training than in the computer vision field, both because that RF data collection is cumbersome as it depends on many environmental factors and that RF data cannot be labeled offline since they are not visually intelligible to human eyes. There exists a gap between prior neural networks and the distinct RF data, rendering existing deep wireless sensing systems suboptimal in performance yet over-complicated in model complexity. While there are also other non-visual, complex, and/or high-dimensional data like speech [28, 50], the unique characteristics of RF sensing call for a separate design to push the limit of deep wireless sensing.

We present SLNet, a novel neural network architecture with a *spectrogram analysis-deep learning co-design* for RF data. Rather than performing spectrogram analysis separately from deep learning, SLNet couples them tightly based on an in-depth understanding of their respective limitations in processing RF data. By doing so, SLNet significantly boosts the effectiveness and efficiency of deep wireless sensing. It consists of three major modules:

Learning-Assisted Spectrogram Enhancement: Many

[*] Zheng Yang is the corresponding author and Yi Zhang is the first student author.

[1] RF data can certainly be visualized in many ways. However, we argue that RF data itself is not visually intelligible like images to humans.

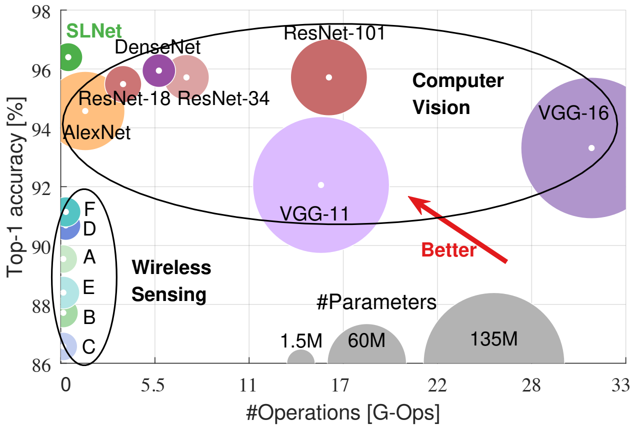


Figure 1: A comparison between SLNet and the state-of-the-art neural networks for WiFi-based gesture recognition task. CV models are accurate but bigger, while existing networks for wireless sensing are relatively small but less accurate. SLNet achieves the highest performance on wireless sensing tasks while reducing computing and memory consumption for practical applications. The radii of the circles represent the number of model parameters. (References: A: [90], B: [22], C: [87], D: [84], E: [30], F: [46])

wireless sensing approaches, either model-driven or data-driven, employ the Fast Fourier Transform (FFT) on a time series of RF data to obtain time-frequency spectrograms of human activities. FFT suffers from errors due to an effect known as leakage, when the block of data is not periodic (the most common case in practice), which results in a smeared spectrum of the original signal and further leads to misleading data representation for learning-based sensing: First, the side lobes "pollute" the spectrograms as they are not from actual human motions but simply the results of spectral leakage. Second, human activities typically contain multiple frequency responses that may be severely affected by the leakage, leading to a "blurred" spectrogram with mixed lobes. Classical approaches reduce leakage by windowing, which cannot eliminate leakage entirely. In effect, they only change the shape of the leakage with different windowing functions to achieve a trade-off between temporal and frequency resolutions. Differently, utilizing learning-based methods promises to push the boundaries beyond classical limitations and, in turn, provide high-fidelity spectrograms for further learning tasks. SLNet introduces a spectrogram enhancement network (§3.1) to learn the best function to minimize or nearly eliminate the leakage, thereby outputting an unparalleled spectrum with high accuracy.

Multi-Resolution Spectrogram Fusion: The frequency resolution of FFT depends on its window size, *i.e.*, the length of the input data block. Using a larger window promises higher resolution, but only generates a more accurate spectrum when the underlying frequency is quasi-static within the window. In contrast, applying a shorter window improves the responsiveness to fast-changing frequencies, but immediately loses high resolution. Therefore, instead of balancing between conflicting goals of resolution and responsiveness by finding a fixed window length, SLNet employs multiple windows jointly and generates a hologram of multi-resolution spectrograms, which then serves as multi-channel inputs that a neural network can adaptively learn from (§3.2).

Polarized Convolutional Network: The hologram is like an image by format, with each spectrogram serving as a "color" channel. Thus it is straightforward to employ convolutional neural networks (CNN) to extract underlying features from it. Invented for visual data, CNN mainly learns local features irrespective of global locations of objects in an image, allowing images to be shifted. Unfortunately, the locality property makes CNN inappropriate for spectrogram learning, as the global locations, *i.e.*, frequencies, are correlated with the physical properties of a person's activities, which is not shift-invariant. To preserve global discrimination, we propose a Polarized Convolutional Network (PCN, §3.3). First, we polarize the spectrograms via specially modulated phase information, making them locally unaltered while globally differentiated. Then we design a special convolutional operator to extract features from the polarized and thus complex-valued spectrogram. Compared to CNN, PCN preserves the local features and the global discrimination simultaneously and thus boosts the learning performance. Based on this, we further adopt a compression network for feature deduction and build a task-adaptive network that can be flexibly customized for different sensing tasks.

We implement SLNet on commodity off-the-shelf (COTS) WiFi devices and evaluate its performance for four human-centered sensing applications, i.e., gesture recognition, gait-based person identification, fall detection, and breathing rate estimation. Extensive experiments are conducted in four typical indoor environments including a classroom, a hall, an apartment, and an office. Our results show that SLNet achieves 96.6% accuracy on gesture recognition, 98.9% accuracy on gait identification, 99.8%/97.2% precision/recall for fall detection, and an average error of 2.4 BPM for multi-person breath estimation. Experimental comparisons with over 10 state-of-the-art deep learning models demonstrate that SLNet achieves the highest accuracy with the fewest model parameters and computation operations, as illustrated in Fig. 1, making it more practical and preferable for edge devices (e.g., home routers).

Contributions: SLNet presents a spectrogram analysis-deep learning co-design network distinctively customized for deep wireless sensing on the time series of high-dimensional, complex-valued RF data. We envision that SLNet inspires tailored deep-learning architectures that are generalizable to multiple tasks and environments of wireless sensing. Further, we believe the techniques introduced in SLNet, including SEN and PCN, are applicable to many fields involving time–frequency signal analysis and spectrogram learning. SLNet is open-sourced here [41].

2 PRIMER

2.1 Preprocessing of RF Data

CSI reflects the channel through which wireless signals propagate. When a person performs an activity, his or her impact on the channel is encoded in CSI, and the activity can thus be inferred from the CSI. Suppose that the person creates L propagation paths between the transmitter and the receiver, the measured CSI is [36]:

$$H(t) = H_s + \sum_{l=1}^{L} \alpha_l(t) e^{j2\pi \int_{-\infty}^{t} f_{D_l}(u)\mathrm{d}u} + n(t), \tag{1}$$

where α_l, f_{D_l} are the complex attenuation and Doppler frequency shift of the signal of the l-th path, H_s is the static part of the channel between the transmitter and the receiver, and n is the additive Gaussian noise.

To recognize the activity of the person, the raw temporal CSI signals are usually transformed into spectrograms via Short-Time Fourier Transform (STFT):

$$S(f,t) = \mathrm{STFT}[H(t)] = \mathrm{FFT}[\varpi] * \sum_{l=1}^{L} \alpha_l(t)\delta(f - f_{D_l}) + N(f,t), \tag{2}$$

where ϖ represents the windowing function in the time domain, $*$ the convolution operation, and δ the impulse function. N is the frequency response of the Gaussian noise. In Eq. 2, $\sum_{l=1}^{L} \alpha_l(t)\delta(f - f_{D_l})$ reflects the activity of the person. However, it is distorted by the windowing effect of $\mathrm{FFT}[\varpi]$ and the noise N. As a result, the data fidelity of a spectrogram in representing a person's activity is impaired. To remove these negative effects in the spectrogram and make the frequency components of interest prominent, a spectrogram enhancement network is developed in SLNet.

Hereafter, we refer to the 2-D output of STFT as *spectrogram* and the 1-D output of FFT as *spectrum*.

2.2 Complex-Valued Neural Network

The neural network acts as one of the most powerful tools in solving various cognitive problems, such as image classification [24], speech enhancement [16], and text translation [31]. A neural network consists of layers of neurons that generate responses according to their inputs. As shown in Fig. 2a, a neuron calculates the sum of the input $\mathbf{x}$ weighted with parameters $\mathbf{w}$ and the bias and applies a nonlinear activation function σ (e.g., tanh) to generate the output x', i.e., $x' = \sigma(\sum \mathbf{wx} + b)$. Recently, neural networks have been used for applications of wireless sensing, such as activity classification [22], gait identification [91], and gesture recognition [90]. However, the phase information of CSI is less exploited or even abandoned in the existing approaches. According to Eq. 1, the CSI phase also encodes important information related to the person's activity, which, once exploited, can benefit the recognition process. Thus, instead of using a real-valued neural network, SLNet devises a complex-valued neural network, whose neuron processes complex values and fits the complex-valued spectrogram of CSI. As shown in Fig. 2b, a complex-valued neuron consists of two real-valued neurons, which process the real and imaginary parts of the input, respectively. Specifically, suppose the input is $\mathbf{z} = \mathbf{x} + i\mathbf{y}$, the weight is $\mathbf{w} = \mathbf{w}_1 + i\mathbf{w}_2$, and the bias is $\mathbf{b} = \mathbf{b}_1 + i\mathbf{b}_2$, then the output of the complex neuron is $z' = x' + iy'$, where $x' = \sigma(\mathrm{Re}(\sum \mathbf{wz} + b))$ and $y' = \sigma(\mathrm{Im}(\sum \mathbf{wz} + b))$.

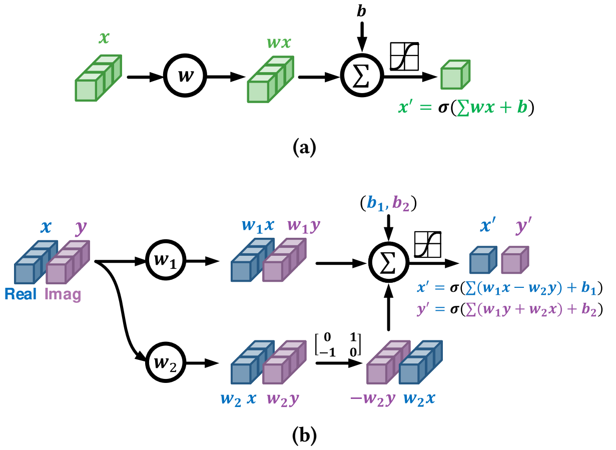

b
x
w
wx
$\sum$
$x' = \sigma(\sum wx + b)$
(a)
(b_1, b_2)
x y
Real Imag
w_1
w_1x w_1y
$\sum$
x' y'
$x' = \sigma(\sum(w_1x - w_2y) + b_1)$
$y' = \sigma(\sum(w_1y + w_2x) + b_2)$
w_2
$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$
$w_2\,x$ w_2y
$-w_2y$ w_2x
(b)

Figure 2: Comparison between (a) real-valued and (b) complex-valued neurons.

3 SLNET ARCHITECTURE

SLNet is designed as a customized spectrogram learning framework assisted with deep learning for RF data applications. It identifies the limitations of the standard signal processing methods for wireless signals and employs specifically designed deep learning modules to overcome them. Fig. 3 shows the workflow of SLNet, which consists of four parts. First, the *spectrogram enhancement network (SEN)* takes as input a spectrogram transformed from wireless signals via STFT, removes the spectral leakage in the spectrogram, and recovers the underlying actual frequency components. Second, the *Fusion* module combines SEN-enhanced spectrograms with various temporal and frequency resolutions to form a hologram of spectrograms. To coherently combine all spectrograms, SLNet modulates them with linear phases,

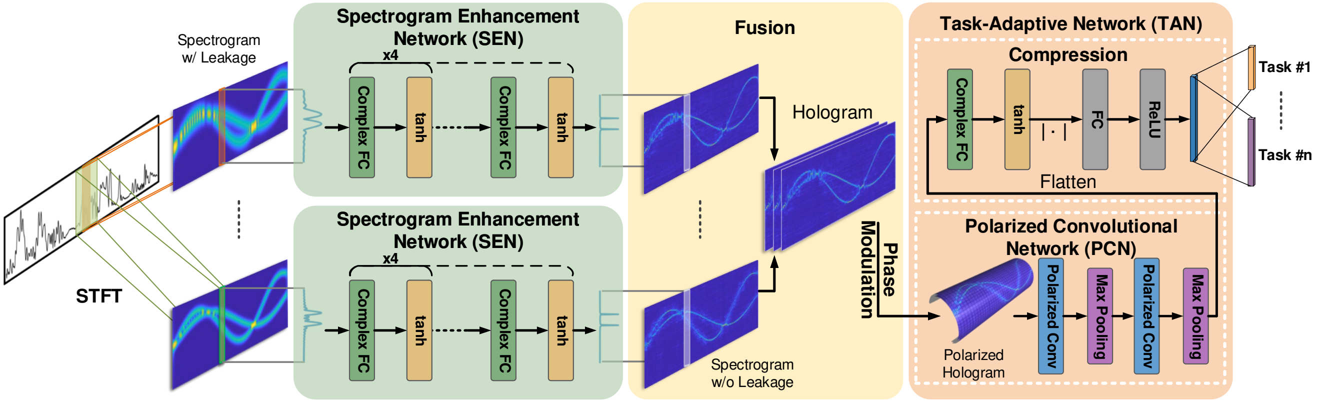
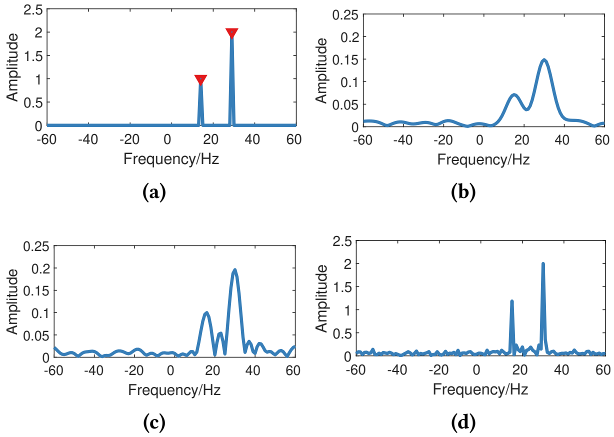


Figure 3: Overview of SLNet. The temporal CSI signal is transformed into spectrograms via a bank of STFT operators with different temporal and frequency resolutions. Each spectrogram is fed into the SEN to remove spectral leakage. Then, a hologram of spectrograms is generated by stacking all enhanced spectrograms and modulating them with linear phases. Next, the hologram is processed with the PCN to generate feature maps, and the compression networks to generate abstract features for specific learning tasks.



(a)

(b)

(c)

(d)

Figure 4: Illustration of spectral leakage. (a) The ideal frequency spectrum with discrete frequency components. (b) The measured frequency spectrum obtained via FFT. (c) The frequency components recovered via least mean square regression. (d) The frequency components recovered from SLNet's SEN.

and the result is termed a polarized hologram. Third, the *polarized convolutional network (PCN)* module processes the hologram to obtain feature maps with general representativity. Finally, we adopt a *compression* network for feature deduction and build a *task-adaptive network (TAN)*, which can be flexibly adapted for different sensing tasks.

3.1 Spectrogram Enhancement Network

Standard signal processing transforms temporal CSI signals to a time-frequency spectrogram via STFT. A certain STFT operator truncates the time series of signals using a sliding window with a fixed length. However, the truncation results in the windowing effect, which convolves the ideal frequency spectrum with a sinc function and creates spectral leakage in the frequency domain. Some classical windowing functions (like Hamming or Gaussian window [13]) can be multiplied with the truncated signal to mitigate the spectral leakage, but none of them completely removes the leakage. Fig. 4a illustrates an ideal frequency spectrum with two frequency components at 15 and 30 Hz. As shown in Fig. 4b, the estimated frequency spectrum obtained via STFT and Gaussian window has significant spectral leakage and additive Gaussian noise. Formally, suppose the ideal and estimated frequency spectrums are $\mathbf{s}$ and $\hat{\mathbf{s}}$ respectively. We have:

$$\hat{\mathbf{s}} = \mathbf{A}\mathbf{s} + \mathbf{n}, \tag{3}$$

where $\mathbf{n}$ represents the additive Gaussian noise vector and $\mathbf{A}$ is the convolution matrix of the windowing function in the frequency domain. Based on Eq. 2, the i-th column of $\mathbf{A}$ is:

$$\mathbf{A}_{(:,\mathrm{i})} = \mathrm{FFT}(\varpi) * \delta(i). \tag{4}$$

The spectral leakage significantly distorts the frequency spectrum, producing unwanted side lobes and inaccurate frequencies and amplitudes. For example, when two frequency components are close to each other, their spectral leakage interacts, and the weaker component becomes less prominent, as shown in Fig. 4b. Such spectral leakage is caused by the truncation of the STFT operation and is not relevant to the sensing targets, and is essential to be removed before applying the frequency spectrum to sensing tasks.

Given the relation between the ideal and estimated spectrum as in Eq. 3, it is straightforward to recover the ideal spectrum via the least mean square (LMS) regression:

$$\mathbf{s} = \mathrm{argmin}_{\mathbf{s}} ||\hat{\mathbf{s}} - \mathbf{A}\mathbf{s}||_2. \tag{5}$$

However, the LMS regression tends to output suboptimal

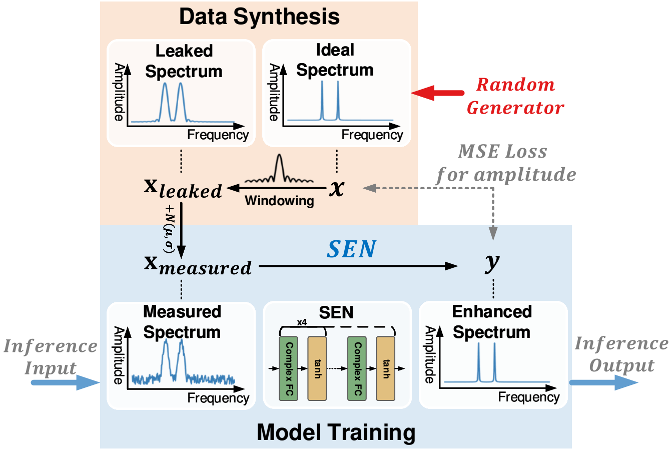


Figure 5: Data synthesis and training process of SEN.

solutions with inaccurate side peaks, as shown in Fig. 4c, due to the existence of Gaussian noises. In contrast, the ideal frequency spectrum of CSI signals tends to be sparse, due to the sparsity of moving objects exposed in the wireless channel [90]. By adding the l_0-norm regularization, the recovered spectrum is closer to the ideal spectrum. However, the l_0-norm regularization makes the computational complexity of the problem exponential to the dimension of the frequency spectrum $\mathbf{s}$ [6], which is thus intractable. Besides, the sparse Fourier transform that aims to solve this issue also suffers from high complexity [14].

To efficiently recover the ideal spectrum, we resort to the neural network. Compared with the optimization methods, the learning-based method offloads the computation efforts to the training phase and enables efficient linear computation in the testing phase. In addition, the neural network can regress arbitrary functions and is resistant to noises thanks to its continuity in the hidden space. To achieve it, we develop the dedicated network SEN. As shown in Fig. 5, the SEN takes as input the measured complex-valued frequency spectrum and outputs the recovered spectrum. The SEN consists of four complex-valued fully connected layers with the hyperbolic tangent activation function.

To train the SEN, the training dataset has to embrace the complexity of the frequency spectrum, which is extremely high due to the wide amplitude and phase range of wireless signals and random channel noises. For example, the frequency of interest for human-centered sensing is within $[-60, 60]$ Hz and usually, at most five frequency components can be observed for major reflections from the human body [55]. That is, after normalizing the signal amplitude, a spectrogram can consist of 1 to 5 frequency components, whose amplitudes, phases, and frequencies are in $[0, 1]$, $[0, 2\pi]$, $[-60, 60]$ Hz, respectively. As collecting data with labeled ground truth from real scenarios is challenging, we instead synthesize the training data, which turns out to be sufficiently effective. As shown in the upper part of Fig. 5, we randomly generate ideal spectrums with 1 to 5 frequency components, whose amplitudes, phases, and frequencies are uniformly drawn from their ranges of interest.

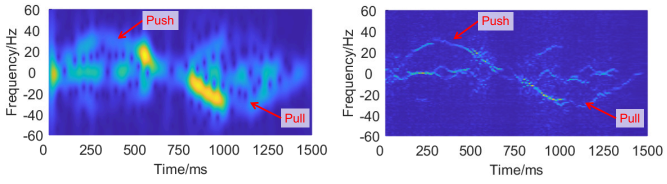
Then, the ideal spectrum is converted to the leaked spectrum following the process in Eq. 3 to simulate the windowing effect and random complex noises. The amplitude of the noise follows a Gaussian distribution, and its phase follows a uniform distribution in $[0, 2\pi]$. The SEN takes the leaked spectrum as input and outputs the enhanced spectrum close to the ideal one. Thus, we minimize the L_2 loss $L_{\text{SEN}} = ||\text{SEN}(\hat{\mathbf{s}}) - \mathbf{s}||_2$ during training. During inference, the spectrums measured from real-world scenarios are normalized to $[0, 1]$ and fed into the SEN to obtain the enhanced spectrum for further processing. Fig. 6 shows an example of the spectrogram when a person performs a gesture. As is shown, the frequency components caused by the pushing and pulling gesture are clearly recovered by the SEN.

3.2 Multi-Resolution Spectrogram Fusion

The SEN refines the frequency spectrum of the CSI signals by recovering its underlying frequency components. Thus, it assumes that the frequency components remain quasi-static during the sliding window where the frequency spectrum is generated. However, the Doppler frequency shifts induced by human activities keep changing, which may violate this assumption. For example, in the fall detection scenario, the speed of the human body changes between 0 m/s and 5 m/s, creating significant variations in signal frequencies. To illustrate its impact on SEN, we use an example of two components with changing frequencies, as shown in Fig. 7a. The measured spectrogram with a sliding window of 251 ms is shown in Fig. 7b and the refined spectrogram from SEN in Fig. 7c. While SEN correctly distinguishes the two components in the first half of the spectrogram, it fails to recover them clearly in the second half, due to the rapidly changing frequencies of the components.

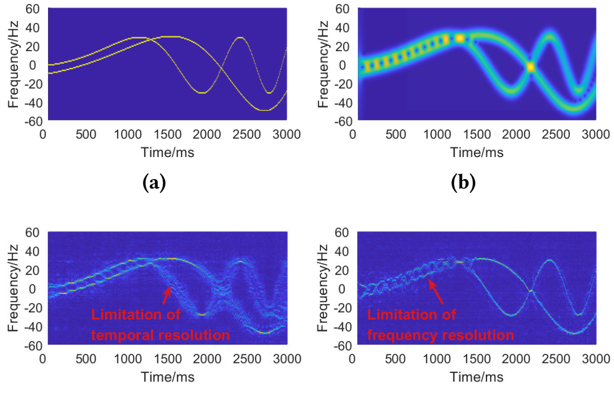
One straightforward solution is to use a shorter sliding window, during which the rapidly changing frequencies can be approximately viewed as quasi-static. However, using a shorter sliding window reduces the frequency resolution, which limits the ability of the SEN to remove the spectral leakage of the two close frequency components. Fig. 7d shows the output of the SEN using a sliding window of 125 ms. With a shorter sliding window, the SEN recovers the second half of the spectrogram with fast-changing frequencies. However, it does not separate well the close frequency components in the first half of the spectrogram, due to the coarse frequency resolution of the short sliding window.

To overcome the limitation of the temporal and frequency



(a) (b)

Figure 6: Illustration of the spectrogram of a pushing and pulling gesture. (a) The measured spectrogram and (b) the enhanced spectrogram from SEN.



(a) (b) (c) (d)

Figure 7: Illustrations of SEN-enhanced spectrograms. (a) The ideal spectrogram with two frequency components. (b) The measured spectrogram from STFT with a sliding window size of 251 ms. (c) The enhanced spectrogram with a window size of 251ms. (d) The enhanced spectrogram with a window size of 125 ms.

resolutions of the spectrogram, instead of using a fixed sliding window, SLNet employs a bank of sliding windows with different lengths (similar to [73]). For each sliding window, the corresponding spectrogram is processed via the SEN that is pre-trained with the synthesis spectrograms with the same window length. All SEN-enhanced spectrograms are then concatenated as multiple channels to form a hologram of spectrograms. As each spectrogram encodes useful information for a certain range of temporal and frequency resolutions, we further resort to the neural network to adaptively combine all the spectrograms.

3.3 Task-Adaptive Network

SLNet employs the Task-Adaptive Network to further adapt the hologram of enhanced spectrograms to various sensing tasks, such as gesture recognition, gait identification, and fall detection. As shown in Fig. 3, the TAN consists of two modules, the PCN module that captures high-level feature maps of the hologram and the compression module that reduces feature dimension for specific tasks.

Polarized Convolutional Network. A hologram can be treated as an image where each spectrogram spanning in 2-D time and frequency dimension is as one of its “color” channels, and, by doing so, a CNN [15, 24] can be applied to extract the underlying features of the hologram. However, the solution is not optimal, as explained below.

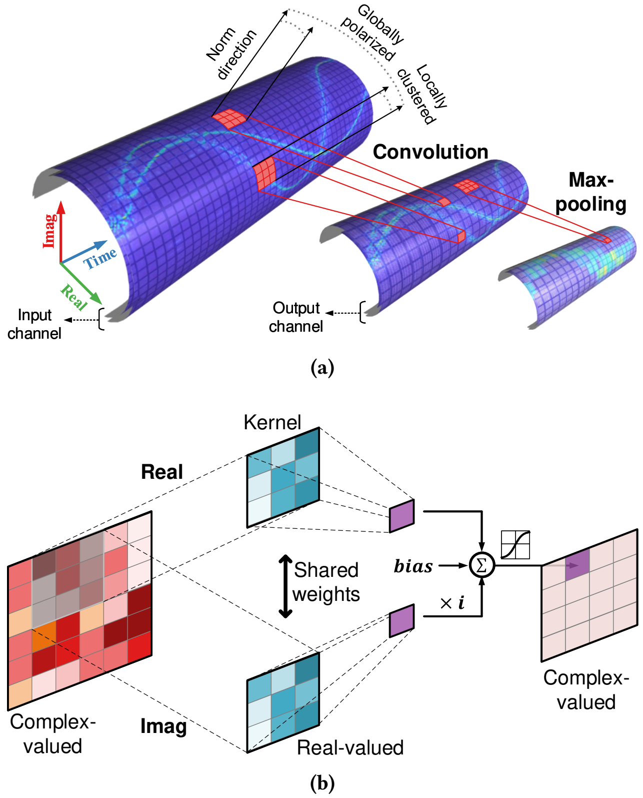
Each neuron in CNN only takes a local field of the input to generate the output. All the neurons in each layer share the same weights to ensure that the local features are preserved irrespective of their global locations. As a result, CNN is particularly tailored for visual data since it focuses on local dependencies and is invariant to global shifts of objects in images. This shift-invariant property makes CNN inappropriate for spectrogram processing, as the global locations, i.e., frequencies, of the frequency components are correlated with the physical properties of the person’s activities. In another word, a shift along the frequency dimension means a change in the moving status of the person, which is highly possible due to a different activity. Besides, the local patterns of the spectrogram capture the instant motion status of the target, which is still needed for sensing tasks. Hence, it is necessary to develop a new model that simultaneously preserves local dependency and global discrimination.

We propose to modulate the spectrograms with phase information, which is discarded in existing wireless sensing models, to explicitly encode global locations in the spectrogram while retaining its local correlations. Specifically, it is expected that the adjacent frequency components have similar phases while the distant ones have discriminative phases. Thus, we modulate the spectrogram with phases that vary linearly along the frequency dimension, i.e., the modulated phase of the i-th frequency bin is:

$$\phi_i = i\frac{\phi_h - \phi_l}{M} + \phi_l, \tag{6}$$

where ϕ_h and ϕ_l are the phases modulated to the lowest and highest frequency bins, and M is the number of total frequency bins. As a result, global discrimination is introduced along the frequency dimension while the shift-invariant property is preserved along the time dimension. Note that we apply the proposed polarized phase modulation, rather than incorporating the original phase information because the raw phase contains significant errors due to carrier frequency offsets and timing offsets, *etc.* [37, 57].

The frequency components modulated with phases can be viewed as polarized in a 2-D complex plane. To process the polarized spectrograms, we propose the PCN network. PCN consists of two pairs of convolutional layers and max-pooling layers, which are responsible for higher-level feature extraction and dimension reduction, respectively. As shown in Fig. 8a, the polarized hologram is applied with a convolutional layer to extract local features followed by



(a)

(b)

Figure 8: The (a) structure and (b) convolutional operations of the Polarized Convolutional Network.

a max-pooling layer to reduce the feature dimension. The elements within each kernel are locally clustered, while the elements within different kernels are globally polarized to have better global discrimination benefiting from the linear phase introduced. In practice, several convolutional layers can be cascaded to obtain higher-level features of the input spectrograms. The input channel represents the number of fused spectrograms in the hologram or the number of output channels in the previous convolutional layer. For each convolutional layer, Fig. 8b illustrates the convolutional operations in complex domain. The real-valued kernels are convolved with the real and imaginary parts of the spectrograms separately and combined with a bias to get the final complex-valued output. The max-pooling layer downsamples the features with the maximum amplitude and outputs the complex-valued features.

Feature map compression. SLNet further adopts a compression network to reduce the high dimensions of the feature maps generated by the PCN and obtains a condensed representation of features for specific tasks. The compression network consists of a complex-valued fully connected (FC) layer with the *tanh* activation function and a real-valued FC layer with the *ReLU* activation function. To connect two FC layers, SLNet calculates the absolute value of the output from the complex-valued FC layer and inputs it to the real-valued FC layer. The output features can be further fed into additional FC layers customized for different tasks. For example, an FC layer with N output units followed by a softmax activation function can be used for a gesture classification task with N gestures, while an FC layer with 1 output unit followed by a sigmoid layer can be used to predict the likelihood of human fall for the fall detection task.

We employ the pre-trained SENs to obtain the enhanced spectrograms and feed them into the TAN for training. The L_2 loss between the output of the TAN and the ground-truth label is minimized, and the RMSprop optimizer is used. During inference, SLNet takes as input the measured CSI spectrograms and outputs the prediction result.

4 IMPLEMENTATION & EXPERIMENTS

4.1 Implementation

Hardware. SLNet collects CSI measurements from commodity Intel 5300 WiFi Network Interface Cards (NICs) equipped on off-the-shelf mini-computers. The three antennas of the NIC are separated apart by half of the signal wavelength, i.e., 2.85 cm. The operating system of the mini-computer is Ubuntu 10.04 with Linux CSI Tool [12] installed to log CSI readings. The NIC is set to operate on channel 165 with a center frequency of 5.825 GHz. We set all the receivers to work on monitoring mode and inject the transmitter to broadcast at a rate of 1,000 packets per second. All the devices are connected to a router and remotely controlled. We employ a workstation equipped with an NVIDIA GeForce 2080Ti GPU to host the DNN model.

Software. We implement SLNet mainly for benchmark analysis. The data[2] is collected with a Linux shell script, and CSI measurements are preprocessed [36] with Matlab. The dataset [70] and code [69] is available to public. Instructions to use this dataset can be found in our released tutorial [68]. The PyTorch [34] library is adopted to implement the custom complex-valued neurons. Raw CSI is preprocessed in a similar way as in [90]. The SEN module is trained offline with randomly generated spectrums. A total of 5,000 epochs is used to train the SEN models, each of which contains 100 batches × 128 instances of generated spectrums. We pre-train an SEN for each resolution of the spectrogram. Three resolutions are used with window sizes of 125 ms, 251 ms, and 501 ms, respectively. The TAN is trained by the data measured from real deployment scenarios and enhanced by the pre-trained SEN. All the models are trained with Adam optimizer at a learning rate of 0.001. Batch normalization and dropout techniques are applied during training. In practice, the model can be trained offline, except for the use case of

[2] All experiments that involve humans satisfy our IRB requirements.

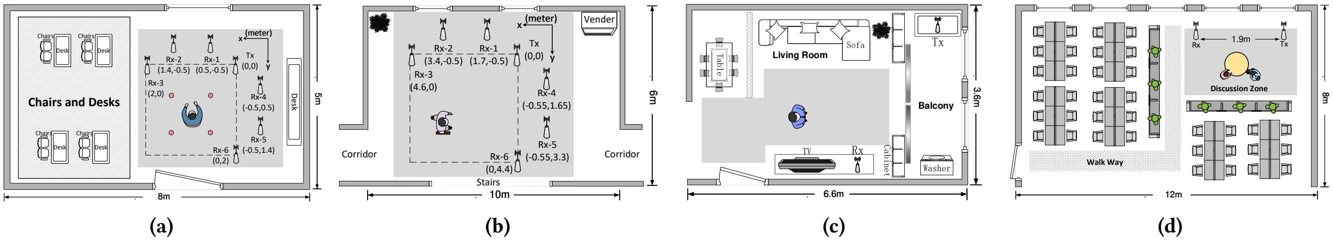


Figure 9: Experimental settings established in SLNet. (a) Classroom for gesture recognition. (b) Hall for gait identification. (c) Apartment for fall detection. (d) Office for breath estimation.

gait recognition where a user needs to register himself first.

4.2 Experiments

We evaluate SLNet in four WiFi sensing applications, including gesture recognition, gait identification, fall detection, and breath estimation. We mainly focus on WiFi CSI sensing in this paper and leave it for the future to explore SLNet's potential for other modalities like acoustic sensing.

Gesture recognition. Device-free gesture recognition [2, 90] is one of the core enablers for human-computer interaction. To evaluate the performance of SLNet for gesture recognition, we conduct experiments in a classroom (sketched in Fig. 9a). One WiFi transmitter and six receivers are placed at a height of 110 cm to capture the motion of human arms. The users are asked to stand at the five marked positions and face the second, third, or fourth quadrant. Eight users (6 males and 2 females) participate in this experiment, with heights varying from 155 cm to 185 cm and ages from 22 to 28. They perform 16 gestures, including 6 sign gestures (push and pull, sweep, clap, slide, draw a circle, and draw zigzag), and 10 input gestures (draw digits 0 to 9). We collect a total of 6,000 data samples (8 people × 6 gestures × 125 instances) for the sign gestures and 5,000 samples (2 people × 10 gestures × 250 instances) for the input gestures. The sign gestures are used in §5.1, and the input gestures are used in the other evaluations. We use the ratio between the number of correctly recognized gestures and the number of all samples as metrics.

Gait identification. Gait has been exploited [18, 64] for human identification. To evaluate the performance of SLNet for gait identification, we conduct experiments in a hall (sketched in Fig. 9b). The devices are placed similarly to that in the gesture experiment. The users are asked to walk freely across the center of the area with eight directions separating 45 degrees apart. Eleven users (7 males and 4 females) participate in this experiment, and their heights vary from 155 cm to 183 cm, and their ages vary from 20 to 26. We collect a total of 3,600 data samples, among which 2,800 samples (7 people × 8 directions × 50 instances) are from 7 users, and 800 samples (4 people × 8 directions × 25 instances) are from the other 4 users. We use the ratio between the number of correctly identified gait samples and the number of all samples as metrics.

Motion	Types	Sub variations
Fall	sit-then-fall, lose-balance, kneel-then-fall, trip walk-then-fall, slip	forward, backward, lateral, on-position
Normal	walk, sit-down/stand-up, run, bend-and-pickup, squat, dance, open/close door, open/close fanner	

Table 1: Fall and normal activities in SLNet

Fall detection. Fall is a major cause of impairment among senior citizens, and some works [33, 52] have tried to detect falls with wireless signals. To evaluate the performance of SLNet for fall detection, we conduct experiments in an apartment (sketched in Fig. 9c). A pair of WiFi devices are placed at the height of 135 cm and 40 cm, respectively, in the living room and the balcony at a distance of 4.5 m. We recruit a voluntary family with 5 members (two males and three females with heights varying from 160 cm to 181 cm and ages varying from 20 to 50). The observed fall and normal activities are described in Tab. 1. When collecting fall data, we require the users to wear protective gear, and the floor is covered with foam. Additionally, we augment the dataset by leveraging a manikin [25] to fall. In total, we collect 2,000 normal instances and 556 fall instances, among which 300 falls are from the manikin, and 256 are from the five users. We use precision and recall as metrics [45].

Breath estimation. Breath rate [67, 78, 83] is an important vital sign that can indicate the condition of physical health. To evaluate the performance of SLNet for breath estimation, we conduct experiments in an office (sketched in Fig. 9d). A pair of WiFi devices are placed at a height of 1 meter to capture the signal reflection from seated participants in the discussion zone. We recruit three participants with heights varying from 172 cm to 185 cm and ages varying from 22 to 26. For each experiment, two of the participants sit in the chairs and breathe naturally. We collect a total of 19 groups of data with a duration of approximately 44 minutes. We use the Breath Per Minute (BPM) error between the estimated respiration rate and the ground truth as metrics.

We first conduct experiments on all tasks to demonstrate the generality for multiple tasks. Then we carry out ablation and parameter study with gesture recognition as the example task due to the space limit. Unless otherwise stated, the results below are obtained on a 10-fold validation basis where we randomly split the datasets into training and testing parts.

5 EVALUATION

5.1 Comparison Study

5.1.1 Comparison between learning models. To validate the effectiveness of the whole SLNet for wireless sensing, we compare it with 12 typical neural models used in different modalities, including WiFi sensing, FMCW-radar sensing, acoustic sensing, computer vision, and other tasks that leverage Complex-Valued Neural Networks (CVNNs). It is worth noting that the implementations of these networks differ slightly from those in the cited works. As the networks presented in those citations are designed for tasks beyond WiFi sensing, we borrow the backbone architectures and customize them for our own tasks and datasets. The comparison is performed with multiple metrics in terms of model complexity and recognition performance. For model complexity, we evaluate the number of parameters of the models, which is perceived as an effective estimate of the memory requirements and training overhead. For recognition performance, the metrics are discussed in §4.2.

WiFi (4 baseline models): We compare a hybrid CNN-RNN model similar to [23, 90] with three convolutional layers, a GRU layer, and four FC layers; an adversarial model [8, 22] with three convolutional layers as the feature extractor, two FC layers as the activity recognizer, and two FC layers as domain discriminator; an encoder-decoder model with ten FC layers as in [39, 79]; and the time-frequency feature learning model introduced in STFNet [73].

FMCW (2 baseline models): Similar to [87], we design an adversarial model with three convolutional layers and a GRU layer as the feature extractor, two FC layers as the activity recognizer, and two FC layers as the domain discriminator. We further compare a CNN model with three convolutional layers and four FC layers as in [84, 86]. The model is applied to the real and imaginary parts of complex-valued features separately.

Acoustic (1 baseline model): We compare an RNN model with a GRU layer and six FC layers as in [30].

Vision (3 baseline models): We compare three baseline models including VGG-11 [40], resnet-18 [15], and densenet [20]. The input and output layers are reshaped to accommodate our tasks.

CVNN (2 baseline models): We compare two **CVNN** networks that are not originally designed for wireless sensing tasks. Similar to [17, 32], we implement an encoder-decoder model with five complex-valued FC layers and three real-valued FC layers. Similar to [46], we evaluate a CNN model with two complex-valued convolutional layers, two complex-valued FC layers, and two real-valued FC layers.

Modality	Ref.	Gesture	Gait	Fall[1]	Para[2]
WiFi	[23, 90]	90.6%	95.1%	92.8%, 96.3%	1.07M
	[8, 22]	89.0%	96.6%	96.4%, 84.3%	2.72M
	[39, 79]	84.3%	83.3%	96.8%, 93.8%	5.77M
	[73][3]	78.9%	70.9%	95.5%, 96.8%	0.06M
FMCW	[87]	88.0%	95.4%	96.0%, 96.0%	1.06M
	[84, 86]	91.6%	96.4%	99.7%, 95.7%	2.76M
Acoustic	[30]	89.6%	95.4%	90.6%, 98.3%	6.08M
Vision	[40]	88.3%	90.1%	95.3%, 95.3%	128.8M
	[15]	91.9%	96.6%	97.0%, 95.6%	11.18M
	[20]	91.0%	97.7%	99.8%, 96.3%	6.96M
CVNN	[17, 32]	72.3%	96.0%	95.2%, 93.7%	115.6M
	[46]	92.0%	96.3%	98.4%, 93.8%	2.94M
WiFi	**SLNet**	**96.6%**	**98.9%**	**99.8%, 97.2%**	**1.48M**

Table 2: Comparison against 12 baseline models. [1] The two metrics are precision and recall. [2] Number of parameters in Million. [3] Trained with 10,000 epochs to converge.

The input of the baseline model [73] is CSI amplitude with a size of $(T, 30, C)$, where T represents the time snapshots of data samples, *30* represents the number of subcarriers, and C represents the number of WiFi antennas. The input of the other baseline models is raw DFS with a size of $(121, T, C)$, where 121 represents the frequency bins within $[-60, 60]$ Hz. For the signals collected by multiple antennas, we perform PCA analysis on the subcarriers of all three antennas and use the principle components for spectrogram1 analysis.

Tab. 2 presents the performance of the baseline models and SLNet. Three key observations can be derived from the results. First, the models used in computer vision tasks achieve better performance than most of the other baseline models. This is because the vision models are heavily parameterized, which endows them with strong representation capabilities. However, for wireless sensing tasks, a less parameterized neural network is preferable due to the cumbersome data collection and the lack of the wide availability of public datasets. SLNet is designed for this purpose. Second, the models that work in complex domain [17, 32, 46, 84, 86] achieve better performance than those real-valued models. This verifies our assumption that the phase of wireless signals embodies valuable information. SLNet strives to exploit this information with its custom neurons. Third, the advantage of SLNet over baseline models is more significant for the gesture and gait tasks than the fall detection task. This is because fall detection is a binary classification problem that is simpler than the other tasks. SLNet is advantageous in complicated

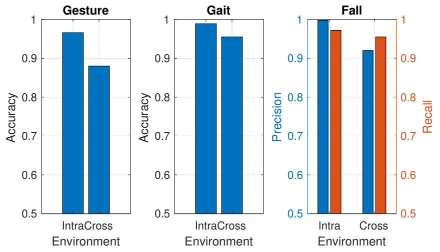


Figure 10: Performance across environments.

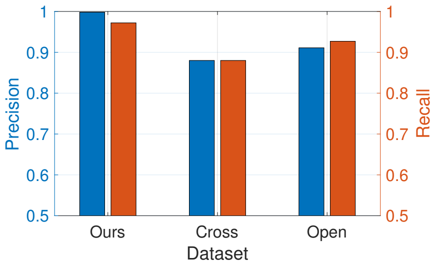


Figure 11: Performance on open dataset [33].

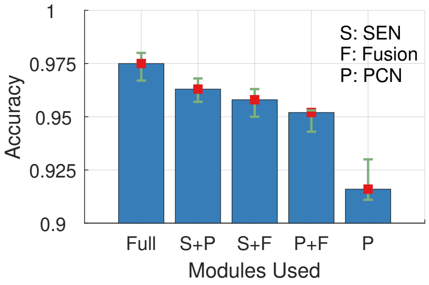


Figure 12: Ablation study on the modules of SLNet.

human sensing scenarios. The most relevant work to SLNet is STFNet [73], which designs customized neural operations to process sensor and RF data. However, while the performance on fall detection is comparable with the other models, it is not desirable for gesture and gait identification tasks. This is because STFNet is built upon the traditional STFT that suffers from spectral leakage. This leakage, when appearing in the spectrogram with clustered frequency components (typical for gesture and gait motion), will distort and even mislead the learning process when searching for the underlying signal structures. The lightly parameterized model in STFNet makes this problem even more challenging. On the contrary, SLNet alleviates this leakage with SEN before learning the hidden features with task-specific networks, ensuring high-fidelity motion representations.

5.1.2 Performance for unseen environments. Wireless sensing systems are prone to environmental changes when deployed in various environments. In this experiment, we evaluate the performance of SLNet when it is applied in *unseen* environments/users after training. Specifically, for the gesture recognition task, we set up another system in an office, which has different layouts and sizes with the classroom illustrated in Fig. 9a. Four volunteers participate in the experiments, and we collect a total of 3,000 gesture samples. For the gait identification task, we collect 600 instances of walking samples from three volunteers in a discussion room, which has a smaller size and more furniture compared with the hall illustrated in Fig. 9b. For the fall detection task, we collect 500 walking samples and 200 falling samples in an office room with different layouts and sizes from the apartment in Fig. 9c. For each task, we train SLNet from scratch with the data collected from one site and test it with the data collected from another.

Fig. 10 demonstrates the performance. As is shown, when deployed in unseen environments without any model adaptation, the performance of SLNet slightly decreases but is still encouraging. SLNet is based on the spectrograms of wireless signals, making it more robust to the surrounding static objects. However, this also makes it prone to changes in relative locations and orientations between users and devices. This is because the frequency components in RF spectrograms are induced by the Doppler shifts, which depend on the moving direction and locations. SLNet is not particularly designed to resolve this problem, yet we believe it can be further addressed by domain adaptation mechanisms [10, 22, 90].

5.1.3 Performance on open datasets. We further evaluate SLNet on a publicly available open datasets. We mainly study fall detection using the dataset released in [33], since it is nearly impossible to find open datasets that have the same types of gestures for gesture recognition or have the same users for gait recognition. This dataset has a total of 181 clearly annotated fall samples and 297 samples of normal activities collected from five rooms of a typical apartment. We compare three settings: 1) We only use our own dataset and split it into non-overlapped train and test parts; 2) We train the model with our dataset and test it with the open dataset; 3) We only use the open dataset and split it into separate train and test parts. The results in Fig. 11 show that SLNet achieves close to 90% precision and recall when trained on our dataset and tested on the open dataset. Despite being degraded, we believe the performance is still encouraging, as the testing data are from completely different and unseen settings with different users, environments, devices, types of falls, etc. Compared with the performance in [33], the precision of our system increases by around 5%, demonstrating the effectiveness of the proposed spectrograms learning pipeline. Considering the promising performance of SLNet in both intra-domain and cross-domain scenarios, We believe SLNet points a valuable direction to applying wireless sensing systems for real-world applications.

5.2 Ablation Study

SLNet consists of three key modules, i.e., SEN, Fusion, and PCN. To validate their effectiveness, we perform an ablation study for these modules. To do so, we remove it from SLNet while adapting the two modules to the input and output

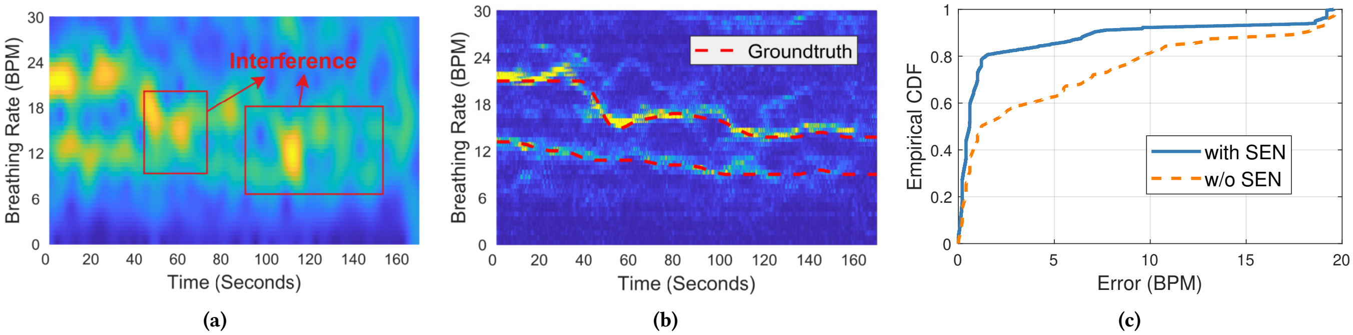


Figure 13: Accuracy for breath estimation. (a) The raw spectrogram from traditional FFT. Spectral leakage causes severe interference for the close frequency components, making it hard to differentiate the breath rates of different people. (b) The enhanced spectrogram with SEN. Frequency components can be clearly discriminated. (c) The accuracy of breath rate estimation with raw and enhanced spectrograms.

format. Specifically, for the SEN module, the originally leaked spectrograms are used as the input of the Fusion and PCN modules. For the Fusion module, only the spectrograms with a window size of 251 ms are enhanced by the SEN and used as the input of the PCN. For the PCN module, the output of the Fusion module is fed to two conventional CNN layers and one max-pooling layer, followed by one FC layer as the output layer. In addition, we further remove both the SEN and the Fusion modules to evaluate their joint performance.

As shown in Fig. 12, the accuracy decreases from 97.5% to 96.2%, 95.6% and 95% when the Fusion, PCN and SEN are removed respectively. The accuracy further decreases to 91.2% when only the PCN is used. The result of the ablation study demonstrates the effectiveness of the three modules of SLNet. It is also worth noting that the benefit brought by the SEN module is more significant than others, meaning that the spectral leakage problem cannot be neglected in wireless sensing tasks, and SLNet successfully resolve the problem.

Multi-Person Breathing Rate Estimation Performance. Even though SLNet is designed for motion recognition tasks that attempt to leverage deep learning schemes, its components are valuable beyond that scope. For example, the SEN module can be used to mitigate the spectral leakage induced by the Fourier transform, which could potentially boost the performance of sensing systems that involve spectral analysis. To validate the effectiveness of SEN, we design a human breath estimation experiment in this part.

Breath estimation plays an important role in healthcare, and some recent works [51, 78] exploit the feasibility of using WiFi signals to estimate the respiration rate. A typical way to do this job is to convert the time domain CSI measurements into a frequency domain spectrogram and characterize respiration rates with the prominent frequency components. However, when multiple people breathe concurrently, the spectral leakage problem will severely blur the spectrogram components and make it difficult to discriminate the respiration of different people. With the SEN module, we envision that the leakage will be mitigated or even eliminated, contributing to improved respiration rate estimation accuracy.

In this experiment, two participants sit in chairs and breathe naturally. One of them has just finished some exercise. To obtain ground truth, each of the participants has a smartphone tied to his chest to measure the acceleration of the body induced by breath movements. We apply STFT with a window width of 6,000 (60 seconds) on the acceleration measurements and detect peaks in the spectrogram to represent the respiration rate of each participant. We downsample CSI to 10 Hz and apply STFT with a window width of 251 (25.1 seconds) to get the spectrograms of CSI measurements. We then apply SEN to the spectrograms and pinpoint the two most prominent peaks therein to characterize the respiration rates of the two participants.

Fig. 13a and Fig. 13b demonstrate the raw and enhanced spectrograms of WiFi. As can be seen, the raw spectrogram is severely distorted by the leakage effect, and the frequency components corresponding to the two participates are blurred. By applying SEN, two distinct frequency components can be observed and approximate ground truth very well. Fig. 13c presents the empirical CDF of the respiration rate estimation error. With SEN applied, the average error is 2.4 BPM and the 80%-tile error is 1.4 BPM. Without SEN, the performance deteriorates to 5.0 BPM for average error and 9.5 BPM for 80%-tile error. This experiment demonstrates SEN's strong capability of removing spectral leakage. This merit makes it especially suitable for human-centered sensing tasks, where the human-induced frequency components are tightly clustered and demand to be discriminated.

5.3 Parameter Study

5.3.1 Impact of training epochs of the SEN. In practice,

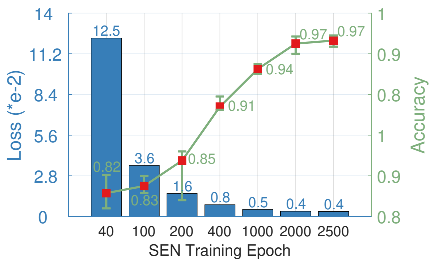


Figure 14: Impact of SEN training epochs.

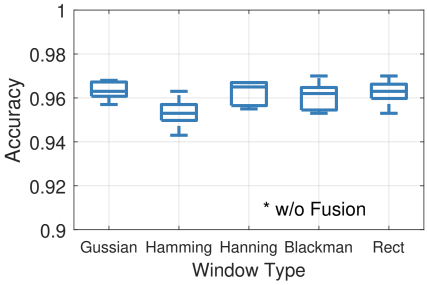


Figure 15: Impact of window type.

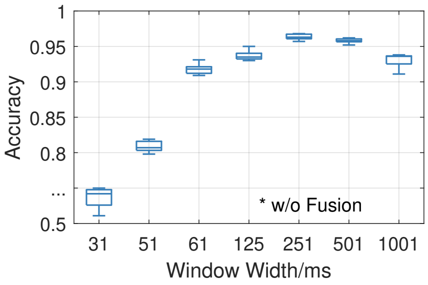


Figure 16: Impact of window width.

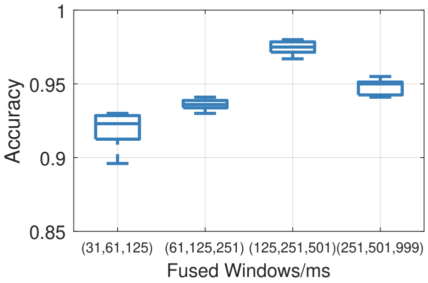


Figure 17: Impact of fused window width.

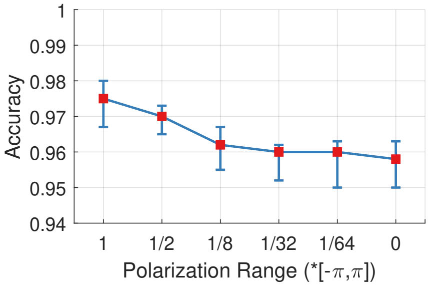


Figure 18: Impact of phase modulation range.

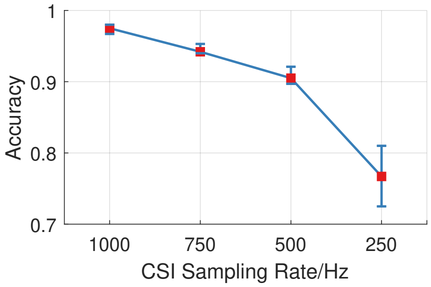


Figure 19: Impact of CSI sampling rate.

we randomly generate data samples during each training epoch. With more training epochs, the SEN should capture the underlying structure of spectrograms better and improve the system recognition accuracy. To reveal the relationship between training epochs and performance, we train the SEN with different numbers of training epochs from 40 to 2500 and integrate them into SLNet. For each SEN, the TAN module is retrained from scratch. We record the validation loss of the SEN and the recognition accuracy of the overall SLNet. As shown in Fig. 14, the validation loss for SEN training tumbles and the overall accuracy proliferates when the epochs increase from 40 to 2,000. The performance becomes stable when the SEN is trained with more epochs.

5.3.2 Impact of window type. Some window functions [13] have been proposed to suppress spectral leakages, such as Hamming, Hanning, and Blackman windows. In this experiment, we evaluate the system performance with regard to these windows. For each window function, we train an independent SEN module with spectrograms calculated with it. The TAN module is then trained from scratch with different SEN modules. We record the overall recognition accuracy concerning different window functions. As shown in Fig. 15, different window functions have little impact on the performance of SLNet, demonstrating the effectiveness of SEN for the removal of spectral leakage.

5.3.3 Impact of window width. In this experiment, we evaluate the impact of the window width on the system performance. Specifically, we train an SEN module for each window width. The Fusion module of SLNet is removed to evaluate each window width. The TAN modules are retrained accordingly, and the overall recognition accuracy is recorded. As shown in Fig. 16, when window width increases from 31 ms to 251 ms, the overall accuracy proliferates from 74% to 96%, but tumbles to 92.5% when window width further increases to 1001 ms. It is because a very small window leads to a coarse-grained frequency resolution, while a very large window cannot capture the rapid change of frequency components in the time domain. The result reveals that a width of 251 ms is the best for the gesture recognition task. However, it is noted that a single-resolution spectrogram for wireless sensing tasks is not the optimal solution, as verified by the ablation study.

5.3.4 Impact of combinations of different window widths. In this experiment, we evaluate the impact of different combinations of window widths. For each combination, we use the corresponding SEN modules to enhance the spectrograms and combine them as holograms. The TAN module is retrained for each combination. The overall recognition accuracy with different combinations is reported. As shown in Fig. 17, the best combination of window widths is $(125, 251, 501)$ ms and the worst is $(31, 61, 125)$ ms. The combination of three windows outperforms either one of these windows independently. The performance could be further improved with more window widths fused at the cost of increased computational complexity, which is limited in edge devices in practice. For SLNet, the combination of $(125, 251, 501)$ ms is adopted to achieve a trade-off between performance and complexity.

5.3.5 Impact of polarization range of PCN. The PCN module of SLNet is designed to extract both local and global features simultaneously from the holograms. In this experiment, we evaluate the impact of the range of the linear phase modulated to spectrograms in SEN. Specifically, the phases are set to be within $k * [-\pi, \pi]$, where k changes from 0 to 1. For each phase range, we retrain the TAN model from scratch. The overall recognition accuracy concerning different phase ranges is reported. As shown in Fig. 18, the accuracy decreases from 97.5% to 96.0% when the polarization range decreases from $[-\pi, \pi]$ to $[0, 0]$, which reveals that the PCN with phase modulation is effective in spectrogram-based sensing tasks.

5.3.6 Impact of CSI sampling rate. The impact of the CSI sampling rate is evaluated in this part. Specifically, we downsample the original CSI streams (1,000 Hz) before spectral analysis and input the corresponding spectrograms in SLNet. Both SEN and TAN are retrained for each downsampling rate. As shown in Fig. 19, when the CSI sampling rate decreases from 1,000 Hz to 500 Hz, the accuracy gradually decrease from 97.5% to 90% and further tumbles to 76% with a sampling rate of 250 Hz. It is because, with a lower sampling rate, the CSI signals have a poorer temporal resolution and cannot capture the rapidly changing frequency components. In addition, the signal-to-noise ratio decreases with the reduced number of samples for spectral analysis. These factors together deteriorate the recognition performance of SLNet.

6 RELATED WORK

Model-based wireless sensing. Model-based wireless sensing works [2, 4, 38, 49, 52, 60, 62, 74] try to establish quantitative relations between wireless signals and human activities via non-learning based approaches. Many applications have been explored and enabled, including gestures [2, 35, 44, 47, 75], walking [56, 57, 64, 76], falls [19, 21, 33, 45], and respiration [1, 27, 58, 61, 78], and tracking [3, 37, 63, 66], etc. These approaches have the benefit of being interpretable and usually efficient. For example, WiGest [2] empirically builds a link between received signal strength (RSSI) and hand-moving patterns. SMARS [78] exploits breathing estimation by periodicity finding. However, these approaches are constrained by the coarse-grained signal and motion models and are approaching performance limits in real environments. More works are seeking learning-based schemes for better performance, and SLNet is one among them.

Learning-based wireless sensing. Early works mainly rely on signal processing and employ traditional machine learning [48, 56, 57, 59, 64, 76, 77]. With the impressive achievements in computer vision using deep neural networks, more effort [9, 10, 22, 45, 54, 71, 72, 79, 80, 82, 85, 86, 88, 90] has been put into applying deep learning models in wireless sensing tasks. Among them, Widar3.0 [81, 90] leverages CNN and RNN networks to learn from its novel motion feature BVP. RFPose [84], RFPose3D [86], and RFAvatar [85] use CNN models to capture human skeleton and mesh of body. Many works [10, 22, 49, 79, 90] employ sophisticated network architectures like adversarial learning, transfer learning, and meta-learning to solve the environment-dependency problem of wireless sensing, while others aim to reduce cumbersome data collection for training [7, 11, 42, 53, 65]. Some works *e.g.*, [84–86] on FMCW sensing, have further considered customized models for the unique properties of RF data. Existing works either learn from the time series of raw CSI, with both amplitude and phase, or convert them into the frequency-domain representation or other feature space. Despite some time-domain approaches for speech separation [29, 43], recent works like STFNets [73], which extends DeepSense [72], and UniTS [26] both pursue and demonstrate superior performance of temporal-spatial learning with STFT operators. Noticing phase encodes essential spatial information, complex-valued neural networks [5] have been explored in the DL community [17, 32, 46] and exploited especially for radar sensing [89], acoustic sensing and speech processing [28, 50].

7 CONCLUSION

This paper presents SLNet, a spectrogram analysis-deep learning co-design for deep wireless sensing. We demonstrate SLNet's remarkable performance in gesture recognition, gait recognition, fall detection, and breath estimation, showing the highest accuracy and lowest computation compared to the state-of-the-art models. We believe SLNet is a unique deep-learning framework for WiFi sensing. At the same time, the techniques can be used, jointly or separately, to augment the spectrogram quality and enhance learning performance for many applications in signal estimation, frequency analysis, sensing with acoustic/millimeter-wave signals, etc.

ACKNOWLEDGMENTS

We thank our shepherd Dr. Srikanth Kandula and the anonymous reviewers for their feedback, which greatly improved the paper. This work is supported in part by the NSFC under grants No. 61832010, No. 62202262, and No. 62222216, and Hong Kong RGC ECS under grant 27204522.

REFERENCES

[1] Heba Abdelnasser, Khaled A Harras, and Moustafa Youssef. 2015. UbiBreathe: A ubiquitous non-invasive WiFi-based breathing estimator. In *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. 277–286.

[2] Heba Abdelnasser, Moustafa Youssef, and Khaled A Harras. 2015. Wigest: A Ubiquitous Wifi-based Gesture Recognition System. In *Proceedings of IEEE INFOCOM*.

[3] Fadel Adib, Zachary Kabelac, and Dina Katabi. 2015. Multi-Person Localization via RF Body Reflections. In *Proceedings of USENIX NSDI*.

[4] Kamran Ali, Alex X Liu, Wei Wang, and Muhammad Shahzad. 2017. Recognizing keystrokes using WiFi devices. *IEEE Journal on Selected Areas in Communications* 35, 5 (May 2017), 1175–1190.

[5] Joshua Bassey, Lijun Qian, and Xianfang Li. 2021. A survey of complex-valued neural networks. *arXiv preprint arXiv:2101.12249* (2021).

[6] Holger Boche, Robert Calderbank, Gitta Kutyniok, Jan Vybíral, et al. 2015. *Compressed sensing and its applications*. Springer.

[7] Hong Cai, Belal Korany, Chitra R Karanam, and Yasamin Mostofi. 2020. Teaching rf to sense without rf training measurements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22.

[8] Xi Chen, Hang Li, Chenyi Zhou, Xue Liu, Di Wu, and Gregory Dudek. 2020. FiDo: Ubiquitous Fine-Grained WiFi-Based Localization for Unlabelled Users via Domain Adaptation. In *Proceedings of ACM WWW*.

[9] Zhe Chen, Tianyue Zheng, Chao Cai, and Jun Luo. 2021. MoVi-Fi: Motion-robust vital signs waveform recovery via deep interpreted RF sensing. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 392–405.

[10] Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. 2020. RF-Net: A Unified Meta-Learning Framework for RF-Enabled One-Shot Human Activity Recognition. In *Proceedings of ACM SenSys*.

[11] 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.

[12] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. 2011. Tool Release: Gathering 802.11n Traces with Channel State Information. *SIGCOMM Comput. Commun. Rev.* 41, 1 (2011), 53.

[13] Fredric J. Harris. 1978. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. IEEE* 66, 1 (1978), 51–83.

[14] Haitham Hassanieh. 2018. *The Sparse Fourier Transform: Theory and Practice*. Association for Computing Machinery and Morgan & Claypool.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015).

[16] Geoffrey Hinton, li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Phuongtrang Nguyen, Tara Sainath, and Brian Kingsbury. 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine* 29, 6 (2012), 82–97.

[17] Akira Hirose and Shotaro Yoshida. 2012. Generalization Characteristics of Complex-Valued Feedforward Neural Networks in Relation to Signal Coherence. *IEEE Transactions on Neural Networks and Learning Systems* 23, 4 (2012), 541–551.

[18] 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 ACM CHI*.

[19] Yuqian Hu, Feng Zhang, Chenshu Wu, Beibei Wang, and K. J. Ray Liu. 2020. A WiFi-based Passive Fall Detection System. In *Proceedings of IEEE ICASSP*.

[20] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. 2016. Densely Connected Convolutional Networks. *CoRR* (2016).

[21] Sijie Ji, Yaxiong Xie, and Mo Li. 2022. SiFall: Practical Online Fall Detection with RF Sensing. In *Proceedings of the Twentieth ACM Conference on Embedded Networked Sensor Systems*. 563–577.

[22] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, Wenyao Xu, and Lu Su. 2018. Towards Environment Independent Device Free Human Activity Recognition. In *Proceedings of ACM MobiCom*.

[23] 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 ACM MobiCom*.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of NIPS*.

[25] Laerdal. Accessed 2021. Resusci Anne QCPR Manikin. https://laerdal.com/us/products/simulation-training/resuscitation-training/resusci-anne-qcpr/. (Accessed 2021).

[26] Shuheng Li, Ranak Roy Chowdhury, Jingbo Shang, Rajesh K Gupta, and Dezhi Hong. 2021. UniTS: Short-Time Fourier Inspired Neural Networks for Sensory Time Series Classification. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 234–247.

[27] Jian Liu, Yan Wang, Yingying Chen, Jie Yang, Xu Chen, and Jerry Cheng. 2015. Tracking vital signs during sleep leveraging off-the-shelf wifi. In *Proceedings of the 16th ACM international symposium on mobile ad hoc networking and computing*. 267–276.

[28] Yi Luo, Zhuo Chen, Nima Mesgarani, and Takuya Yoshioka. 2020. End-to-end microphone permutation and number invariant multi-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6394–6398.

[29] Yi Luo and Nima Mesgarani. 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing* 27, 8 (2019), 1256–1266.

[30] Wenguang Mao, Mei Wang, Wei Sun, Lili Qiu, Swadhin Pradhan, and Yi-Chao Chen. 2019. RNN-Based Room Scale Hand Motion Tracking. In *Proceedings of ACM MobiCom*.

[31] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of ICLR*.

[32] Nils Moenning and Suresh Manandhar. 2018. Complex- and Real-Valued Neural Network Architectures. In *International Conference on Learning Representations (openreview)*. https://openreview.net/forum?id=HkCy2uqQM

[33] Sameera Palipana, David Rojas, Piyush Agrawal, and Dirk Pesch. 2019. FallDeFi: Ubiquitous Fall Detection Using Commodity Wi-Fi Devices. In *Proceedings of ACM IMWUT*.

[34] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[35] Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel. 2013. Whole-home gesture recognition using wireless signals. In *Proceedings of the 19th annual international conference on Mobile computing & networking*. 27–38.

[36] Kun Qian, Chenshu Wu, Zheng Yang, Yunhao Liu, and Kyle Jamieson. 2017. Widar: Decimeter-level passive tracking via velocity monitoring with commodity Wi-Fi. In *Proceedings of ACM MobiHoc*.

[37] Kun Qian, Chenshu Wu, Yi Zhang, Guidong Zhang, Zheng Yang, and Yunhao Liu. 2018. Widar2.0: Passive Human Tracking with a Single

Wi-Fi Link. In *Proceedings of ACM MobiSys*.

[38] Kun Qian, Chenshu Wu, Zimu Zhou, Yue Zheng, Zheng Yang, and Yunhao Liu. 2017. Inferring Motion Direction Using Commodity Wi-Fi for Interactive Exergames. In *Proceedings of ACM CHI*.

[39] Cong Shi, Jian Liu, Hongbo Liu, and Yingying Chen. 2017. Smart User Authentication through Actuation of Daily Activities Leveraging WiFi-Enabled IoT. In *Proceedings of ACM MobiHoc*.

[40] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[41] SLNet. 2022. https://github.com/SLNetRelease/SLNetCode. (2022).

[42] Ruiyuan Song, Dongheng Zhang, Zhi Wu, Cong Yu, Chunyang Xie, Shuai Yang, Yang Hu, and Yan Chen. 2022. RF-URL: unsupervised representation learning for RF sensing. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 282–295.

[43] Daniel Stoller, Sebastian Ewert, and Simon Dixon. 2018. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185* (2018).

[44] Sheng Tan and Jie Yang. 2016. WiFinger: Leveraging commodity WiFi for fine-grained finger gesture recognition. In *Proceedings of the 17th ACM international symposium on mobile ad hoc networking and computing*. 201–210.

[45] Yonglong Tian, Guang-He Lee, Hao He, Chen-Yu Hsu, and Dina Katabi. 2018. RF-Based Fall Monitoring Using Convolutional Neural Networks. *Proceedings of ACM IMWUT* (2018).

[46] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, João Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. 2018. Deep Complex Networks. (2018). arXiv:1705.09792

[47] Raghav H Venkatnarayan, Griffin Page, and Muhammad Shahzad. 2018. Multi-user gesture recognition using WiFi. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 401–413.

[48] Raghav H. Venkatnarayan, Griffin Page, and Muhammad Shahzad. 2018. Multi-User Gesture Recognition Using WiFi. In *Proceedings of ACM MobiSys*.

[49] Aditya Virmani and Muhammad Shahzad. 2017. Position and Orientation Agnostic Gesture Recognition Using WiFi. In *Proceedings of ACM MobiSys*.

[50] Anran Wang, Maruchi Kim, Hao Zhang, and Shyamnath Gollakota. 2022. Hybrid neural networks for on-device directional hearing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11421–11430.

[51] Fengyu Wang, Feng Zhang, Chenshu Wu, Beibei Wang, and K. J. Ray Liu. 2020. Respiration Tracking for People Counting and Recognition. *IEEE Internet of Things Journal* 7, 6 (2020), 5233–5245.

[52] Hao Wang, Daqing Zhang, Yasha Wang, Junyi Ma, Yuxiang Wang, and Shengjie Li. 2017. RT-Fall: A Real-Time and Contactless Fall Detection System with Commodity WiFi Devices. *IEEE Transactions on Mobile Computing* 16, 2 (2017), 511–526.

[53] Jie Wang, Qinhua Gao, Xiaorui Ma, Yunong Zhao, and Yuguang Fang. 2020. Learning to sense: Deep learning for wireless sensing with less training efforts. *IEEE Wireless Communications* 27, 3 (2020), 156–162.

[54] Mei Wang, Wei Sun, and Lili Qiu. 2021. MAVL: Multiresolution Analysis of Voice Localization. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*. 845–858.

[55] Wei Wang, Alex X. Liu, Shahzad Muhammad, Kang Ling, and Sanglu Lu. 2017. Device-Free Human Activity Recognition Using Commercial WiFi Devices. *IEEE Journal on Selected Areas in Communications* 35, 5 (2017), 1118–1131.

[56] Wei Wang, Alex X Liu, and Muhammad Shahzad. 2016. Gait Recognition Using WiFi Signals. In *Proceedings of ACM UbiComp*.

[57] Wei Wang, Alex X. Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2015. Understanding and Modeling of WiFi Signal Based Human Activity Recognition. In *Proceedings of ACM MobiCom*.

[58] Xuyu Wang, Chao Yang, and Shiwen Mao. 2017. TensorBeat: Tensor decomposition for monitoring multiperson breathing beats with commodity WiFi. *ACM Transactions on Intelligent Systems and Technology (TIST)* 9, 1 (2017), 1–27.

[59] Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu. 2014. E-eyes: Device-free Location-oriented Activity Identification Using Fine-grained WiFi Signatures. In *Proceedings of ACM MobiCom*.

[60] Yuxi Wang, Kaishun Wu, and Lionel M. Ni. 2017. WiFall: Device-Free Fall Detection by Wireless Networks. *IEEE Transactions on Mobile Computing* 16, 2 (2017), 581–594.

[61] Chenshu Wu, Zheng Yang, Zimu Zhou, Xuefeng Liu, Yunhao Liu, and Jiannong Cao. 2015. Non-invasive detection of moving and stationary human with WiFi. *IEEE Journal on Selected Areas in Communications* 33, 11 (2015), 2329–2342.

[62] Chenshu Wu, Zheng Yang, Zimu Zhou, Kun Qian, Yunhao Liu, and Mingyan Liu. 2015. PhaseU: Real-time LOS identification with WiFi. In *Proceedings of IEEE INFOCOM*.

[63] Chenshu Wu, Feng Zhang, Yusen Fan, and KJ Ray Liu. 2019. RF-based inertial measurement. In *Proceedings of the ACM Special Interest Group on Data Communication*. 117–129.

[64] Chenshu Wu, Feng Zhang, Yuqian Hu, and K. J. Ray Liu. 2020. GaitWay: Monitoring and Recognizing Gait Speed Through the Walls. *IEEE Transactions on Mobile Computing* (2020).

[65] Rui Xiao, Jianwei Liu, Jinsong Han, and Kui Ren. 2021. OneFi: One-Shot Recognition for Unseen Gesture via COTS WiFi. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 206–219.

[66] Yaxiong Xie, Jie Xiong, Mo Li, and Kyle Jamieson. 2019. MD-Track: Leveraging Multi-Dimensionality for Passive Indoor Wi-Fi Tracking. In *Proceedings of ACM MobiCom*.

[67] Xiangyu Xu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Linghe Kong, and Minglu Li. 2019. BreathListener: Fine-Grained Breathing Monitoring in Driving Environments Utilizing Acoustic Signals. In *Proceedings ACM MobiSys*.

[68] Zheng Yang, Yi Zhang, Guoxuan Chi, and Guidong Zhang. 2022. Hands-on Wireless Sensing with Wi-Fi: A Tutorial. (2022). https://arxiv.org/abs/2206.09532

[69] Zheng Yang, Yi Zhang, Kun Qian, and Chenshu Wu. 2023. SLNet Release Code. https://github.com/SLNetRelease/SLNetCode. (2023).

[70] Zheng Yang, Yi Zhang, Guidong Zhang, and Yue Zheng. 2020. Widar 3.0: WiFi-based Activity Recognition Dataset. (2020). https://doi.org/10.21227/7znf-qp86

[71] Zheng Yang, Yi Zhang, and Qian Zhang. 2022. Rethinking Fall Detection With Wi-Fi. *IEEE Transactions on Mobile Computing* (2022).

[72] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. 2017. DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing. In *Proceedings of ACM WWW*.

[73] Shuochao Yao, Ailing Piao, Wenjun Jiang, Yiran Zhao, Huajie Shao, Shengzhong Liu, Dongxin Liu, Jinyang Li, Tianshi Wang, Shaohan Hu, Lu Su, Jiawei Han, and Tarek Abdelzaher. 2019. STFNets: Learning Sensing Signals from the Time-Frequency Perspective with Short-Time Fourier Neural Networks. In *Proceedings of ACM WWW*.

[74] Nan Yu, Wei Wang, Alex X. Liu, and Lingtao Kong. 2018. QGesture: Quantifying Gesture Distance and Direction with WiFi Signals. *Proceedings of ACM IMWUT* 2, 1 (2018), 23.

[75] Nan Yu, Wei Wang, Alex X Liu, and Lingtao Kong. 2018. QGesture:

Quantifying gesture distance and direction with WiFi signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–23.

[76] Yunze Zeng, Parth H Pathak, and Prasant Mohapatra. 2016. WiWho: WiFi-Based Person Identification in Smart Spaces. In *Proceedings of ACM/IEEE IPSN*.

[77] Shuangjiao Zhai, Zhanyong Tang, Petteri Nurmi, Dingyi Fang, Xiaojiang Chen, and Zheng Wang. 2021. RISE: Robust wireless sensing using probabilistic and statistical assessments. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 309–322.

[78] Feng Zhang, Chenshu Wu, Beibei Wang, Min Wu, Daniel Bugos, Hangfang Zhang, and KJ Ray Liu. 2019. SMARS: Sleep monitoring via ambient radio signals. *IEEE Transactions on Mobile Computing* 20, 1 (2019), 217–231.

[79] Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Tapio Nurmi, and Zheng Wang. 2018. CrossSense: Towards Cross-Site and Large-Scale WiFi Sensing. In *Proceedings of ACM MobiCom*.

[80] Yi Zhang, Zheng Yang, Guidong Zhang, Chenshu Wu, and Li Zhang. 2021. XGest: Enabling Cross-Label gesture recognition with RF signals. *ACM Transactions on Sensor Networks (TOSN)* 17, 4 (2021), 1–23.

[81] 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* (2021).

[82] Yi Zhang, Yue Zheng, Guidong Zhang, Kun Qian, Chen Qian, and Zheng Yang. 2021. GaitSense: towards ubiquitous gait-based human identification with Wi-Fi. *ACM Transactions on Sensor Networks (TOSN)* 18, 1 (2021), 1–24.

[83] Mingmin Zhao, Fadel Adib, and Dina Katabi. 2016. Emotion Recognition Using Wireless Signals. In *Proceedings of ACM MobiCom*.

[84] 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 IEEE/CVF CVPR*.

[85] Mingmin Zhao, Yingcheng Liu, Aniruddh Raghu, Hang Zhao, Tianhong Li, Antonio Torralba, and Dina Katabi. 2019. Through-Wall Human Mesh Recovery Using Radio Signals. In *Proceedings of IEEE/CVF ICCV*.

[86] 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 ACM SIGCOMM*.

[87] 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 *Proceedings of ACM ICML*.

[88] Tianyue Zheng, Zhe Chen, Shuya Ding, and Jun Luo. 2021. Enhancing RF sensing with deep learning: A layered approach. *IEEE Communications Magazine* 59, 2 (2021), 70–76.

[89] Tianyue Zheng, Zhe Chen, Shujie Zhang, Chao Cai, and Jun Luo. 2021. MoRe-Fi: Motion-robust and Fine-grained Respiration Monitoring via Deep-Learning UWB Radar. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 111–124.

[90] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2019. Zero-effort cross-domain gesture recognition with Wi-Fi. In *Proceedings of ACM Mobisys*.

[91] Han Zou, Yuxun Zhou, Jianfei Yang, Weixi Gu, L. Xie, and C. Spanos. 2018. WiFi-Based Human Identification via Convex Tensor Shapelet Learning. In *Proceedings of AAAI*.