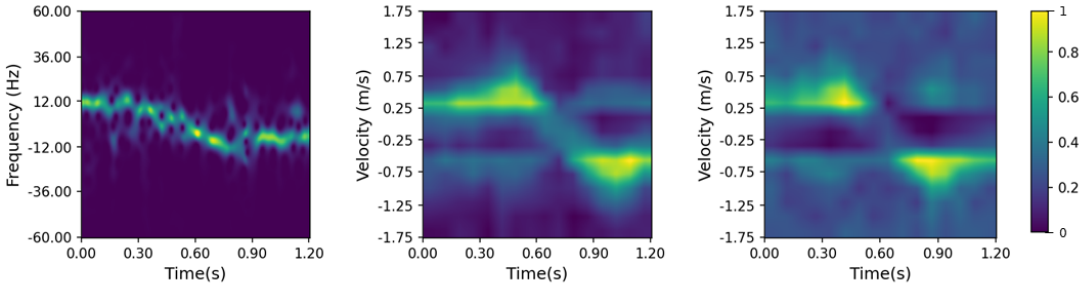
This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3349378

WiFi2Radar: Orientation Independent Single Receiver WiFi Sensing via WiFi to Radar Translation

Isura Nirmal, *Member, IEEE,* Abdelwahed Khamis, *Member, IEEE,* Mahbub Hassan, *Senior Member, IEEE,* Wen Hu, *Senior Member, IEEE,* Rui Li, and Avinash Kalyanaraman



(a) WiFi CSI Spectrogram (b) Radar Doppler Spectrogram (c) Spectrogram converted from (a) using our deep learning translator

Fig. 1: Converting noisy WiFi CSI to high-precision radar Doppler using deep learning. (a) WiFi CSI spectrogram of a Leg Swing activity performed by a subject at 20^0 orientation to a single WiFi receiver; (b) High-precision Doppler spectrogram produced by a TI IWR1443 [1] mmWave radar while observing the same activity simultaneously from the front; (c) Spectrogram converted from (a) using our deep learning translator trained with both WiFi and the corresponding radar data of the same activity. (b) and (c) have a structural similarity index measure (SSIM) of 0.84.

***Abstract*—Recent research has demonstrated the huge potential of WiFi for contactless sensing of human activities. Unfortunately, such sensing is highly sensitive to the relative orientation between the user and the WiFi receivers. To overcome this problem, existing solutions deploy multiple WiFi receivers at precise positions to capture orientation-independent view of the human activity. Orientation independent single receiver WiFi sensing is still considered an open problem. In this paper, we propose a deep neural network architecture that uses radar data during training to learn high-precision Doppler features of human activities from the noisy channel states observed by a single WiFi receiver. Once trained with radars, the network can be used to detect human activities at any arbitrary orientations based only on WiFi signals. Using extensive experiments with millimeter wave radars, we demonstrate that the proposed approach, called WiFi2Radar in this paper, significantly outperforms state-of-the-art for detecting human activities in untrained orientations using only a single WiFi receiver. Our results show that WiFi2Radar can detect orientation-independent human activities with up to 91% accuracy, which outperforms the state-of-the-art by 19%.**

***Index Terms*—WiFi sensing; Device Free; Orientation Independent; Doppler**

I.Nirmal, M.Hassan, W.Hu and R.Li are with the School of Computer Science and Engineering UNSW, 2052,Sydney,Australia
E-mail:(b.isuranirmal@unsw.edu.au,mahbub.hassan@unsw.edu.au, wen.hu@unsw.edu.au,rui.li@unsw.edu.au)

A.Khamis is with CSIRO, Brisbane,Australia
E-mail:(abdelwahed.khamiss@csiro.au)

A.Kalyanaraman is with Cisco Innovation Labs,San Jose, California
E-mail:(avkalyan@cisco.com)

I. Introduction

Wireless modalities (i.e., WiFi, mmWave, RFID, and UWB [2]) for human sensing [3] have attracted significant interest. Recent research demonstrates substantial potential for WiFi-based contactless sensing of human activities [4], [5]. WiFi's pervasive deployment distinguishes it from other modalities. Additionally, it underscores the sensitivity of such sensing to the relative orientation between the human and the WiFi receiver [6], [7].

There are several existing approaches to address the orientation or other "domain" issues of WiFi sensing, where a domain refers to heterogeneity in subjects, orientations, rooms, or any other variables that significantly impact WiFi signals. Widar 3.0 [8] is a proposed approach that fuses signals from multiple WiFi receivers installed at precise locations around the subject to obtain an orientation-independent feature of the human activity, called *BVP (Body Velocity Profile)*. It was shown that deep neural networks trained with BVP were used to achieve a remarkable accuracy of 90% for orientation-independent gesture recognition when 5 WiFi receivers are used. While Widar 3.0 is clearly a step forward in improving the accuracy of WiFi sensing, the requirement of **multiple receivers** in a specific geometry is an impediment to pervasive WiFi sensing. Authors of [7] demonstrated that it is possible to design special gestures that are less sensitive to orientations, but such an approach would not work for detecting many natural and popular activities. Finally, the work in Environment

Independent (EI) [9] has demonstrated that **adversarial deep learning** can improve WiFi sensing accuracy across different environments by forcing the network to extract features that are not dependent on the environment. While adversarial learning appears to be a useful tool to improve orientation-independent WiFi sensing with a single receiver, its accuracy was found to be limited to around 80%. In addition, the design in [9] was found to be useful when some unlabeled data from the target environment was fed to the network, which makes its performance uncertain for pervasive use cases where the trained network may be expected to work in entirely unseen environments. Orientation independent single receiver WiFi sensing, therefore, remains an open challenge.

This paper proposes and explores the feasibility of deep learning to translate WiFi observations of human activities to their corresponding observations from an FMCW (Frequency Modulated Continuous Wave) radar. The motivation here is many-fold. First, radars use an ultra-wide bandwidth of 4 GHz [10] compared to only 20 MHz in typical WiFi use scenarios. This enables radars to observe human movements at much finer resolutions, improving their ability to recognise a specific activity despite environmental noise accurately. Second, radars use tailor-made wireless waves, such as sharp chirps in FMCW radars [10], explicitly designed to sense objects. In contrast, WiFi uses OFDM (Orthogonal Frequency-Division Multiplexing) modulation, optimised for communication but not for sensing. Third, radars use synchronised transmitting and receiving antennas within the same device, making it possible to accurately measure the time-of-flight of the signal reflected from the target.

To realise our concept, we design a multi-component neural network architecture that we call `WiFi2Radar`. A key component of our architecture is a **translator** that is trained to translate WiFi observations of human activities to their corresponding radar observations. Figure 1 shows an example output of our `WiFi2Radar` implementation in converting noisy WiFi CSI to clearer radar images. As WiFi observations are affected by orientations, the translator faces the challenge of producing a radar image that can be used by a classifier to detect human activities accurately without being influenced by orientations. We addressed this challenge by designing an adversarial component, which seeks to learn radar features that are both discriminative with respect to activities but agnostic to the orientation. A key advantage of our adversarial design against that of EI [9] is that it does not require any data from the target domain during training, making it truly orientation independent for pervasive deployment.

To demonstrate the feasibility of `WiFi2Radar`, we simultaneously collect data from a single transmitter-receiver pair of commodity WiFi devices as well as from a commercial millimeter wave radar for 3-6 human activities repeated many thousands of times from 12 different orientations in three different environments. Our results show that `WiFi2Radar` can detect human activities from unseen orientations with up to **91%** accuracy, which outperforms EI [9] by **19%**. Given that our deep learning architecture is similar to that of the EI with the exception of WiFi-to-Radar translation, these experimental results clearly demonstrate the feasibility and benefits of such translations.

Our contributions are three-fold:

- We present the first attempt of using radar as a training aid to learn high-precision motion features from a single pair of commodity WiFi devices. We demonstrate its feasibility through supervised machine learning that builds on the principles of deep learning based translation.
- We design and implement a novel neural network architecture that integrates a `WiFi2Radar` translation component with an activity classifier within an adversarial domain adaptation framework. The proposed design performs significantly better than the baseline activity recognition adversarial models while removing their inherent limitations, namely the reliance on multiple WiFi receivers and target domain data availability.
- We validate and quantify the efficacy of the proposed deep learning architecture on a real activity recognition dataset. Our results show 19% improvement compared to the state-of-the-art approach [9]. Our evaluation also shows that the model has a negligible performance decrease (less than 4%) for unseen subjects and environments.

The rest of the paper is structured as follows. We present the methodology and system model in Section II followed by its evaluation in Section III. We review the state-of-the-art in Section IV. In Section V, we discuss the possible future directions before concluding the paper in Section VI.

II. Methodology and System Model

This section discusses the Doppler acquisition of WiFi and current challenges followed by preprocessing steps to transform Doppler features into a format desirable for Deep learning. We then discuss our System Model of `WiFi2Radar` for an application scenario (i.e., Activity Recognition). Finally, we evaluate the specific architecture for the `WiFi2Radar` to justify the design choice of an adversarial based architecture for an activity recognition use case.

A. WiFi Doppler Preliminaries

Doppler frequency shift [11] measures the negative or positive shift of the observed frequency due to the relative motions of the source or the target as shown in Figure 2. This is observed in both sound waves and electromagnetic waves. Specifically, the amount of frequency change Δf for electromagnetic waves can be calculated as:

$$\Delta f = \pm v f / c, \tag{1}$$

where f is the carrier frequency, c is the speed of light and v is the velocity of the motions. However, this type of straightforward estimation of Doppler is infeasible in WiFi based sensing due to the following reasons.

First, there exists no means for quantifying the rate of change in path length due to the well known random phase error issue [12], which is a limitation that **hinders commodity WiFi from directly capturing the Doppler shift** [13]. Previous efforts resorted to indirect estimation means by evading the noisy phase. Examples include, employing custom

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3349378

WiFi hardware that can capture accurate phase [14] or using alternative amplitude measurements [13] or phase differences across antennas [15] whose values are a *"function"* of the speed of path length change. However, these systems require specialised hardware. Despite these efforts, precise Doppler acquisition from unmodified WiFi is mostly unexplored and yet to be realised.

Second, the bi-static nature of WiFi (i.e. non co-located sender and receiver) makes the path length changes conditional on the motion's position and orientation with respect to the transmitter and the receiver. Thus a specific motion will yield **variant Doppler patterns** at the receiver depending on the configuration [16] which complicates the learning task in activity and gesture recognition applications. One approach that was followed for unifying Doppler patterns is *Invariant Doppler Acquisition*. For example, Widar 3.0 [8] uses variant Doppler profiles observed at multiple receivers to approximate the invariant Doppler (called Body Velocity Profile or BVP [8]) that would have been observed at the user's location (body coordinates). However, this necessitates predefined multiple WiFi receiver spatial deployment considerations and the knowledge of device locations. Another line of research shifts the burden to the machine learning side through models that align the patterns internally in the embedding space. Here, adversarial domain adaptation techniques [9] have attracted a lot of attention. Yet, they assume access to target domain data during training, in addition to the requirements of multiple WiFi receivers. It is generally desirable to have a "sealed" model ready to operate in a new context without re-training.

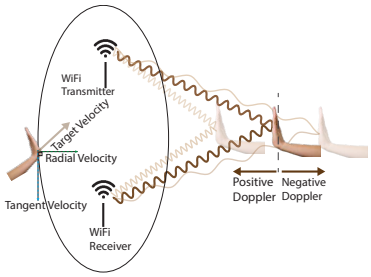


Fig. 2: Velocity components and Doppler Effect in WiFi sensing

Figure 2 further shows a typical WiFi setup with a T_x and R_x pair, where the area of interest can be considered as the foci of an ellipse and a human subject performs motions of interest (e.g., activities) on the ellipse. Here, the reflected signal path change is caused by the radial velocity of the human motions and contains some information on human motion velocities [16]. As mentioned earlier, the Doppler information generated from commodity WiFi devices lacks the full Doppler information, the performance when identifying the specific human motions has its limitation when inferring motions with direction as multiple human velocities generate the same radial velocity components.

In the literature, this limitation is addressed by adding multiple receivers in specific configurations to capture WiFi signals and concatenating the Doppler spectrograms [8], [15]. Chen et al. [17] introduced a novel deep learning-based pipeline aimed at enhancing low-light photographs marred by noise, successfully restoring them to their original clarity. This advancement effectively addresses the issue of noise in images produced by cost-effective CMOS sensors.

Inspired by Chen et al. [17] and considering signal modelling, the pertinent Doppler information derived from WiFi signals becomes obscured and intertwined with various other Doppler components that may not contribute to the gesture recognition task. These unwanted components can be seen as noise. However, our ground truth Doppler data boasts superior Doppler acquisition capabilities (due to synchronised transceivers and higher bandwidth), capturing precise Doppler information relevant to the intended gesture. Therefore, we propose a similar learning-based approach for the restoration of noisy and incomplete Doppler information extracted from WiFi signals, utilising more precise Doppler information.

After the successful acquisition of high-precision Doppler spectrogram information, we extend our work by proposing a generalised human-sensing system via Adversarial Domain Adaptation that can perform well in both in-domain (source domain or orientation) and cross-domain (target domain or orientation) setups, requiring **zero target domain orientation data**.

B. CSI Preliminaries

Computer vision-based deep learning workflow techniques [17] inspire the proposed approach. In this subsection, we introduce a signal processing pipeline to extract Doppler features from raw WiFi data and generate spectrograms that can be used as input to the deep learning model.

The WiFi radio signal is continuously affected by the motion in the background. The CSI (Channel State Information) measurements contain the CFR (Channel Frequency Response) values of the wireless communication channels. We define $X(f,t)$ and $Y(f,t)$ as the transmitted and received signal, respectively, and the following Eq. (2) shows how the two signals are related.

$$Y(f,t) = H(f,t) \times X(f,t), \tag{2}$$

where each $H(f,t)$ entry is the complex CFR value. In the IEEE 802.11n standard, CSI is obtained for each of the subcarriers used in OFDM (Orthogonal Frequency Division Multiplexing) subcarriers. The dimensions of CSI measurements are:

$$m \times N_{Tx} \times N_{Rx}, \tag{3}$$

where m is the number of OFDM subcarriers, and N_{Tx} and N_{Rx} are the numbers of transmitting and receiving antennas respectively. Since CSI is measured during packet reception, this adds a time axis to the CSI dimension and finally CSI is represented as **a time series**. Henceforth, we refer CSI stream with *x-axis* being CSI data collection time period which we discuss in the coming section. The CSI streams can be extracted from COTS devices such as Intel 5300 NIC [18], which is used in this paper. Specifically, we choose a 5 GHz WiFi band with 40 MHz bandwidth in 802.11n. Intel 5300

NIC measures the CSI in 30 subcarriers. Since N_{Tx} = 1 and N_{Rx} = 3. In our setup, we have 90 CSI streams altogether.

C. From CSI Measurement to Doppler Spectrogram

Raw CSI streams from COTS WiFi radio are noisy. Therefore, we apply an extensive preprocessing and signal processing pipeline to extract Doppler information from the CSI streams. Specifically, we choose a major part of the signal processing pipeline proposed in *WiDance [15]* to generate Doppler spectrograms. We note that *WiDance's* signal processing pipeline makes use of the phase information to generate Doppler information preserving its direction (positive direction as towards and negative direction as backward to the receiver). We summarise the preprocessing steps from *WiDance* as follows.

CSI ($H(f,t)$) can be modelled as the superimposition of the frequency response of the multiple propagation paths in the radio environment. We define $\alpha_k(t)$ and $\tau_k(t)$ as the complex attenuation factor and propagation delay for the path k respectively, and there are K number of significant propagation paths. Let $e^{j\epsilon(f,t)}$ be the timing alignment offset, sampling frequency offset, and carrier frequency offset due to the lack of synchronisation between transceivers and hardware imperfections. Then, we can derive $H(f,t)$ as follows:

$$H(f,t) = \Big(\sum_{k=1}^{K} \alpha_k(t) e^{-j2\pi f \tau_k(t)}\Big)\, e^{j\epsilon(f,t).} \tag{4}$$

Frequency offset causes the random phase shifts between measured CSI and needs to be sanitised before the signal processing pipeline. It's assumed that frequency offsets are equal in all the antennas in the same NIC card, the conjugate multiplication of CSI measurements between two antennas of the same NIC will eliminate the random phase offsets. We select two antennas (out of three available in the Intel 5300) in the receiver (60 CSI streams) to perform the conjugate multiplication. Then the variance of the CSI amplitudes of all the antennas are compared and the two with the highest variance are selected.

The resulting data structure after the conjugate multiplication is then filtered out with a bandpass filter (2 - 60 Hz) to remove the impulses and burst noises caused by the environment. Additionally, the filtering will remove the potential noise from the product of antennas dynamic responses ("Cross Terms" of Eq(6) in [15]). Principle Component Analysis (PCA) [19] is applied to further de-noise and reduce the dimensionality after the above step. The first principle component is selected as it contains the major and consistent power variations induced by human motions.

Finally, STFT (Short Time Fourier Transform) [20] is applied to the first principle component to generate a time-frequency spectrogram. Figure 1(a) shows a spectrogram generated by the pipeline outlined above for leg swing motion. This is 1D matrix (single channel) where *x-axis* represents the 1 sec of CSI data captured during the data collection time period after applying a segmentation algorithm to extract start-end of the activity. Here, we use 121 frequency bins (-60 to 60 Hz) to capture frequencies and the colours represent the energy in dBs (normalised to 0-1 range) in y-axis.

Figure 15 in the appendix provides an illustration of the flowchart representing the process.

D. System Model

Figure 3 illustrates the proposed adversarial based neural network architecture of `WiFi2Radar`, integrated with an application (i.e., activity recognition). Here, we use a U-Net neural network model as the `WiFi2Radar` Translator to improve the quality of the Doppler information generated by a COTS WiFi receiver. Furthermore, a domain (i.e., orientation) discriminator is introduced to extract orientation invariant Doppler information from the WiFi measurements. Finally, the U-Net translation and domain discrimination components in the proposed architecture seamlessly integrate with an example application (i.e., human activity recognition) so that our U-Net feature extractor can retain the features that are important for the application.

In the following sections, we will discuss each component in detail where we specifically evaluate the U-Net based `WiFi2Radar` Translator Module to validate `WiFi2Radar` translation is possible and motivate the application. In Section III, we further discuss `WiFi2Radar`'s design choices.

E. Deep Learning Architecture

As discussed earlier, [17] proposed to use a U-Net [21] style model to denoise and enhance the quality of low-light images. During such low-light conditions, the number of photons is so small that the Signal to Noise Ratio (SNR) of an image becomes too low to be visible. Similarly, the SNR of the Doppler spectrograms produced by the WiFi CSI measurements is very low. Therefore, we aim to use a similar U-Net model to improve the quality of the Doppler spectrograms.

a) U-Net as `WiFi2Radar` Translator: A U-Net [21] is fundamentally an autoencoder-type neural network, featuring a paired encoder (contracting path) and decoder (expansive path). The inherent symmetry between these paths gives rise to its distinctive U-shaped architecture. In our U-Net model, the contracting path consists of multiple blocks, each comprising two convolutional neural network (CNN) layers, accompanied by a Rectified Linear Unit (ReLU) activation function and subsequently a Max Pooling layer. A consistent dropout rate of 0.1 is applied across all blocks, contributing to regularisation. The contracting path has a depth of 5 blocks.

The bottleneck block in our U-Net model consists of two CNN layers. A Block in the expansive path contains two CNN layers followed by a transpose convolution layer. The depth is the same as that of the contracting path (i.e., five blocks). More importantly, the corresponding blocks in contracting and expansive paths have cross-connections (i.e., copy and crop) that share the learnt parameters. The dimensions of the input and output of our U-Net model are identical, i.e., 512 x 512 resolution images (i.e., Doppler spectrograms)

Our U-Net model-based CSI to precise Doppler translator takes a CSI spectrogram image produced by WiFi as the input

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3349378

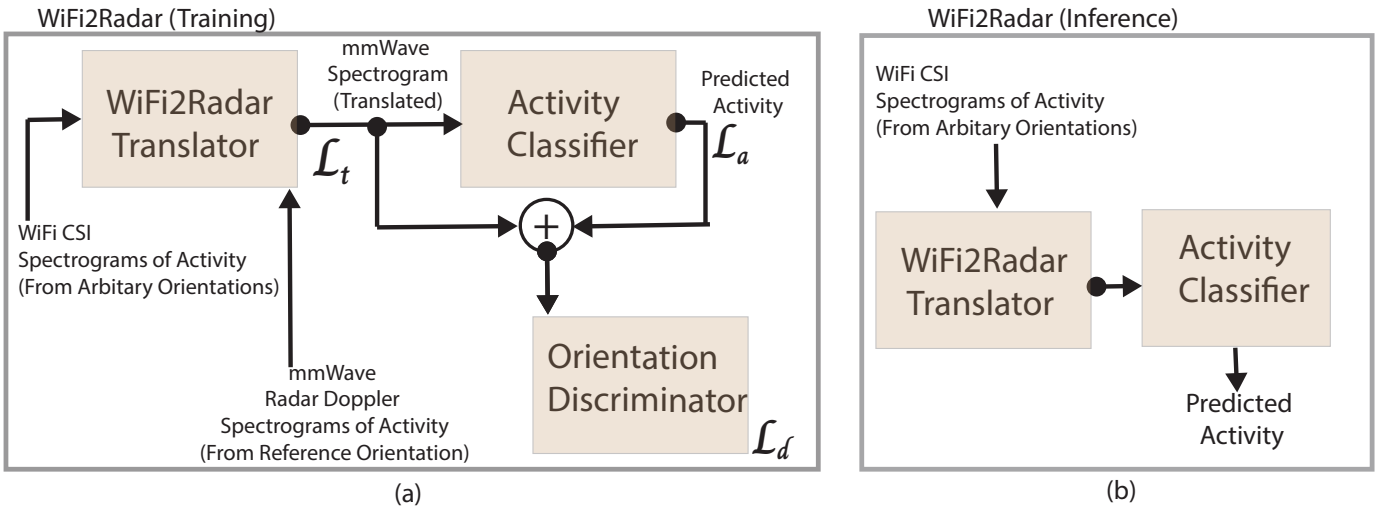


Fig. 3: The proposed deep learning architecture, which features a `WiFi2Radar` translator, an orientation discriminator to reduce orientation-dependence, and an application specific model (i.e., human activity classifier). These three components are seamlessly integrated by combining their respective loss functions (i.e., $\mathcal{L}_t$, $\mathcal{L}_d$ and $\mathcal{L}_\alpha$) together during each learning iteration. Millimeter wave radar data is used only during (a) training , while (b) inference is done based on WiFi signal alone.

and a precise reference Doppler spectrogram as the output during the training time. The translated CSI spectrogram will be used as the input (features) for the other two components in the proposed architecture.

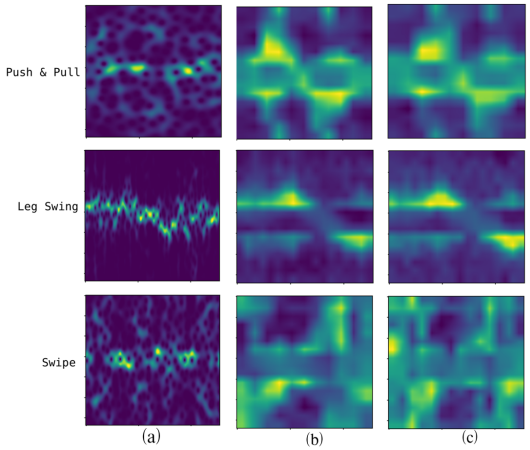


Fig. 4: Samples of U-Net performance: (a) WiFi CSI spectrograms of different activities as input to the U-Net (b) corresponding mmWave radar spectrograms (c) U-Net translated spectrograms

Performance of the U-Net translator: To evaluate the performance of the U-Net translator, we have used our primary dataset consisting of three activities, Push & Pull, Leg Swing back and forth, and Swipe, collected from three different environments involving two subjects. We use 70% of the data to train the U-Net model and the remaining 30% for testing. Figure 4 shows some samples of generated spectrograms using a trained U-Net model. Here, the rows represent different activities, and the columns represent the spectrograms generated by WiFi, ground truth and the proposed U-Net model, respectively. The figure shows that the spectrograms produced by the WiFi CSI (the first column in the figure) directly are indeed very noisy (i.e., low SNR) similar to the images in low light conditions in [17]. This is the direct result of CSI after the noise reduction techniques we discussed in Section II-C. But those (the third column in the figure) produced by the U-Net component of the proposed architecture in Figure 3 are smooth and similar to the ground truth (the second column in the figure). We have also measured the average similarity between ground truth and generated spectrograms by the U-Net with the Structural Similarity Index (SSIM) [22] (see Section III for more details). Our results show that the average performance of the proposed U-Net model in all test spectrograms is 0.87 (standard deviation or std: 0.13), 0.88 (std: 0.35) and 0.77 (std: 0.36) for the three different activities, respectively, which are significantly better than those of WiFi spectrograms, e.g., 0.02 (std: 0.007), 0.05 (std: 0.08), and 0.03 (std: 0.09) respectively.

b) Performance of U-Net with unseen orientations: We continued our experiment by translating the WiFi spectrograms of an unseen orientation during the training time. Unsurprisingly, this resulted in a drastic drop of SSIM performance of 0.19 (std: 0.06), 0.23 (std: 0.01) and 0.15 (std: 0.05) for the three different activities respectively. This observation had led us to remedy the performance of U-Net as discussed in next sections.

c) Domain Discriminator: Orientation Invariant Doppler: As previously mentioned, Doppler measurements exhibit orientation dependence, with receivers in different orientations yielding distinct Doppler frequency measurements for the same motion due to the directional nature of speed and motion. Consequently, the U-Net component introduced earlier does not inherently produce orientation-independent Doppler measurements. Training the U-Net with observations from each orientation would be labor-intensive. To address this, our architecture incorporates domain adaptation into the U-Net component, inspired by techniques outlined in [23]. Notably, our model deviates from [23] by eliminating the need for unlabeled data from target ("unseen") domains,

enhancing user-friendliness. This capability is facilitated by the Doppler information translation facilitated by the U-Net component.

To achieve orientation-invariant Doppler extraction, we introduce a deep learning component to the earlier U-Net, tasked with classifying the specific orientation (i.e., degrees relative to the reference direction) of the performed motion. Since our objective is not to retain features distinguishing orientations, we seek to maximise incorrect orientation classifications. This goal is realised through the incorporation of a domain discriminator, following the design proposed in [23]. The **domain discriminator** component includes a fully connected layer with a subsequent softmax activation function. The input for the domain discriminator comprises concatenated features from translated spectrograms and predicted activity labels. During training, a Gradient Reversal Layer (GRL) [23] is employed in the backward pass, enforcing maximisation of the domain discrimination loss.

d) Application-Specific Component: Activity Classifier: The previously established domain (orientation) invariant Doppler features form a robust foundation for various WiFi-based radio sensing applications, notably human activity recognition. A common approach involves a sequential setup, incorporating an additional neural network classifier (e.g., activity classifier) that utilises the orientation-invariant precise Doppler spectrogram as input to yield activity classifications. However, this straightforward approach may inadvertently lead to the loss of application-specific features crucial for activity recognition during Doppler spectrogram generation. To address this, our architecture integrates the activity recognition component directly, as illustrated in Figure 3.

Our activity classifier adopts a hybrid architecture, combining Convolutional (4 layers) and Gated Recurrent Unit (GRU) layers (2 layers) to extract both spatial and temporal features from the translated Doppler spectrograms for activity classification. A dropout layer with a 0.4 dropout rate is strategically placed between the 3rd and 4th Convolutional layer, as well as after the GRU layers. The default activation function for the neural layers in the classifier is the Rectified Linear Unit (ReLU) [24]. Lastly, a fully connected layer precedes the generation of activity probabilities through the Softmax activation function [24].

Our results, discussed in Section III, demonstrate that our proposed approach effectively preserves application-specific features, outperforming the naive (sequential) approach with a notable increase in activity classification accuracy (80% vs. 67%). Notably, the Signal-to-Noise Ratio (SSIM) of the spectrograms produced by our approach (0.42) is lower than that of the naive approach (0.77). This discrepancy underscores that our proposed approach sacrifices Doppler spectrogram quality to enhance application accuracy by retaining essential features while discarding some elements critical for spectrogram quality.

F. Model Constraints

Figure 3 shows that three different constraints (i.e., $\mathcal{L}_t$, $\mathcal{L}_d$ and $\mathcal{L}_a$) are introduced in our model to enforce the goals of different components discussed earlier. In this section, we will define them formally.

Specification, the translation loss (L_t) is a binary cross-entropy:

$$\mathcal{L}_t = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(\hat{y_i}) + (1-y_i)\log(1-\hat{y_i})), \quad (5)$$

where $\hat{y}$ and y represent the pixel values in the predicted Doppler spectrogram produced by the U-Net component in our architecture and those in the ground truth spectrograms respectively. N is the number of spectrograms. We normalise the values in y and $\hat{y}$ to a range between [0-1] during the preprocessing stages by following [25].

The domain (orientation) classification (discrimination) loss (L_d) is a categorical cross-entropy function:

$$\mathcal{L}_d = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M} y_{im}\log(\hat{y}_{im}), \quad (6)$$

where y_{im} and $\hat{y}_{im}$ are the predicted and ground truth orientation labels respectively. M is number of orientation class labels.

Similarly, the activity classification loss (L_a) in our model is also a categorical cross entropy function:

$$\mathcal{L}_a = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log(\hat{y}_{ic}), \quad (7)$$

where C is the number of activity class labels. y_{ic} and $\hat{y}_{ic}$ are the predicted and ground truth activity class labels respectively.

We note the - operation in Eq. (7) but not in Eq. (6) because the aim of our orientation classification is to maximise the incorrect orientations and discard those features that can classify an orientation correctly (i.e., orientation or domain dependent) as discussed earlier.

Finally, the overall loss function is depicted below.

$$\mathcal{L} = \alpha\mathcal{L}_t + \beta\mathcal{L}_a + \lambda\mathcal{L}_d, \quad (8)$$

where λ is the gradient reverse layer introduced in [23], and α and β consecutively defines the coefficients, which are empirically set to 1. We use Adam [26] optimiser with learning rate $1e{-}6$ for the model training.

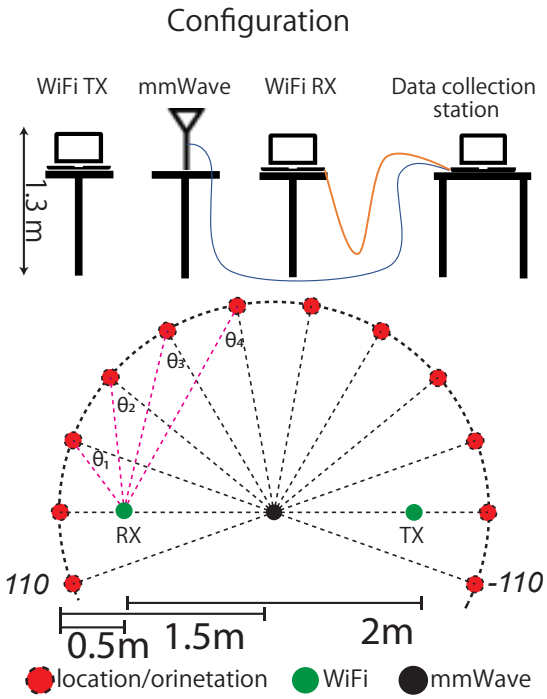
Appendix 13 portrays the overall process of the system model.

III. Evaluation

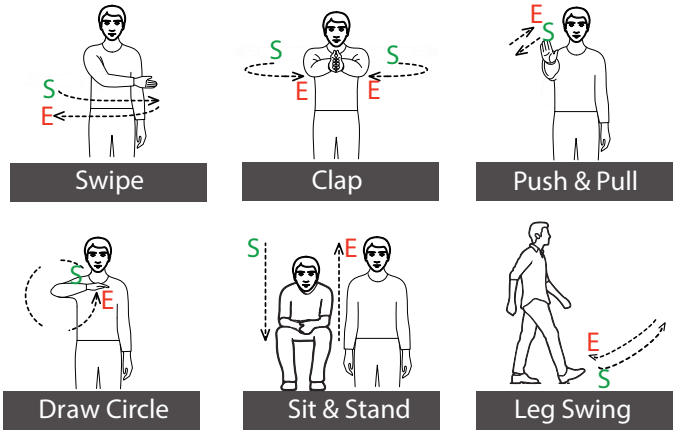
In this section, we evaluate the proposed `WiFi2Radar` for human activity recognition using real subjects for unseen orientations (domains), i.e., **the orientation it is tested on is not used in training**. We benchmark `WiFi2Radar` against the state-of-the-art approach, EI [9], which also uses adversarial deep learning based on the [23] architecture but without the WiFi to radar translation. Because EI was designed with specific algorithms to take advantage of any unlabelled data that may be available from the target domain, we employ two different benchmarks for it. The first EI benchmark assumes 10% unlabelled data from the target domain, while the second assumes truly unseen domain, i.e., 0% data from the target orientation. The second EI benchmark is equivalent to our

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3349378

(a) Data Collection Environments



(b) Experiment Layout



(c) Activity Set

Fig. 5: Data collection environments, experiment layout, and the activity set considered in the evaluations.

proposed `WiFi2Radar` approach, which also does not use any data from the target domain. EI [9] follows similar preprocessing pipeline as `WiFi2Radar` where the Frequency Domain features are generated from the CSI data series as a input to the deep model. We do aware that the modern mmwave based sensing solutions including Pantomime [27], M-gesture [28] where the mmwave being the main sensing modality and supporting up to 21 gestures. `WiFi2Radar` uses mmwave as a training aid only and despite the popularity of the mmwave radars, still the commercial products are not available in the consumer electronics and not deployed widely compared to WiFi. Furthermore, contemporary state-of-the-art techniques such as RF-net [29] take a different approach by circumventing signal translation methods, with a primary focus on reducing data labelling requirements and adaptability to diverse environments. While this approach complements our work, given that RF data collection can be resource-intensive and model-dependent, our research predominantly relies on WiFi-to-Radar translation. This choice motivated us to benchmark `WiFi2Radar` against a prominent WiFi-based gesture sensing study featured in the existing literature.

For `WiFi2Radar`, we also explore its,

- Generalisation capability to unseen subjects and radio environments.
- Robustness against the WiFi Tx-Rx deployment layout.
- Performance with different deep translation networks, namely a generic autoencoder and U-net.

Finally, we evaluate the scenario where the activity classifier is both trained and tested with radar data, which provided an accuracy of 97% and could be considered as the 'upper limit' for `WiFi2Radar` translation approach.

A. *Experimental Setup and Data Collection*

Figure 5b illustrates the dataset we collected firsthand. During our experimental setup, we conducted data collection concurrently by capturing both WiFi and radar data while the subjects were engaged in various activities across 12 different

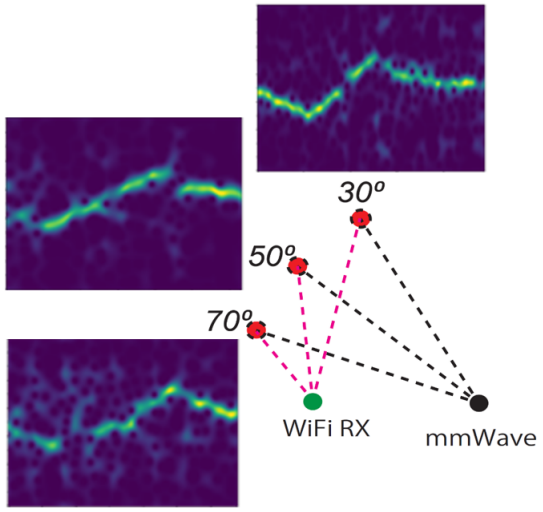


Fig. 6: Variation of WiFi Doppler with the orientation

orientations. For the primary data collection, we positioned two WiFi-enabled laptops (one serving as the WiFi transmitter and the other as the receiver) at a distance of 2 meters from each other. Between these laptops, we mounted a TI IWR1443 millimeter-wave FMCW (Frequency Modulated Carrier Wave) radar, which operated in the 77-81GHz frequency band.

Both laptops were equipped with Intel 5300 wireless Network Interface Cards (NICs) and were configured to operate in the IEEE 802.11n 5 GHz band, specifically channel 100, with a 5,500 MHz centre frequency and 40 MHz bandwidth. While the transmitter laptop (Tx) employed a single antenna, the receiver laptop (Rx) used three antennas. The Tx laptop transmitted data packets at a rate of 200 packets per second, and the Rx laptop was configured with a CSI capture tool [18] to collect Channel State Information (CSI) data from the received packets. We ensured synchronisation between the Rx laptop and the mmWave radar using the Network Time Protocol (NTP).

The subjects performed activities from 12 different locations, so the orientations of adjacent locations from the subject to the middle point of the two laptops were separated by 20° . Data collection points were located 1.5m to 2m away from the ground truth collection device (i.e., mmwave radar). The radar was always facing the subject to capture the human motion within its field of view. Although subject's orientaiton is always perpendicular to the ground truth collection device's boresight direction, the angle between the WiFi receiver device, and the subjects' orientation (θ) is always different which introduces different radial velocities when the motion occurs. Figure 6 further visualises the variation of the doppler when Leg Swing activity is performed in selected orientation.

Therefore, the distance between a subject and one of the WiFi transceivers was between 0.5m to 2.5m. The subjects performed the same activity (among Hand Push & Pull (P&P), Leg Swing back and forth (LS), Hand Swipe (HS), Clap, Draw a Circle vertically and Stand & Sit (S&S)) repeatedly for up to 5 minutes in one session with an approximately 2-second interval between each repetition. We have selected the set of gestures by first analysing the state-of-the-art [8], [9], [27], [30] and aligning them with following consideration.

One of the reason being we specifically selected the set of gestures to test was that in smart home applications like person interacting with an appliance such as TV or a smart home automation device. Smart devices are inherently built in WiFi sensors therefore can be reused for sensing as well. In these use cases, the user is usually facing towards to the appliance in order to see the response from the device but doesn't usually have to be in the boresight angle and in a specific distance as well. However, we have noticed the possible orientation/position combinations are usually circular to the appliance. Smart contactless screens are also popular even in public places as it's easy to interact and safe in pandemic situations. the screen can be activated when a person approaches and swipe ,draw circle and clap would interact with it.

Second reason is that the gestures should cover wide range body movements. For example, Push and Pull using the dominant hand is a regular activity with horizontal arm movement, while Swipe is a vertical arm movement. We incorporated Leg moving back and forth to cover the lower body movements. Repeated Leg moving back and forth may be considered as walking. Additionally, Clap, Draw a Circle and Sit/Stand activities are performed in one of our environments (R3) to incorporate duel hand motion, circular motion and full-body motion types. Thus, we believe the choice of our activity covers a wide variety of body motions that a human perform in front of a smart appliances. Please note that we do use mmwave sensor for ground truth collection only and used for training our deep model.We recruited four subjects with varying physical attributes, both female and male with the average weight of 62.5 kg, the average height of 167.8cm and the average age of 30.5 years.[1]

To obtain Doppler ground truth information to train our model, we leveraged a commercially available mmWave FMCW [31](Frequency Modulated Carrier Wave) radar [1] operating on 77 GHz to 80 GHz frequency. Every 125 ms, the FMCW radar transmitted one frame containing radio chirps over a wide bandwidth of 4 GHz, which gives the radar a resolution of 3.75 cm, i.e., it can differentiate reflections that are bounced back from different parts of the human body separated by as little as 3.75cm. Therefore, the radar was able to produce a 1D energy vector (the closer the distance, the higher the energy) of the reflectors, which is called range-Fast Fourier Transform (range-FFT). By leveraging the phase difference of the received neighbouring chirps, we computed the radial velocity of the reflector in a specific distance. By stacking the range-FFTs of multiple chirps, we can calculate the fine-grain Doppler information of the human body motion via a second FFT across different ranges (i.e., the columns in the stacked range vectors or FFTs). The output of the second FFT was a 2D Range Doppler Matrix (RDM). A row-wise averaging is performed on RDMs to extract the Doppler

[1]Human data collection is approved by UNSW Human Ethics HC17823.

information that matches the range-less Doppler captured by WiFi only. Thus, at each time step, an RDM is collapsed into a 1D Doppler array. Finally, we stitched the consecutive 1D Doppler vectors to obtain a Doppler matrix (2D spectrogram), which can be used as the input of `WiFi2Radar` neural network model using the method in Section II-C during training. Figure 1(b) illustrates an 2D Range-Doppler Spectrogram when the range collapse is applied (thus, the x-axis becomes the time dimension). We note that in the FMCW literature and products, values 2D Range-Doppler matrix is often reported in meters and meters/sec since the Doppler has the dimensions of velocity. Therefore, we follow the same terminology when presenting the ground truth with the units of measurements in this paper. The Figure 15 in the appendix presents the flowchart depicting the process.

We collect data from three different radio environments or rooms, open office area (R1), seminar room (R2), and computer lab (R3), as shown in Figure 5a.While the data collection setup and geometry remain the same for all environments, the rooms themselves are different in terms of their size, layouts, furniture, and background activities. Daytime data collection in R1 continued while other office occupants were present in the open space, which meant increased WiFi usage in the area as well as uncontrolled background movement of other people in the background. In contrast, no human was present in the other two rooms but their sizes were smaller than R1. To capture realistic patterns of changing ambient conditions, we spread the data collection over several days and different time slots within the day, i.e., work hours (9:00-18:00) vs. after hours (18:00- 24:00). Additionally, we did not exhaust the data collection for all orientations in one day but rather split it over 3-5 days. Moreover, a followup data collection round specific to unseen subjects (Sec.III-D) and varying transceiver locations (Sec.III-E) was performed 3 months later.[2].

One subject performed all 6 activities in R3, but only three activities, push-pull, leg-swing, and swipe, in the other two rooms. A second subject performed three activities, push-pull, leg-swing, and swipe, only in R1. For each combination of orientation and activity, a subject repeated the activity for 5 minutes within one session with approximately 2-sec gap between the repetitions. Each subject completed two 5-minute sessions, thus 10 minutes of data were collected from each subject for a given orientation-activity combination in each room. We further incorporated the testing data from two extra subjects, who performed three activities (i.e., push-pull,leg-swing,and swipe) in 2 orientations with varying WiFi transceiver setups, i.e., Tx-Rx distance between 1m and 3m.

Therefore, our final data set contains 1,800 minutes of human motion data, which contains approximately 32.4 million WiFi CSI samples and 1.3 million radar frames. After segmentation, we collected a total of 27 041 activity spectrograms for both WiFi and radar to be used for deep learning. [3]

[2]During this period, other uncontrollable factors (e.g. change in chairs/furniture locations) probably altered the original wireless environment. This presents a more realistic domain shift that we leverage to test the robustness of `WiFi2Radar`

[3]We plan to release the data set as a publicly available repository along with the publication.

TABLE I: `WiFi2Radar`

Architecture	Performance	
	Accuracy	Avg.SSIM
$WiFi2Radar$	80.53	0.42
$WiFi2Radar_{pre-training}$	67.29	0.77
$WiFi2Radar_{No\ U-Net}$	60.24	0.12
$WiFi2Radar_{No\ Translation}$	58.41	N/A

B. Training and WiFi2Radar

In Section II, we have already discussed the performance of the U-Net in achieving the precise Doppler translation from WiFi and its limitations on working with unseen domains. In this section, we introduce each component to U-Net with the performance evaluation while justifying the architecture and performance improvement. For clarity, we include the proposed architecture we derive at the end and the ablation architectures in Figure 7 (a). Table I summarises the accuracy ($accuracy = \frac{TP+TN}{TP+TN+FP+FN}$) and SSIM difference between the ground truth and translated spectrograms based on the R1 dataset. `WiFi2Radar` archived 80.53% accuracy in this experiment. We further explain the approach of the ablation study for building `WiFi2Radar` as follows.

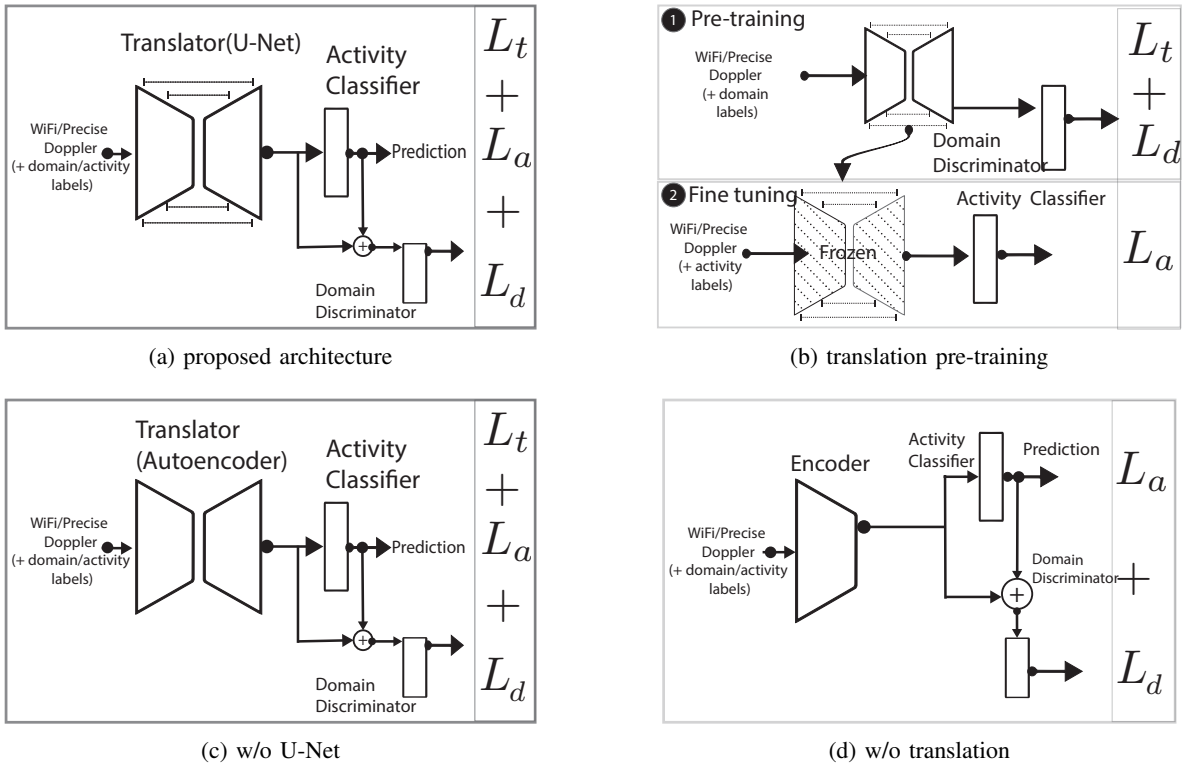
In Figure 7 (b), we streamline the training strategy into a two-player adversarial game, wherein the translation module undergoes pre-training in conjunction with the domain discriminator. Subsequently, during the fine-tuning stage, the frozen translator module's output is utilized. This particular investigation yielded an accuracy of 67% when subjected to leave-one-out domain testing across 12 domains. A detailed analysis of the visual quality of the generated mmWave spectrograms is presented in Figure 8 (Row 1). Unfortunately, this approach resulted in a loss of distinguishable features from activities, leading to a decline in classification performance from 80.53% to 67%.

The two-player pre-training stage is characterised by a dual objective: generating orientation-invariant Doppler and remaining unaware of subsequent classification goals. Consequently, the classification task lacks features crucial for distinguishing between different activities, thereby contributing to the observed reduction in classification accuracy.

While skip connections enhance the translation capabilities of U-Net by preserving essential residual information from the encoder, they simultaneously introduce heightened computational complexity and extended training times. To better understand this trade-off, we explore the use of conventional autoencoders, which have proven effective in similar translation tasks [33], as an alternative to U-Net without incorporating skip connections between the encoder and decoder components.

In our study using the R1 dataset, we replaced the U-Net in `WiFi2Radar` with a conventional autoencoder. This autoencoder comprises a standard 4-layer convolutional neural network (CNN) — featuring 2 CNN layers for the encoder and 2 transposed convolutional layers for the decoder. In leave-one-out evaluations across 12 orientations, we achieved an average accuracy of 60.24% with the conventional autoencoder, absent of skip connections. As an illustrative example,

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3349378



(a) proposed architecture

(b) translation pre-training

(c) w/o U-Net

(d) w/o translation

Fig. 7: Various ablations were considered to evaluate (a) the proposed model including (b) pre-training the translation module prior to gesture classification (c) replacing U-Net with a conventional autoencoder.(d) removing the translation component. **GRL** denotes Gradient Reverse Layer [32].

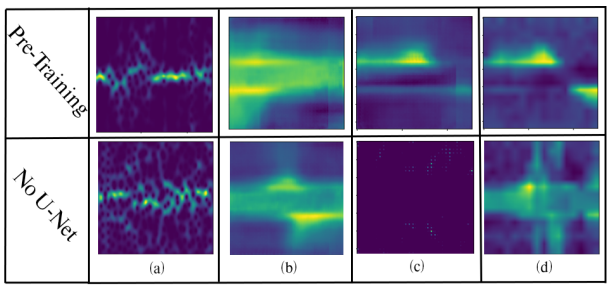


Fig. 8: WiFi Spectrogram (a), Spectrogram generated from Proposed Architecture (b), Comparison Architecture (c) and Ground Truth (d)

focusing on the swipe activity, our results demonstrate that the U-Net generates a more accurately translated image compared to the autoencoder, as depicted in Figure 8 (Row 2).

These experiments further confirmed the pivotal role of skip connections in realising the envisioned capabilities of `WiFi2Radar`. While the incorporation of skip connections introduces additional complexity, it is noteworthy that the training process can be conducted offline in the cloud. As such, the supplementary complexity due to skip connections is not anticipated to pose a bottleneck for the widespread deployment of `WiFi2Radar`.

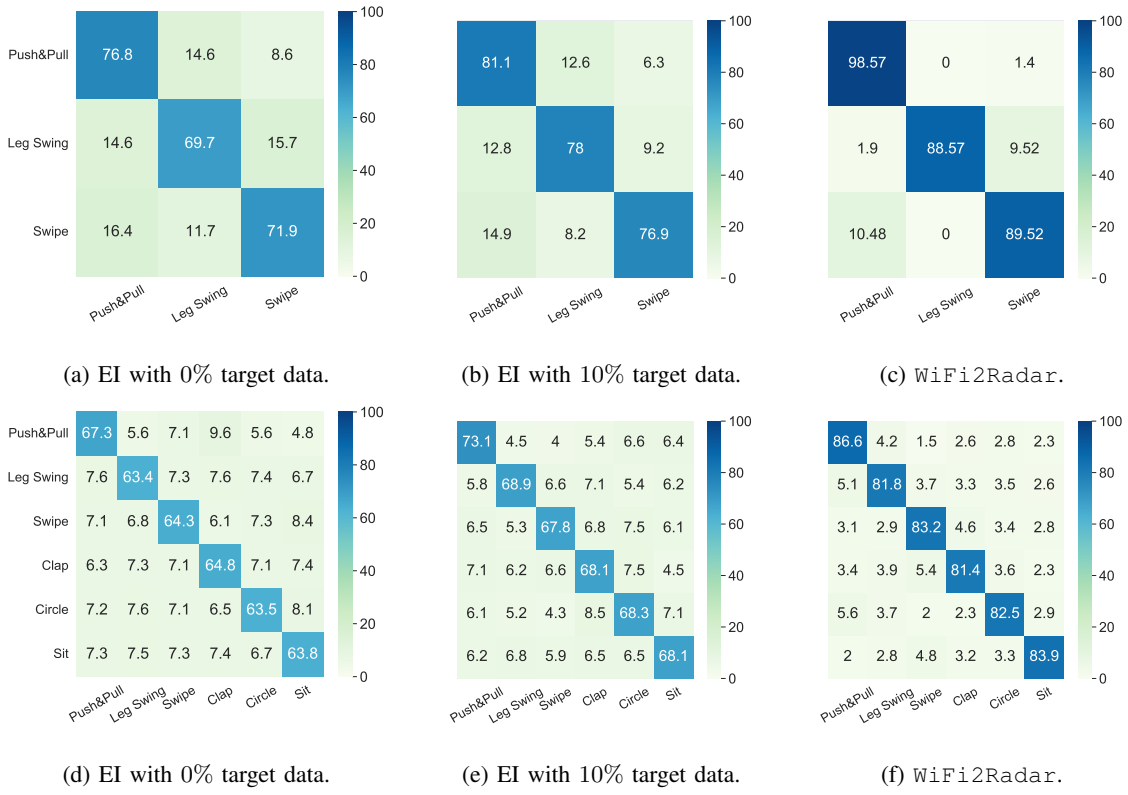
In Figure 7 (d), we undertake an experiment where we eliminate the translation module entirely to assess the impact of mmWave's contribution to our model. In this configuration, we discard the U-Net-based translation module and instead employ the generic Domain Adversarial Neural Network (DANN) architecture without unlabeled data from the unseen domain. Utilising WiFi Doppler Spectrograms as inputs along with their corresponding labels, we aim to extract features.

Upon subjecting this model to leave-one-out domain testing across 12 domains, we achieve an average accuracy of approximately 58%. This outcome affirms that the incorporation of mmWave not only successfully recovers Doppler information but also enhances overall model performance.

C. Benchmarking against EI

Using the 3-activity datasets from three rooms and two subjects, Figure 10 compares the performance of `WiFi2Radar` against the two versions of EI, i.e., 0% (henceforth referred to as EI-0) and 10% (EI-10) unlabelled data used from target domain based on leave-N-out evaluation, where N orientations are used for testing and $12 - N$ for training. Note that `WiFi2Radar` never uses any data from the testing domain and hence EI with 0% is a more fair benchmark for it. As expected, we find that the accuracy of all three models increases with the number of domains (orientations) used in the training. However, `WiFi2Radar` outperforms the EI benchmarks significantly in each configuration. For example, when the number of domains used in the training is 11, the average accuracy of `WiFi2Radar` is 91.5%, which is 18.7% higher than E1 0% benchmark. Even with the unfair advantage

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3349378



(a) EI with 0% target data. (b) EI with 10% target data. (c) WiFi2Radar.

(d) EI with 0% target data. (e) EI with 10% target data. (f) WiFi2Radar.

Fig. 9: Confusion matrices of WiFi2Radar and EI. (a)-(c) for 3 activities from all rooms and (d)-(f) for 6 activities from R3.

TABLE II: WiFi2Radar vs. EI for 3 and 6 activity classes

No. of Activity classes	EI-0		EI-10		WiFi2Radar		Gain against EI-0
	ACC	F1	ACC	F1	ACC	F1	ACC/F1
3 (P&P, LS, Swipe)	72.8±3.6	77.7±0.03	78.6±2.3%	81.2.5±0.02	92.2±5.5%	92.2±0.06	+19.4/+14.5
6 (P&LS, Swipe, Clap, DC, S&S)	64.5±2.3%	71.7±0.02	68.3±0.5	72.9±0.02	83.2±1.9 %	86.9±0.02	+18.7/+15.2

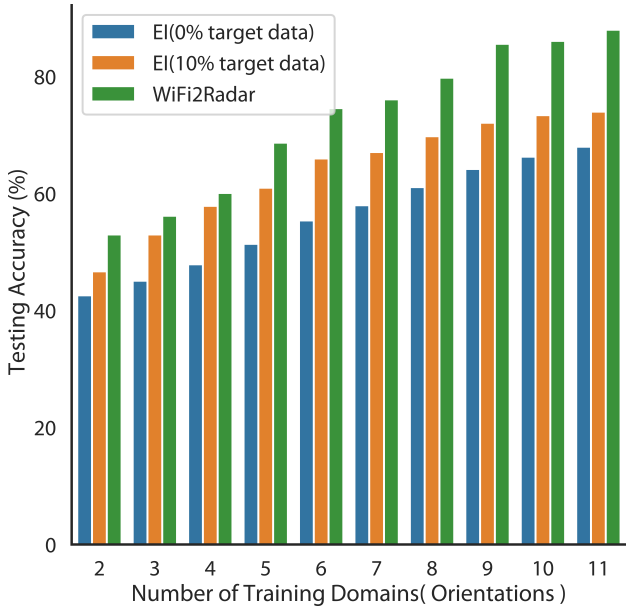


Fig. 10: Accuracy comparison for WiFi2Radar vs. EI.

of 10% unlabelled target domain data given to EI, the proposed WiFi2Radar outperforms EI by 12.8%.

We have introduced F1 score as another performance metric for this experiment.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

$$F1\ \text{Score} = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}} \tag{11}$$

Similar performance gains are also achieved for F1 scores (Information on how we calculated the F1 scores are given in Eq.9, Eq.10, and Eq.11). Table II shows the performance of WiFi2Radar against the EI-0 and EI-10 also with different levels of cardinality.

To further investigate the potential superiority of WiFi2Radar over EI when scaling up the number of activity classes, Table II presents a performance comparison using the 3- and 6-activity datasets while 11 orientations are used for training. As expected, accuracy experiences

a decrease across all models as the number of gestures increases. However, it is noteworthy that WiFi2Radar maintains an approximate 18% accuracy advantage over the EI 0% benchmark, regardless of the number of gestures. In terms of F1 score, WiFi2Radar also sustains an approximate 15% lead over the EI 0% benchmark, regardless of the number of gestures.

The Sit activity is a full body movement, which we anticipated a better performance than the rest of the activities that involve parts of the body only. However, Push & Pull activity still performs the best in the activity wise classification because of the reflected signal path change rate is higher than the Sit activity as we discussed in Section II-A earlier.

D. Robustness against unseen environments and subjects

Our main contribution is to produce a WiFi to Radar Doppler feature translation-based approach before further extending it to an orientation independent activity recognition application. Therefore, our evaluation mainly focused on orientation invariant activity recognition. However, real-world applications involve deploying the motion-sensing systems in many different radio environments interacting with people with various physical characteristics. In order to demonstrate that the Doppler features used in WiFi2Radar are robust to different radio environments and people, we further evaluated our model's performance with different "seen" and "unseen" radio environments and participants.

Tables III and IV show the performance of WiFi2Radar across different rooms and subjects, respectively. We used 11 orientations from two rooms for training and the 12th orientation from the third room for testing for cross-room experiments. Only data from a single subject (S1) for three activities were used for these experiments.

Similarly, for cross-subject experiments, two subjects' data from only R1 was used to train the model, where 11 orientations from one subject were used for training, and the 12th leave-out orientation is used for the testing "seen" subject performance. Finally, we tested the same with three different subjects' data using "unseen subject". Here, we used leave-out orientation from unseen subjects during the test time.

Tables III and IV show that the overall accuracy of WiFi2Radar drops by approximately 3% only when tested across different environments and subjects, despite being trained with a small number of rooms (2 in this case) and subjects (only one in this case). Thus, the proposed WiFi to radar translation for human activity recognition performs well across different orientations, and different environments and subjects.

E. Varying Tx-Rx Distances

In real-world scenarios, WiFi transceivers are dispersed in the deployment environment so that the distance between them varies. With the increase of propagation distance, the signal to noise ratio of WiFi signal degrades, thus the features extracted from the received signals may be less distinguishable. Although our contribution is an orientation independent WiFi sensing model, we further tested the robustness of WiFi2Radar with different distances between a Tx and a Rx. To this end, we conducted an experiment with different distances between Tx-Rx to 1m, 2m and 3m in the R1 environment where subject one performed the activities. The ground truth collection radar and subject was maintained at 1.5 m distance. Table V shows the performance of both in-dataset and cross-dataset. For the seen dataset performances, our model predicts the activity in a similar performance regardless of the training dataset.

We've observed a slight (2%) performance decrease compared to the in-dataset results, indicating our model's resilience to varying transmitter-receiver (Tx-Rx) distances during deployment. A closer look reveals that accuracy drops more significantly with the distance, unseen datasets, which is reasonable given the reduced signal-to-noise ratio (SNR) resulting in noisy WiFi Doppler images. This observation is consistent with testing our 2m + 3m trained model on a closer dataset (1m), where performance dropped by 1.5%.

Training on both near and far datasets and testing on mid-distance data showed only a modest 0.5% performance reduction. This result can be attributed to the wider range of Doppler samples with varying SNR, enhancing the model's ability to learn meaningful representations.

TABLE III: Cross Room Accuracy

Train Datasets	**Test Dataset (Unseen)**	**Seen Env.**	**Unseen Env.**
R1+R2	R3	91.6%	89.4%
R2+R3	R1	92.8%	88.5%
R1+R3	R2	91%	87.9%
Avg.		**91.8%**	**88.6%**

TABLE IV: Cross Subject Accuracy

	Seen Sub.	**Unseen Sub. 1**	**Unseen Sub. 2**	**Unseen Sub. 3**	**Unseen Sub. 4**
S1	92.6 %	-	88.3%	89%	89.2%
S2	91.2	87.4%	-	87.8%	88.5%
Avg.	**91.9%**	**87.4%**	**88.3%**	**88.4%**	**88.9%**

F. Analysing the embeddings of the Intermediate Layers

In order to further explore how our model works, we have used a popular method for visualising and analysing the learned representation of the intermediate layers of the deep model. We aim to explore the model's learning capability with the constraints we have introduced to train the model. As discussed in the section, we are forcing the overall network to predict the activity labels but penalising the domain prediction

TABLE V: Varying Tx-Rx Distances

Train Datasets	**Test Dataset (Unseen)**	**Seen Env.**	**Unseen Env.**
1m+2m	3m	92.5%	88.6%
2m+3m	1m	89.3%	87.8%
1m+3m	2m	89%	88.5%
Avg.		**90.3%**	**88.3%**

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3349378

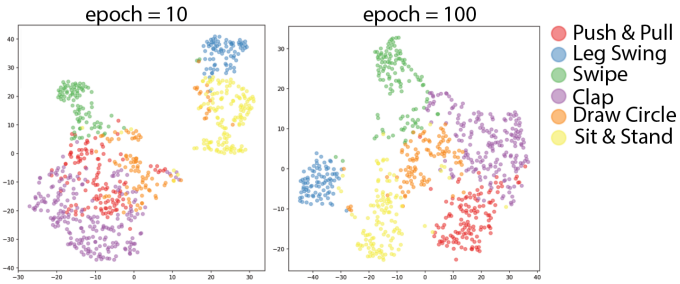


Fig. 11: Embeddings of the Classifier component

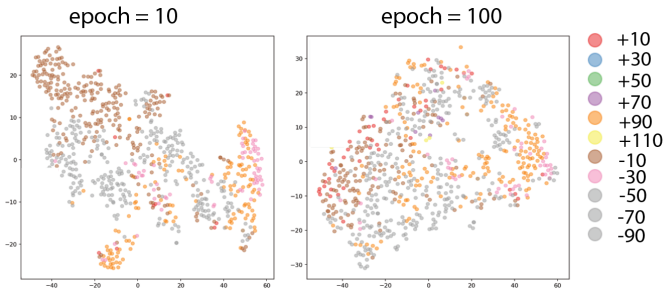


Fig. 12: Embeddings of the Discriminator component

(orientation labels). As a result, however, we only see the end probability output of the network using accuracy as a metric but not how the network learned.

Deep models' layer activations are multidimensional and sparse, beyond the human capacity to interpret and identify the patterns. Therefore, we are using a well-known statistical analysis method called t-SNE [34] to convert the higher dimension learned representation into two dimension space while preserving the learned features where it is easier to interpret and analyse. This is often referred to as embeddings of the deep neural network learnings.

In our experiment, we analyse the last fully connected layers of embeddings of our network's discriminator and classifier components. We project the embeddings into 2D space and assign the labels to analyse the clustering of the embeddings in epochs 10 and 100.

Figure 11 shows the embeddings for epochs ten and 100, where the colours correspond to the activity label. Initially, two main clusters exist in the embeddings, and intra-clusters have overlaps in the boundaries. For example, the model created one cluster for leg swing and sit&stand activities. However, once the model has been trained, we can see a clear boundary between leg swing and sit&stand activities in the embeddings. The rest of the activities are clustered together. However, the overlap between the rest of the activities is minimal (especially between the Push&Pull and clap activities).

Figure 12 illustrates the embeddings of the discriminator component's final layer. Since the discriminator's goal is opposite to the activity classifier (increasing the loss), we can observe the de-clustered embeddings compared to the final stages of the model training.

IV. Related Work

A decade of research has been inventing an ample number of WiFi-based sensing models [35]–[37], and recently leveraged by Deep Learning [38], [39]. However, one of the key challenges in WiFi sensing is the domain shift problem where the domain being either the sensing environment [9],subject [40] or the subject's position relative to the sensor [41]. More specifically, the sensing model tends to perform poorly when tested in a domain with a data distribution significantly different from the training dataset. Domain shift occurs when the marginal distributions of the source and target data sets differ [37], [42] as the problem is mainly relevant to interaction between model and input, in the wireless sensing literature this is addressed either by *Input Processing*; processing the input to suppress the domain shift impact in the input signal or *Model-based Alignment*; developing the models that can tolerate domain shift while maintaining good performance.

The latter draws mainly on domain adaptation literature that aims at developing models capable of operating in diverse target domains regardless of the distribution data set it was trained in.For example, a model is trained on images captured on a rainy day to identify the road signs but needs to be tested on images captured on a sunny day. Being the task is the same but source and target domains are different making the model perform badly in a new domain it doesn't yet know.

Input Processing WiAG [43] leverages *analytical translation* to generate virtual samples for different locations and orientations in single environment. Extending the analytical translation functions to advanced scenarios (e.g. multi-arm gestures) or other applications isn't straightforward. Thus, subsequent works [8] targeted extracting *domain-independent input*. Widar 3.0 and WiPose leveraged BVP for gesture recognition and pose estimation; respectively. They share the same motivation of our work albeit our methods are different. More specifically, they invariant Doppler that is independent of the receiver configurations. At a high level, their approach aims at reconstructing velocity profile at fixed reference point (i.e. the body coordinates). WiHF [44] addresses this limitation by proposing domain-independent *motion change pattern* features which describes the power distribution over different velocities of the body parts involving the gesture movement. WiGesture [45] proposes a similar feature referred to as MNP (Motion Navigation Primitive) for generating a hand relative view of a gesture. *However, this feature requires multiple WiFi receivers and is tested with figure gestures.* We, on the other hand, translate all Doppler measurements to reference orientation using the proposed deep architecture. A key advantage of our approach is that it can work in a single receiver setup rather than multiple receivers. Making the system more ubiquitous and convenient for use in confined spaces (such as in-car sensing [46]).

Model-based Alignment: In this direction, *transfer learning* was leveraged by CrossSense [30] through roaming models based on MLP (multilayer perceptron). The idea is to translate/roam the wireless features (captured by a localized model trained on a single environment) to a novel environment. The features captured by CrossSense are domain-dependent and the classifier must be re-trained for each new environment. This leads to excessive training effort as the number of environments grows. To mitigate the issue, *adversarial domain adaptation* [40], [47], [48] was investigated and demonstrated

great success. EI [9] proposed to extract environment independent features from multiple wireless sensing modalities with a main goal of of remove the domain specific characteristics from the model feature to achieve alignment. DualConFi [49] proposed a new dual-stream contrastive learning model to process and learn raw WiFi CSI data in a self-supervised manner without requiring a large amount of data to train. This method can complement our work by reducing the effort of labelling. In addition, the novel approach of `WiFi2Radar` can provide a more precise modality to improve WiFi-based sensing, which can also be utilised in dual-stream-based deep learning models. Therefore, incorporating these approaches can potentially lead to improved performance and efficiency in WiFi sensing applications.

Domain Adversarial Neural Network: Many systems that followed the DANN (Domain Adversarial Neural Network) [23], [32] in their core and encompassed three components; feature extractor, activity recogniser and domain discriminator. Traditional Generative Adversarial Networks has the so-called minimax game between the generator and the discriminator. DANN, however, has a feature extractor and activity recogniser that plays a cooperative game. The concatenated features are extracted, and activity labels are fed to the discriminator. The feature extractor and the domain discriminator play a minimax game. At the end of the optimisation, the system generates features that are discriminative to the activity and independent of the domain.

Many domain independent sensing systems used DANN [9], [40], [50] with alterations to the feature extractor and used it with both activities labelled and unlabeled data.Some efforts was followed up by attempts to further upgrade the performance by using multiple discriminators. EUIGR [51] uses two discriminators; one for users and another for environments. WiCAR [52] and CSI-GAN [53] proposed adding even more discriminators. While this in general resulted in a better performance, it was achieved at the expense of more complex training procedure [53]. Our work builds on the advances of deep adversarial-based features alignment and extends its capabilities in a novel way. Our proposed model is based on DANN, with the feature extractor replaced by U-Net style autoencoder.We demonstrated practically that augmenting adversarial architecture with a translation component improves the performance of the underlying architecture without either complicating the training procedure or requiring data from the target domain.

Autoencoder-based Translation. In $WiFi$-2-$Radar$, we leveraged a U-Net autoencoder as a component for WiFi Doppler translation to high-precision Doppler. This is the first time an autoencoder (U-Net is a specialised autoencoder architecture) is used for modality translation to the best of our knowledge. However, the autoencoder architecture is prevalent in deep Learning applications in WiFi due to its capability in feature extraction. Wang et al. [54], [55] discuss the limitations of the manual feature extraction in WiFi as handcrafted features often lead the inference task to lower performance, and they propose a localisation, activity and gesture recognition architecture utilising a 3-layer sparse autoencoder for feature learning. The experimental environments contain 6 radio receiver/transmitter pairs to capture the variances of Received signal strength (RSS). Localisation accuracy of 100% was observed. Gao et al. [56] utilises a similar model for deep feature extraction from image features extracted using colour correlogram, colour autocorrelogram, Gabor filter, and Gray level co-occurrence matrix. Image features were extracted from CSI radio image built from CSI amplitude and calibrated phase information. Finally, this approach is used to locate and recognise the activities performed by the users, and the deep learning-based feature extraction improves approximately 2% accuracy to original accuracy of 93%.

A Fingerprint method for indoor localisation using an autoencoder-based deep extreme learning machine (ADELM) is proposed by Katab et al. [57] for RSS-based localisation. The autoencoder eliminates the need for random weight generation for the extreme learning machine by generating the hidden layer output matrix from the extracted features. ADELM produces a success rate of 92.82% where prior work produces only 86.66% success rate.

Similar to Katab et al., Chang et al. [58] use an autoencoder model for unsupervised pre-training the Deep Neutral Network (DNN) for localisation. This DNN model achieves significant performance improvement of mean distance error 0.47 m and 1.32 m respectively for two target environments. Zhao et al. [59] propose a convolutional autoencoder-based approach to achieve the pre-training of the neural network for localisation. The proposed model achieves near 100% localisation accuracy in a grid of 28 sensor nodes test environment.

In summary, autoencoder based feature extraction has been a popular architecture in wireless sensing since it can improve the performance of the models compared with the manual feature extraction. However, there has not been any autoencoder model for high-precision Doppler feature translation as our proposed model.

V. Discussion and Future Works

In this work, we have demonstrated the applicability of the well-known U-net architecture in translating features from a low-resolution modality, i.e., WiFi, to a high-resolution modality, i.e., mmWave radar, which significantly improved the performance of some basic human gesture recognition task.

One of the limitations of our work is that a single WiFi receiver might not consistently and adequately gather information when the user is positioned behind the data collection device, or when there are obstructions in the scene that hinder data reception. However, we acknowledge that addressing these challenges is a potential avenue for future work.

However, applications, such as controlling smart home devices, where users inherently face the appliance, could greatly benefit from our suggested model.

As part of future research, we plan to investigate the impact of signal attenuation caused by partial occlusion and signal reflection within a limited range on our results. This exploration aims to improve the model's robustness under these conditions and enhance its real-world applicability.

Additionally, our current model has not undergone testing across a range of CSI data collection speeds, which may affect

its performance. Furthermore, the model's ability to accurately classify unseen gestures has not yet been thoroughly evaluated, which presents a potential challenge for future investigations.

It's been discussed in the literature that there's possible performance improvement when it comes to wifi-to-wifi translations collected at different locations [60]. However, the improvements are not significant and consistent. Thus, we haven't tested the possibility of translation between the different wifi pairs, especially regarding the wifi data collected at the boresight direction and other directions which could have been a some improvement.

An interesting future direction would be to explore similar translations between different modality pairs, e.g., WiFi-to-Camera and mmWave-to-Camera. The applications of such translations to more challenging recognition tasks, such as multiple person activity recognition, fine-grained gesture recognition, human-robot interaction detection, etc., would also be worthwhile.

VI. Conclusion

We have proposed and demonstrated the feasibility of deep learning to translate WiFi observations of human activities to their corresponding observations from a millimeter wave radar. We have further shown that when used with appropriate neural networks, such translations can significantly enhance the capabilities of WiFi to recognise human activities performed from any arbitrary orientations. We have tested the effectiveness of our approach with real human experiments and confirmed that the outcomes are valid across different subjects and environments. Since our approach requires the radar only during training while all inferences can be achieved with WiFi signals alone, it can be readily used with any existing WiFi infrastructure.

References

[1] "Iwr1443boost evaluation board — ti.com," 2020. [Online]. Available: https://www.ti.com/tool/IWR1443BOOST

[2] X. Li, S. Chen, S. Zhang, Y. Zhu, Z. Xiao, and X. Wang, "Advancing ir-uwb radar human activity recognition with swin-transformers and supervised contrastive learning," *IEEE Internet of Things Journal*, pp. 1–1, 2023.

[3] J. Xiao, H. Li, M. Wu, H. Jin, M. J. Deen, and J. Cao, "A survey on wireless device-free human sensing: Application scenarios, current solutions, and open issues," *ACM Computing Surveys*, vol. 55, no. 5, p. 1–35, 2023.

[4] Y. He, Y. Chen, Y. Hu, and B. Zeng, "Wifi vision: Sensing, recognition, and detection with commodity mimo-ofdm wifi," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8296–8317, 2020.

[5] S. Tan, Y. Ren, J. Yang, and Y. Chen, "Commodity wifi sensing in ten years: Status, challenges, and opportunities," *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 17 832–17 843, 2022.

[6] H. Wang, D. Zhang, J. Ma, Y. Wang, Y. Wang, D. Wu, T. Gu, and B. Xie, "Human respiration detection with commodity wifi devices: Do user location and body orientation matter?" ser. UbiComp '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 25–36. [Online]. Available: https://doi.org/10.1145/2971648.2971744

[7] K. Niu, F. Zhang, X. Wang, Q. Lv, H. Luo, and D. Zhang, "Understanding wifi signal frequency features for position-independent gesture sensing," *IEEE Transactions on Mobile Computing*, vol. Early Access, pp. 1–1, 2021.

[8] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with wi-fi," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 313–325. [Online]. Available: https://doi.org/10.1145/3307334.3326081

[9] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas *et al.*, "Towards environment independent device free human activity recognition," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018, pp. 289–304.

[10] "Millimeter wave (mmwave) radar sensors — overview — ti.com," 2021. [Online]. Available: https://www.ti.com/sensors/mmwave-radar/overview.html

[11] A. Goldsmith, *Path Loss and Shadowing*. Cambridge University Press, 2005, p. 27–63.

[12] C. Wu, Z. Yang, Z. Zhou, K. Qian, Y. Liu, and M. Liu, "Phaseu: Real-time los identification with wifi," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, 2015, pp. 2038–2046.

[13] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of wifi signal based human activity recognition," in *Proceedings of the 21st annual international conference on mobile computing and networking*, 2015, pp. 65–76.

[14] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Proceedings of the 19th annual international conference on Mobile computing & networking*, 2013, pp. 27–38.

[15] K. Qian, C. Wu, Z. Zhou, Y. Zheng, Z. Yang, and Y. Liu, *Inferring Motion Direction Using Commodity Wi-Fi for Interactive Exergames*. New York, NY, USA: Association for Computing Machinery, 2017, p. 1961–1972. [Online]. Available: https://doi.org/10.1145/3025453.3025678

[16] K. Qian, C. Wu, Z. Yang, Y. Liu, and K. Jamieson, "Widar: Decimeter-level passive tracking via velocity monitoring with commodity wi-fi," in *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, ser. Mobihoc '17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: https://doi.org/10.1145/3084041.3084067

[17] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3291–3300.

[18] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11n traces with channel state information," *ACM SIGCOMM CCR*, vol. 41, no. 1, p. 53, Jan. 2011.

[19] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, p. 559–572, 1901.

[20] E. Sejdić, I. Djurović, and J. Jiang, "Time–frequency feature representation using energy concentration: An overview of recent advances," *Digit. Signal Process.*, vol. 19, no. 1, p. 153–183, jan 2009. [Online]. Available: https://doi.org/10.1016/j.dsp.2007.12.004

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[22] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[23] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[25] A. Creswell, K. Arulkumaran, and A. A. Bharath, "On denoising autoencoders trained to minimise binary cross-entropy," *arXiv preprint arXiv:1708.08487*, 2017.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[27] S. Palipana, D. Salami, L. A. Leiva, and S. Sigg, "Pantomime: Mid-air gesture recognition with sparse millimeter-wave radar point clouds," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 1, mar 2021. [Online]. Available: https://doi.org/10.1145/3448110

[28] H. Liu, A. Zhou, Z. Dong, Y. Sun, J. Zhang, L. Liu, H. Ma, J. Liu, and N. Yang, "M-gesture: Person-independent real-time in-air gesture recognition using commodity millimeter wave radar," *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3397–3415, 2022.

[29] S. Ding, Z. Chen, T. Zheng, and J. Luo, "Rf-net: A unified meta-learning framework for rf-enabled one-shot human activity recognition," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, ser. SenSys '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 517–530. [Online]. Available: https://doi.org/10.1145/3384419.3430735

[30] J. Zhang, Z. Tang, M. Li, D. Fang, P. Nurmi, and Z. Wang, "Crosssense: Towards cross-site and large-scale wifi sensing," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 305–320. [Online]. Available: https://doi.org/10.1145/3241539.3241570

[31] F. Adib, Z. Kabelac, D. Katabi, and R. C. Miller, "3d tracking via body radio reflections," in *11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14)*, 2014, pp. 317–329.

[32] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[33] Y. Wu, Q. Lin, H. Jia, M. Hassan, and W. Hu, "Auto-key: Using autoencoder to speed up gait-based key generation in body area networks," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 1, Mar. 2020. [Online]. Available: https://doi.org/10.1145/3381004

[34] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," *Advances in neural information processing systems*, vol. 15, 2002.

[35] Y. Zeng, P. H. Pathak, and P. Mohapatra, "Wiwho: Wifi-based person identification in smart spaces," in *Proceedings of the 15th International Conference on Information Processing in Sensor Networks*, ser. IPSN '16. IEEE Press, 2016.

[36] D. Huang, R. Nandakumar, and S. Gollakota, "Feasibility and limits of wi-fi imaging," in *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, ser. SenSys '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 266–279. [Online]. Available: https://doi.org/10.1145/2668332.2668344

[37] C. R. Karanam, B. Korany, and Y. Mostofi, "Tracking from one side: Multi-person passive tracking with wifi magnitude measurements," ser. IPSN '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 181–192. [Online]. Available: https://doi.org/10.1145/3302506.3310399

[38] I. Nirmal, A. Khamis, M. Hassan, W. Hu, and X. Zhu, "Deep learning for radio-based human sensing: Recent advances and future directions," *IEEE Communications Surveys Tutorials*, vol. 23, no. 2, pp. 995–1019, 2021.

[39] C. Li, Z. Liu, Y. Yao, Z. Cao, M. Zhang, and Y. Liu, "Wi-fi see it all: Generative adversarial network-augmented versatile wi-fi imaging," ser. SenSys '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 436–448. [Online]. Available: https://doi.org/10.1145/3384419.3430725

[40] M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi, "Learning sleep stages from radio signals: A conditional adversarial architecture," in *International Conference on Machine Learning*. PMLR, 2017, pp. 4100–4109.

[41] Y. Ma, S. Arshad, S. Muniraju, E. Torkildson, E. Rantala, K. Doppler, and G. Zhou, "Location- and person-independent activity recognition with wifi, deep neural networks, and reinforcement learning," *ACM Trans. Internet Things*, vol. 2, no. 1, jan 2021. [Online]. Available: https://doi.org/10.1145/3424739

[42] X. Guo, B. Liu, C. Shi, H. Liu, Y. Chen, and M. C. Chuah, "Wifi-enabled smart human dynamics monitoring," in *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, ser. SenSys '17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: https://doi.org/10.1145/3131672.3131692

[43] A. Virmani and M. Shahzad, "Position and orientation agnostic gesture recognition using wifi," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 252–264. [Online]. Available: https://doi.org/10.1145/3081333.3081340

[44] C. Li, M. Liu, and Z. Cao, "Wihf: Enable user identified gesture recognition with wifi," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 586–595.

[45] R. Gao, M. Zhang, J. Zhang, Y. Li, E. Yi, D. Wu, L. Wang, and D. Zhang, "Towards position-independent sensing for gesture recognition with wi-fi," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 2, p. 1–28, 2021.

[46] Y. Bai, Z. Wang, K. Zheng, X. Wang, and J. Wang, "Widrive: Adaptive wifi-based recognition of driver activity for real-time and safe takeover," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 901–911.

[47] H. Xue, W. Jiang, C. Miao, F. Ma, S. Wang, Y. Yuan, S. Yao, A. Zhang, and L. Su, "Deepmv: Multi-view deep learning for device-free human activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–26, 2020.

[48] H. Kang, Q. Zhang, and Q. Huang, "Context-aware wireless based cross domain gesture recognition," *IEEE Internet of Things Journal*, vol. Early Access, 2021.

[49] K. Xu, J. Wang, L. Zhang, H. Zhu, and D. Zheng, "Dual-stream contrastive learning for channel state information based human activity recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 1, pp. 329–338, 2023.

[50] V. Kakaraparthi, Q. Shao, C. J. Carver, T. Pham, N. Bui, P. Nguyen, X. Zhou, and T. Vu, "Facesense: Sensing face touch with an ear-worn system," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 3, sep 2021. [Online]. Available: https://doi.org/10.1145/3478129

[51] Y. Yu, D. Wang, R. Zhao, and Q. Zhang, "Rfid based real-time recognition of ongoing gesture with adversarial learning," in *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, 2019, pp. 298–310.

[52] F. Wang, J. Liu, and W. Gong, "Wicar: Wifi-based in-car activity recognition with multi-adversarial domain adaptation," in *Proceedings of the International Symposium on Quality of Service*, ser. IWQoS '19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3326285.3329054

[53] C. Xiao, D. Han, Y. Ma, and Z. Qin, "Csigan: Robust channel state information-based activity recognition with gans," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 191–10 204, Dec 2019.

[54] J. Wang, X. Zhang, Q. Gao, H. Yue, and H. Wang, "Device-free wireless localization and activity recognition: A deep learning approach," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 7, pp. 6258–6267, July 2017.

[55] X. Zhang, J. Wang, Q. Gao, X. Ma, and H. Wang, "Device-free wireless localization and activity recognition with deep learning," in *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, March 2016, pp. 1–5.

[56] Q. Gao, J. Wang, X. Ma, X. Feng, and H. Wang, "Csi-based device-free wireless localization and activity recognition using radio image features," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10 346–10 356, Nov 2017.

[57] Z. E. Khatab, A. Hajihoseini, and S. A. Ghorashi, "A fingerprint method for indoor localization using autoencoder based deep extreme learning machine," *IEEE Sensors Letters*, vol. 2, no. 1, pp. 1–4, March 2018.

[58] R. Y. Chang, S.-J. Liu, and Y.-K. Cheng, "Device-free indoor localization using wi-fi channel state information for internet of things," *2018 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–7, 2018.

[59] L. Zhao, H. Huang, X. Li, S. Ding, H. Zhao, and Z. Han, "An accurate and robust approach of device-free localization with convolutional autoencoder," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5825–5840, June 2019.

[60] I. Nirmal, A. Khamis, W. Hu, M. Hassan, and X. Zhu, "Poster abstract: Combating transceiver layout variation in device-free wifi sensing using convolutional autoencoder," in *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, 2020, pp. 335–336.

VII. Appendix

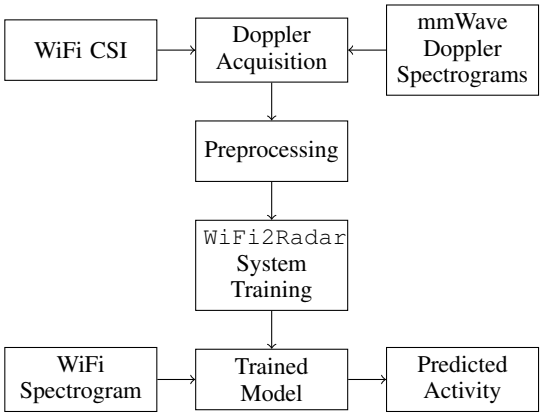


Fig. 13: Flowchart: Overall System Model

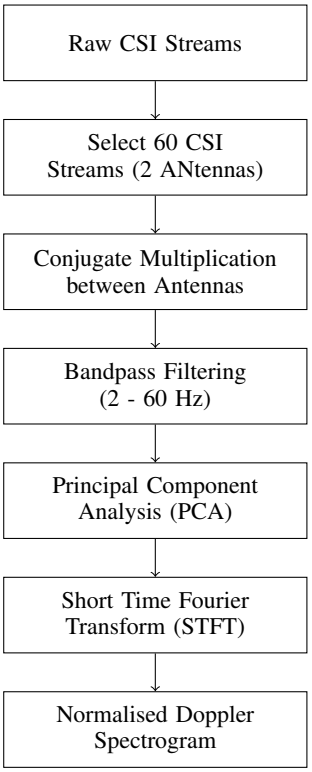


Fig. 14: Flowchart: Preprocessing for CSI Measurement to Doppler Spectrogram

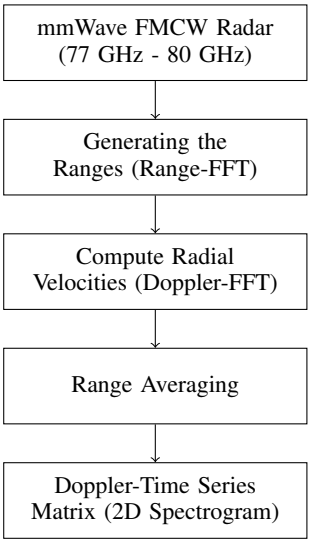


Fig. 15: Flowchart: Obtaining Doppler Ground Truth Information

Isura Nirmal,PhD Currently serving as a Research Associate in the School of Computer Science and Engineering at the University of New South Wales (UNSW) in Sydney, Australia, Isura Nirmal earned his Ph.D. in Computer Science and Engineering from UNSW in 2023. His primary research pursuits encompass wireless sensing, Internet of Things (IoT), and deep learning.

Dr.Abdelwahed Khamis Abdelwahed Khamis is a research scientist at the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Data61 in Australia. Abdelwahed completed his Ph.D. in Computer Science and Engineering from UNSW, Sydney in 2020. His current research interests include AI on the edge, RF and device-free sensing, and Multimodal Sensing.

Mahbub Hassan is a Full Professor in the School of Computer Science and Engineering, University of New South Wales, Sydney, Australia. He has PhD from Monash University, Australia, and MSc from University of Victoria, Canada, both in Computer Science. He served as IEEE Distinguished Lecturer and held visiting appointments at universities in USA, France, Japan, and Taiwan. He has co-authored three books, over 200 scientific articles, and a US patent. He served as editor or guest editor for many journals including IEEE Communications Magazine, IEEE Network, and IEEE Transactions on Multimedia. His current research interests include Mobile Computing and Sensing, Nanoscale Communication, and Wireless Communication Networks. More information is available from http://www.cse.unsw.edu.au/~mahbub.

Wen Hu is currently a Professor with the School of Computer Science and Engineering, University of New South Wales (UNSW). He has published regularly in the top-rated sensor network and mobile computing venues, such as IPSN, SenSys, MobiCom, UbiComp, TOSN, the TMC, TIFS, and the PROCEEDINGS OF THE IEEE. His research interests focus on novel applications, low-power communications, security, and compressive sensing in sensor network systems and the Internet of Things (IoT). He is a Senior Member of ACM and IEEE. He is an Associate Editor of ACM TOSN. He is the General Chair of CPS-IoT Week 2020, and serves on the organizing and program committees of networking conferences, including IPSN, SenSys, MobiSys, MobiCom, and IoTDI.

Rui Li Rui Li is a MPhil student enrolled in the School of Computer Science and Engineering at UNSW, Sydney. His research interests include object sensing, radar technology, and artificial intelligence.

Avinash Kalyanaraman is currently a Researcher with Cisco Innovation Labs (Office of the CTO) in San Jose, California. He received his PhD from the University of Virginia where he built radar sensing systems for smarter indoor human environments. His work has been published in leading sensing and mobile computing venues, such as UbiComp, MobiSys, RTAS, ToSN, etc., including a Best-Paper Runner-Up Award at SECON 2020. His current research involves exploring and building technologies that can simplify and improve the way we work and operate in enterprise buildings