Gesture Recognition with mmWave Wi-Fi Access Points: Lessons Learned

Nabeel Nisar Bhat
IDLab-Faculty of Science
University of Antwerp-imec
Antwerp, Belgium
nabeelnisar.bhat@uantwerpen.be

Rafael Berkvens
IDLab-Faculty of Applied Engineering
University of Antwerp-imec
Antwerp, Belgium
rafael.berkvens@uantwerpen.be

Jeroen Famaey
IDLab-Faculty of Science
University of Antwerp-imec
Antwerp, Belgium
jeroen.famaey@uantwerpen.be

***Abstract*—In recent years, channel state information (CSI) at sub-6 GHz has been widely exploited for Wi-Fi sensing, particularly for activity and gesture recognition. In this work, we instead explore mmWave (60 GHz) Wi-Fi signals for gesture recognition/pose estimation. Our focus is on the mmWave Wi-Fi signals so that they can be used not only for high data rate communication but also for improved sensing *e.g.*, for extended reality (XR) applications. For this reason, we extract spatial beam signal-to-noise ratios (SNRs) from the periodic beam training employed by IEEE 802.11ad devices. We consider a set of 10 gestures/poses motivated by XR applications. We conduct experiments in two environments and with three people. As a comparison, we also collect CSI from IEEE 802.11ac devices. To extract features from the CSI and the beam SNR, we leverage a deep neural network (DNN). The DNN classifier achieves promising results on the beam SNR task with state-of-the-art 96.7% accuracy in a single environment, even with a limited dataset. We also investigate the robustness of the beam SNR against CSI across different environments. Our experiments reveal that features from the CSI generalize without additional re-training, while those from beam SNRs do not. Therefore, re-training is required in the latter case.**

***Index Terms*—Wi-Fi signals, context aware, human activity recognition, gesture recognition, millimeter-wave, CSI, beam SNR, deep neural networks.**

I. Introduction

In the past few years, wireless signals have caught significant attention, and a new field known as wireless sensing has emerged. In the wireless sensing world, most of the focus has been on Wi-Fi signals and, consequently, on Wi-Fi-sensing, primarily due to the availability of such devices. Wi-Fi signals at 2.4 and 5 GHz have been widely used for activity recognition [1]–[4], gesture recognition/pose estimation [5]–[11], human detection [12]–[14], localization [15]–[17], crowd counting [18], biological activities [19], [20], and human identification [21]–[23]. Most of these works exploit physical layer parameters, such as channel state information (CSI), from Wi-Fi devices and use a machine learning/deep learning approach to classify such activities. Compared to camera-based activity recognition, Wi-Fi has improved privacy, works in non-line-of-sight, and does not require a well-lit environment [24]. Moreover, Wi-Fi signals contain significant information about the environment. Due to their low cost, Wi-Fi devices suffer from hardware impairments such as sampling frequency offset (SFO), carrier frequency offset (CFO), multi-path propagation, and fading. Therefore, environment robustness/generalization is one of the key challenges with Wi-Fi sensing.

Among the Wi-Fi standards, much of the focus has been on 2.4 GHz and 5 GHz frequency bands (*e.g.*, IEEE 802.11n and IEEE 802.11ac). However, radio signals at these frequencies have limited resolution due to low bandwidth. Recently, the focus has shifted to mmWave signals (>30 GHz), which offer several advantages compared to sub-6 GHz. Apart from increased data rates for communication, mmWaves have a dominant line-of-sight component with respect to non-line-of-sight [25]. Also, due to larger bandwidth, fine delay resolution of multi-path components can be achieved [26]. Thanks to the narrow-beam or pencil-beam [26], the spatial resolution is very high at mmWave frequencies. Therefore, by moving to mmWave, improved sensing is possible. mmWave radars have been widely used for gesture recognition in human-computer interactions [24], [27] [28]. However, in this work, we exploit commercial-off-the-shelf (COTS) mmWave Wi-Fi as opposed to dedicated radar. Our focus is on joint communication and sensing (JCAS), in the sense that we want to use the same signals for communication as well as for sensing. Using the same signals for communication and sensing gives us several advantages, such as cost and availability.

In this work, we demonstrate gesture/pose recognition capabilities of mmWave Wi-Fi devices, which have not yet been explored much. We extract beam signal-to-noise ratios (SNRs) from the periodic sector sweep algorithm employed by IEEE 802.11ad devices. Compared to CSI, beam SNRs can be considered a direct indicator of channel quality [29]. Based on the beam SNRs, we train a deep neural network (DNN) to extract features and map changes in SNRs to different gestures. We perform multiple experiments with different people, different environments, and different orientations of people. Our gestures are motivated by the extended reality (XR) applications. Since XR applications require high data rates, which can be achieved using mmWave access points [30], our focus is on the sensing capabilities of these devices so that device-free (hands-free) sensing can be achieved in the XR world. Therefore, we investigate the performance of mid-grained beam SNRs and compare it to fine-grained CSI captured from a 5 GHz Wi-Fi access point with an application for gesture recognition. The scope of this research is not to develop new deep learning pipelines for CSI/beam SNR-based gesture recognition. Instead, we use existing convolutional neural network (CNN) architectures and tailor them according to the data and needs to demonstrate the performance of 60 GHz beam SNR. Our experiments with 10 XR-related gestures in natural environments reveal exciting results. Though gesture recognition can be performed reliably even with low-sample beam SNR, we learn some interesting lessons from our trials for future improvement.

The rest of the paper is organized as follows. First, we discuss the related work in Section II concerning CSI and beam SNR. The related work highlights recent developments in Wi-Fi sensing at sub-6 GHz and 60 GHz frequencies. Then in Section III, we describe the hardware setup and the details of the gestures/poses. This is followed by Section IV, where we describe our method, *i.e.,* deep neural network, to extract the features from the data. Finally, we describe the experiments and performance of our method in Section V.

II. Related Work

Activity recognition is the process of identifying the actions of a subject (*e.g.,* human or robot) from a series of observations. It involves detecting movements such as fall, sitting, standing, and running. On the other hand, gesture recognition involves subtle movements like lifting a hand, swiping, and arm movements. The latter is much more challenging than coarse activity recognition. This section reviews the existing literature on gesture recognition based on Wi-Fi. Our focus is on the works which exploit CSI and beam SNR at sub-6 GHz and 60 GHz, respectively.

A. Channel State Information (CSI)

WiG [6] is one of the pioneering works in Wi-Fi sensing. WiG is a low-cost device-free gesture recognition system based on CSI. The authors use a Linux 802.11n CSI tool [31] to extract CSI from Intel 5300 network interface card (NIC). Wavelet denoising is used to clean and smoothen the raw CSI data. After cleaning CSI, outliers are removed, and finally, data is fed to the Support Vector Machine (SVM) classifier. The method achieves 92% accuracy in classifying four different gestures under line-of-sight conditions. Abdulaziz et al. [32] describe a device-free gesture recognition system based on CSI. It uses dynamic time warping to classify different hand gestures. Instead of classification, recent works exploit Wi-Fi signals to construct 2D human poses for fine-grained perception. For example, Wang et al. [33] use RGB camera images as annotations. Using deep learning, 2D body pose coordinates are reconstructed from Wi-Fi signals. The results obtained are quite similar to computer vision based methods on 2D images. However, the study is conducted in a single environment.

Since CSI varies under different environments and the fact that different people perform gestures differently, the focus has recently shifted to robustness and generalization of CSI-based gesture recognition, in addition to multi-people sensing. Jian et al. [34] present EI, a deep-learning-based approach to achieve domain-independent activity recognition. It incorporates an adversarial network consisting of a CNN-based feature extractor. The proposed network can extract common features among different domains. Similarly, the authors propose CrossSense [35], an artificial neural network, to generate virtual data for the new environment from previously collected measurements. The authors use transfer learning to train a network with fewer data from the unseen site. Transfer learning reduces the training cost in unseen environments. Differently, the authors propose a cross-domain gesture-recognition system [9] that involves extracting a domain-independent feature called body-coordinate velocity profile (BVP) from the gestures. With velocity profiles, the authors develop a model trained only once and can adapt to unseen domains and orientations without re-training. Similar to [9], Jiang et al. propose WiPose [36], the first 3D human pose construction from Wi-Fi signals and use a velocity profile for separating posture-specific features from static objects in the environment. WiPose uses a VICON system to generate 3D skeletons as ground truth. It then leverages a DNN to construct 3D skeletons from CSI at 5 GHz. WiPose achieves a 2.83 cm average error in localizing each skeletal joint. Winect [37] is another gesture recognition system that outputs a 3D skeleton of a human body using a 2D angle of arrival information from reflected Wi-Fi signals. Winect is environment independent and does not rely on predefined activities. Notably, it can track free-form movements for human-computer interaction (HCI).

More challenging works involve extending gesture recognition to multiple people simultaneously. Venkatnarayan et al. propose WiMU [38], WiFi-based multi-user gesture recognition. WiMu can recognize up to 6 simultaneously performed predefined gestures with high accuracy, around 90%. It can also identify the start and end times of gestures automatically. However, it can not determine which user performed which gesture.

Recently, focus has shifted to mmWave signals for gesture recognition, particularly for radar-based systems. Ren et al. [28] present a hand gesture recognition prototype with 60 GHz mmWave technology. Radar and communication waveforms are transmitted in time-division duplex (TDD). Range-Doppler information (RDI) is obtained from the Doppler radar and exploited for gesture recognition. RDI is then fed to a CNN+LSTM model, achieving >95% accuracy for gesture recognition.

B. Beam Signal-to-Noise ratio (SNR): 60 GHz

Yu et al. [29] exploited beam SNR at 60 GHz for the first time in Wi-Fi sensing. The authors used beam SNR for pose estimation for a single person in one environment. A neural network was used to train the classifier. The classifier achieves 90% accuracy in identifying the pose correctly. Moreover, quantum transfer learning [39] was used to improve the performance of the beam SNR-based gesture recognition along 7 independent sessions in the same environment. The authors consider the first four sessions as the source domain and the latter as the target domain. Since the environment evolves and

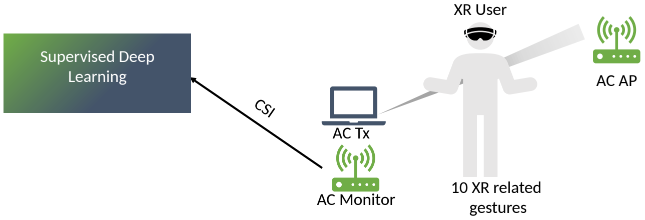

Fig. 1. CSI setup at 5 GHz.

Wi-Fi settings change over time, transfer learning can be used to mitigate this issue. The quantum neural networks (QNN) achieve 90% accuracy for pose recognition when the test data comes from different domains. Apart from these works, not much has been done in the context of COTS mmWave Wi-Fi sensing.

With respect to the prior art, we consider a wider range of gestures motivated by XR applications. These gestures are similar (*e.g.,* left swipe and right swipe) and therefore represent a challenge for deep learning pipelines to extract gesture-specific features. Moreover, we extend our experiments to multiple people and environments to draw broader conclusions about the robustness and generalization of using beam SNR. We show that even with a limited dataset and low sample rate, beam SNR-based gesture recognition can achieve a very high accuracy of around 96.7%. Thanks to the deep neural network presented in Section IV-B and the careful tuning of hyper-parameters, our classifier achieves higher accuracy than the state-of-the-art. However, beam SNR-based gesture recognition falls behind CSI in terms of generalization towards multiple people and environments.

III. Data Collection

This section briefly introduces the concept of CSI and beam SNR. Moreover, we describe our experimental setup at 5 GHz and 60 GHz. Thanks to the Nexmon tool [40], we can extract fine-grained CSI from the 5 GHz device. However, the same can not be accessed from the 60 GHz router without additional overhead. Instead, we use mid-grained beam SNR at 60 GHz, which can be acquired with zero overhead from the mmWave beam training. Finally, we describe the collected data.

A. CSI

Channel state information (CSI) measures the frequency response of the channel. CSI is a measure of changes that a signal undergoes while propagating from transmitter to receiver. CSI changes based on the movement of the transmitter, receiver, or surrounding objects [41].
Mathematically, the wireless channel can be modeled:

$$Y = HX + N \tag{1}$$

where Y represents the received signal, X represents the transmitted signal, H represents the CSI matrix, and N represents the noise vector.
In practice, H is a $4D$ tensor consisting of NxMxKxT dimensions, where N and M correspond to number of transmitting and receiving antennas respectively, K corresponds to number of sub-carriers, T represents the packets in time. Wi-Fi devices estimate CSI with pilot symbols that are known at the transmitter. Moreover, modern Wi-Fi devices use spatial multiplexing and OFDM. Therefore, H changes along space and carrier domain, on top of time domain variations. The CSI matrix is analogous to a digital image where N and M represent changes along space, and K corresponds to color channels. Therefore, deep learning techniques used for images can be applied to CSI-based sensing tasks without much modification.

Our hardware setup at 5 GHz consists of two ASUS routers (RT-AC86U) and an Intel Laptop, both supporting IEEE 802.11ac. One of the ASUS routers functions as an Access Point (AP). Since Wi-Fi devices do not expose CSI measurements to end users, we turn the other ASUS router into a CSI extractor with a firmware patch [40] to access CSI measurements. We set up the system in accordance with Figure 1. So, the setup consists of two active terminals *i.e.*, Intel Laptop and the ASUS AP, while the other ASUS router acts as a passive device. The Intel transmitter and the AP exchange packets, while the patched ASUS router works as a monitor and sniffs the conversation between the AP and the laptop. A user performs gestures in the line-of-sight between the AP and the transmitter. The monitor then extracts the corresponding CSI from the packets sent back by the AP. This situation represents a real-world scenario where two Wi-Fi-enabled devices exchange packets, and a passive monitor listens to the conversation. In practice, the active terminals can be arbitrary devices exchanging Wi-Fi traffic, *e.g.*, streaming XR content using the AP or accessing YouTube via Wi-Fi. In our case, we use only one transmitting antenna ($N = 1$) and one receiving antenna ($M = 1$). The Intel transmitter pings 1000 packets per second of size 3008 bytes to the ASUS AP. The ASUS AP then replies with pong packets of the same size, and the corresponding CSI is captured by the monitor. The CSI is captured according to the package rate. However, the actual capture rate depends on various things, such as the CPU and memory of the device. We collect CSI measurements at a bandwidth of 80 MHz from 256 different sub-carriers.

B. Beam SNR

Most of the research focus in COTS Wi-Fi sensing has been on sub-6 GHz frequency bands, mainly due to the existence of tools that provide access to CSI measurements for Atheros and Intel-based NICs. In contrast, at 60 GHz, few devices give access to channel measurements. The Talon AD 7200 router from TP-Link is the first IEEE 802.11ad compliant Wi-Fi device that was patched [42] to provide access to raw channel measurements in the form of beam SNR. The router has 32 programmable antenna elements in the form of a planar array. The gain and phase of each antenna element can be controlled to change the radiation pattern of the beam. However, such a process results in huge computational costs owing to a large number of permutations and combinations (P). Hence,

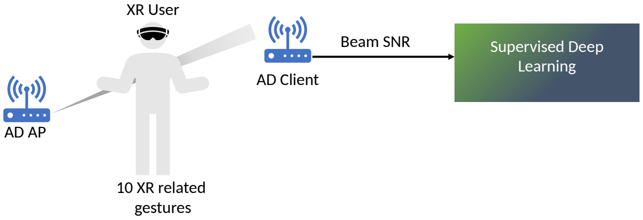

Fig. 2. Beam SNR setup at 60 GHz.

in practice, IEEE 802.11ad routers use a set of predefined patterns ($n<<P$, $n = 36$) known as sectors. 802.11ad devices use a sector sweep algorithm to determine signal strength per sector, and then communication proceeds with the sector with maximum SNR. These devices periodically repeat the sector sweep to react to environmental changes or movements. The sector sweep is triggered at least once per second [43]. In practice, the rate of sector sweep depends on the degree of obstruction along the line-of-sight. If we obstruct the line-of-sight with gestures/poses, the maximum number of sweeps per second could be as high as 10 since it is linked to the beacon interval of 102.4 milliseconds. We leave the beacon interval to the default value, this preserves the concept of JCAS as opposed to increasing the frequency, which can create an additional overhead for communication. The beam SNR is then extracted from the sector sweep frames. The beam SNR can be modeled by the following equation:

$$B_k = 1/\sigma^2 \sum_{n=1}^{N} F_k(\theta_n) G_k \Phi_n = 1 \tag{2}$$

where k is the index of beam pattern, σ^2 is the noise variance, N is the total number of paths, θ_n, and Φ_n are the azimuth angles for transmission and reception, respectively for the n^{th} path. $F_k(\theta_n)$ and $G_k(\Phi_n)$ are beam pattern gains of the transmitting and receiving antennas at the n^{th} path and k^{th} pattern.

Figure 2 shows experimental setup at 60 GHz. We use two Talon AD-7200s, one as an Access Point (AP) and the other as a client. iPerf is set between the two devices in TCP mode, resulting in a data rate of around 1Gbps. We observe that using iPerf, sector sweeps are triggered more frequently than without using any traffic exchange between the two. Moreover, it ensures that the client does not disassociate from the mmWave network during the experiment. In the absence of traffic exchange, the client sometimes disconnects, which is undesirable. Based on the toolset [42], beam SNR measurements from 36 sectors can be extracted from the AP and saved on a local machine. Compared to CSI, beam SNR has a relatively low sample rate which depends on the rate at which the AP performs sector sweep, as mentioned above.

Although the bandwidth is higher at 60 GHz, the lower sample rate and lower number of features in beam SNR are expected to lead to reduced performance of the deep learning algorithms compared to CSI-based gesture recognition.

C. Details of Gestures/Poses

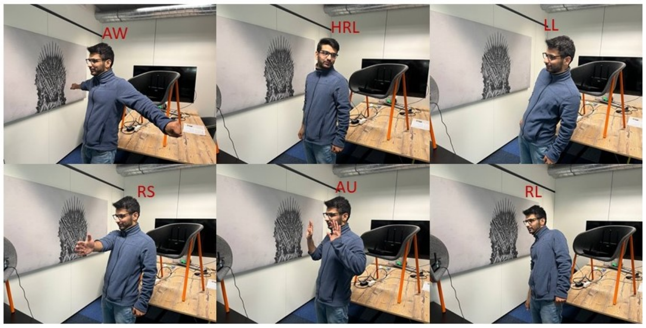

Fig. 3. XR related gestures/poses

We take inspiration from XR applications and perform 10 gestures/poses in the line-of-sight between the AP and the client. The duration of each gesture/pose is set to 15 seconds. We define poses as fixed gestures where the person remains in a particular position for the entire 15 seconds, while the gestures involve continuous movements for the entire duration. We collect the following poses. For the sake of brevity and space constraints, only some of them are shown in Figure 3:

1) Empty (E) pose means that a person stands between the routers and does nothing. This pose serves as a baseline for other gestures/poses.
2) Right Lean (RL) involves leaning to the right.
3) Left Lean (LL) is the counterpart of the RL and therefore involves leaning to the left.
4) Arms Up (AU) is lifting one's arms.
5) Arms Wide (AW) involves opening arms wide horizontally.

In addition, we collect the following gestures:

1) Push (P) involves moving the hand forward.
2) Right Swipe (RS) involves swiping the hand to the right.
3) Left Swipe (LS) is the counterpart of RS and involves moving the same hand in the opposite direction.
4) Head Rotations Left (HRL) involves moving the head to the left.
5) Head rotations Right (HRR) involves moving the head to the right.

To the best of our knowledge, HRL and HRR are unexplored in the related work. These challenging gestures and poses represent common scenarios in different XR applications. The collected dataset is openly available to the community for research and benchmarking[1].

IV. Methodology

First, we patch the firmware of the ASUS and Talon routers to get access to CSI [40] and beam SNR measurements [42], respectively. For 5 GHz, we apply Medium Access Control (MAC) filtering to capture the packets from the expected device. The user then performs each gesture along the line-of-sight between the routers for 15 seconds according to Figures

[1] [https://zenodo.org/record/7813244]

1 and 2. The CSI matrix takes the shape of $x \times 256$, where x represents the number of samples in time, and 256 represents the number of subcarriers. On the other hand, the beam SNR matrix takes the shape of $y \times 36$, where y represents the number of beam SNR samples in time and 36 the number of sectors. We use only amplitude information for the CSI and ignore the phase since our focus is on beam SNRs, so we take the most straightforward approach for CSI. We manually annotate the CSI and beam SNR data for different gestures/poses. Then the raw CSI and beam SNR data are fed to the deep neural network-based classifier. Using deep neural networks has the advantage of directly extracting features from the raw data; there is no need for additional pre-processing.

A. *Deep Learning*

Deep Learning has been widely used for images in solving complex tasks such as image classification, object detection, and image reconstruction. In particular, Convolutional neural networks (CNNs) have been instrumental in facial recognition, biometric authentication, and autonomous driving. CNN's have been widely used for real-world applications because they are outstanding in finding patterns within the data. In the last decade, numerous CNN architectures have been proposed. For example, AlexNet, InceptionNet, VGG, and ResNet have found success in accomplishing challenging tasks. The type of architecture to use depends on the task and objectives like cost and accuracy. Inspired by the performance of neural networks for image processing tasks, CNNs have been widely adopted for Wi-Fi sensing. Recently, 1D CNNs have been used [44] for WiFi-based activity recognition. Since CSI and beam SNR are time-series data, the 1D convolution kernel can capture patterns along the time dimension. Based on our analysis of public datasets like ARIL [44], and WIAR [45], we found that InceptionNet (GoogLeNet) with just 2 inception modules is sufficient to have a significant improvement both in terms of computational cost and accuracy over the proposed architectures [44], [45]. Therefore, we use GoogLeNet with 2 inception modules to learn the CSI and beam SNR features corresponding to different gestures.

B. *Network*

GoogLeNet [46] is a 22 layer deep network proposed for the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC2014). However, since our data and task are less complex, we decided to play with the architecture to find a good compromise between the computational cost and accuracy. We reduce the depth of the network to 7 layers, with only 2 inception modules. We use the same network for the beam SNR and CSI tasks. The rationale behind the inception modules is to use convolutional kernels of different sizes at the same level. As a result, a smaller kernel captures information distributed locally while a larger kernel captures information distributed globally. Figure 4 shows the architecture of the CNN-based classifier. The input is a CSI matrix/beam SNR, and accordingly, the first convolutional layer has either 256 or 36 input channels, respectively. Conv. 1D (7,64) represents the convolutional layer with kernel 7×1 and 64 output channels. The inception 3a and 3b blocks are implemented in the same way as in the original paper [46]. After the Average Pool, the last convolutional layer reduces the number of output channels to 64. These layers extract the features from CSI and beam SNR tensors. The last layer is the fully connected layer that provides an output score for each of the 10 gestures.

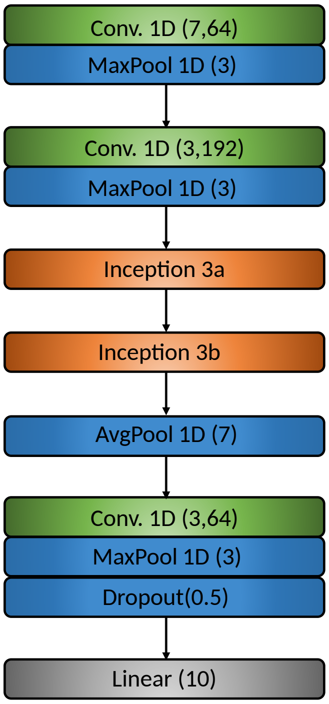

Fig. 4. GoogLeNet: Depth 7

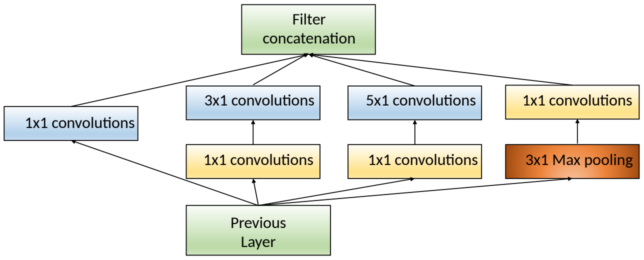

Fig. 5. Inception Module

Figure 5 shows the inception module. The 1×1 convolutions (bottleneck layers) are used to decrease the computational cost by reducing the number of output channels before the input is fed to 3×1 and 5×1 convolutional layers. These convolutions don't change the size of the data. At the end, the outputs from two 1×1, 3×1, and 5×1 convolutions are concatenated across channels.

V. Experiments

We collect 10 gestures/poses from three humans in two different environments. The associated CSI and beam SNR variations are fed to the classifier to learn the features corresponding to these gestures.

A. Testbed and experimental setup

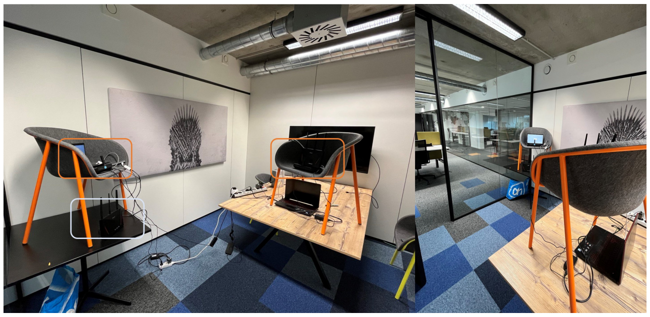

Fig. 6. Testbed: Office Environment

One of the environments is a typical living room (home), and the other is a small office space. The living room has a lot of scatterers or reflectors than the office environment. The dimensions of the home environment are 6 by 5 meters. While the second smaller environment measures 3.45 meters in length and 3 meters in width, for the sake of space constraints, we only show one of the environments *i.e.*, office environment (cf., Figure 6). The room has wooden walls on three sides and a glass panel on one side with a wooden door. Therefore, the reflections are more natural. The IEEE 802.11ad AP and client and IEEE 802.11ac AP and client devices are kept on the chairs, while the IEEE 802.11ac monitor is placed below the chair. The routers are highlighted with an orange box, set 1.26 meters high with a separation of 2 meters. The light blue box indicates the position of the monitor (5 GHz). A similar setup is put in the home environment where routers are placed 1.3 meters high with a separation of 2 meters.

We collect most of the gestures/poses in the home environment and a subset of those in the office environment. A total of 854 gesture instances for beam SNR and CSI are recorded.

B. Hyper-parameter Settings

Hyper-parameters control the learning process of the model. They are user-defined and are used to improve the performance of the model. Tuning the deep learning model can be quite a challenging task. There are no magic rules or settings. These parameters vary from task to task and model to model. After a wide range of trial and error experiments, we use the following settings in our model. We set the number of epochs to 150 for both CSI and beam SNR tasks. We use Adam optimizer with a learning rate of 3e-4 and betas (0.9, 0.999) and epsilon values (1e-8). The batch size is set to 16 and 64 for the SNR and CSI tasks, respectively. We use the scheduler *ReduceLROnPlateau* with patience of 25. The scheduler monitors the learning rate and decreases it if the metric does not improve after a patience number of epochs. This choice of parameters considerably impacts the performance of the deep learning model, especially for the beam SNR task.

C. Experiment 1: Home Environment

The details of the gestures/poses are provided in Section III-C. A single user performs gestures along the line-of-sight between the devices facing the router. Each gesture/pose lasts for 15 seconds. Each gesture is repeated around 50 times so that we have a total of 486 gesture instances of beam SNR and CSI after rejecting invalid data. For the sake of fair comparison, we use the same number of examples for the two tasks. We manually annotate the data and feed it to the deep neural network shown in Figure 4 without additional processing. We split the data randomly into training and test with a standard 75:25 split, respectively. The training set consists of 364 examples, while the test data consisting of 122 gesture instances is left unseen to the classifier.

D. Experiment 2: Office Environment

Three humans perform gestures/poses along the line-of-sight between the devices. In this environment, we collected a total of 221 gestures (excluding the invalid data). We stick to the same hyper-parameter settings and the same train-test split. The training and test data consist of 165 and 56 examples, respectively. Due to the smaller dataset, we expect to see reduced performance in the office environment.

E. Experiment 3: Rotated User Direction

In the subsequent experiments, we combine the data from the two environments. We get a total of 530 and 177 training and test examples (instances), respectively. Also, we test if the deep learning model generalizes to environments or different people without re-training.

In the above two experiments, the user faced the same direction (towards the router) while performing gestures. In an additional experiment, a single user performs gestures rotated 90° with respect to the earlier position *i.e.*, we study how orientation affects the performance of the deep learning model. We collect 147 gestures with 90° rotation in the home environment. We also train on one orientation (0°) and test on another (90°) to see if orientation-independent gesture recognition could be achieved with CSI and beam SNR.

F. Performance of the deep learning model

We collect data from two environments and train the deep learning network with the training data. Then we evaluate the performance on the test data using a tool known as

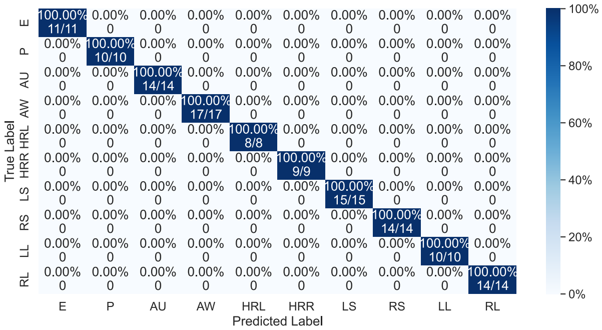

Fig. 7. Performance of DNN on the CSI task: Home Environment.

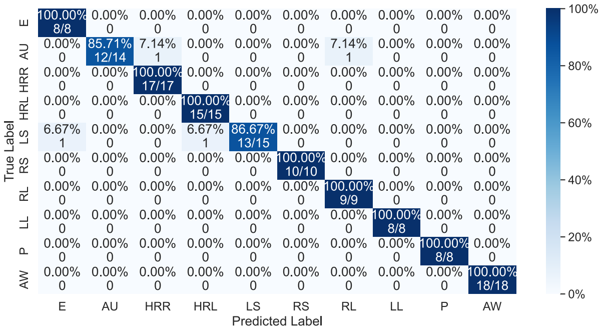

Fig. 8. Performance of DNN on the beam SNR task: Home Environment.

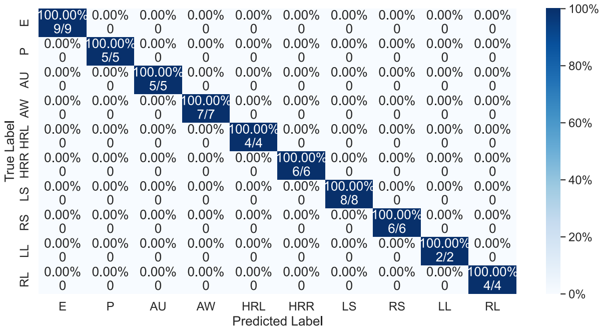

Fig. 10. Performance of DNN on the CSI task: Office Environment.

the confusion matrix. Figures 7 and 8 show the confusion matrices depicting the performance of the network in the home environment for the CSI and beam SNR tasks, respectively. With beam SNRs, we achieve a promising accuracy of around 96.7%, with a per gesture accuracy above 85% even with limited data. Notably, the DNN achieves 100% accuracy with similar gestures such as RL and LL and HRR and HRL. Compared to [29], our DNN-based classifier achieves higher accuracy even with less data. On the other hand, the same DNN achieves 100% accuracy on the CSI task. With CSI, we achieve the same accuracy as [47]. However, instead of an LSTM, we use a CNN, which has a significantly lower computational cost. We believe that the gain in performance for the CSI compared to beam SNR is due to a higher number of features and a higher sample rate. This experiment validates the potential use of mid-grained low-sample beam SNRs for passive gesture recognition in challenging XR applications.

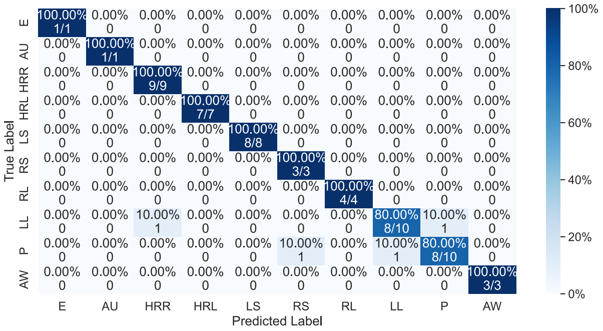

Fig. 9. Performance of DNN on the beam SNR task: Office Environment.

For experiment 2 in the office environment (cf., Figure 9), where 3 users perform gestures, the beam SNR achieves an overall accuracy of around 92.8% accuracy. Also, in this case, the DNN can extract gesture-specific features from similar gestures such as HRL and HRR. The drop in accuracy is because of the smaller dataset. While CSI still outperforms beam SNR and achieves 100% accuracy (cf., Figure 10).

We then test the robustness of our neural network across multiple people and environments. We train on one environment and test on another. We see that DNN on the CSI task can generalize to different people and different environments *i.e.*, a DNN trained in the home environment achieves 100% accuracy in the office environment (unseen environment) as shown in Figure 11. We believe that the neural network can extract deeper and more common features shared across environments from the CSI. While for the beam SNR, the DNN achieves 10% overall accuracy (cf., Figure 12) on the unseen environment. Hence, the DNN can not generalize to

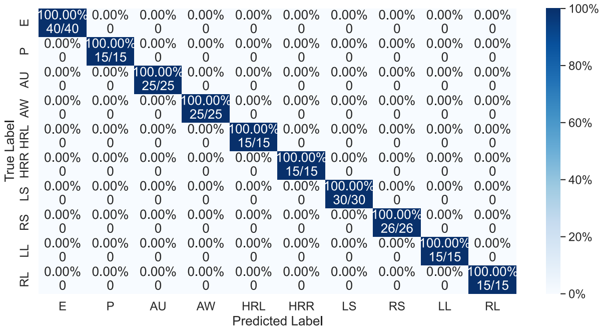

Fig. 11. Generalization of DNN on the CSI task: Train Home Environment and test Office Environment.

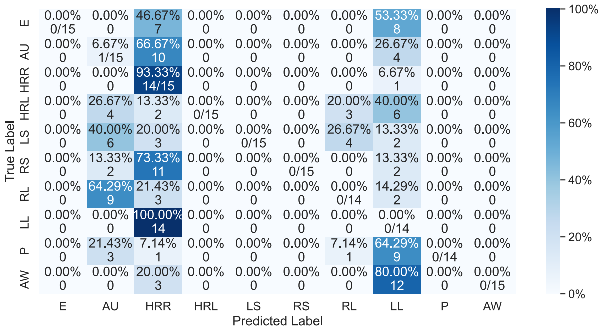

Fig. 12. Generalization of DNN on the beam SNR task: Train Home Environment and test Office Environment.

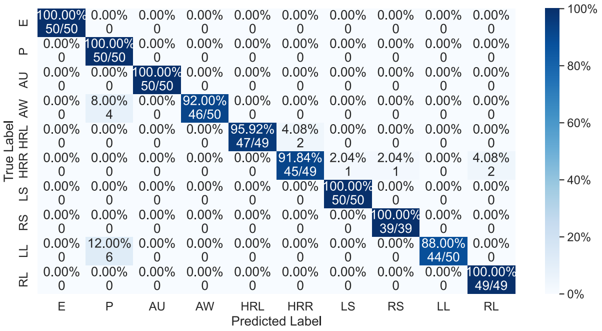

Fig. 13. Generalization of DNN on the CSI task: Train Office Environment and test Home Environment.

different people and environments on which it was not trained. We did a reverse test too *i.e.*, we trained the neural network in the office environment and tested it in the home environment.

In this case, the DNN achieves 96.7% accuracy on the CSI task (cf., Figure 13). The drop in accuracy is due to a lower number of training examples in the office environment with respect to the home. Also, in this scenario, the beam SNR does not generalize across people and environments. This occurs because beam SNR is mid-grained or coarse-grained compared to fine-grained CSI. Moreover, we believe that beam SNR significantly depends on the environment compared to CSI. Therefore, features extracted by the DNN from CSI generalize, while those from beam SNR do not. Thus, the DNN needs the exact example of the person and environment to be able to classify gestures correctly based on the beam SNR task. Hence, for the beam SNR, re-training is required in the unseen environment. Therefore, we collected 10 additional instances per gesture per person in the second environment and fed them to the DNN for the beam SNR task. Figure 14 shows the confusion matrix for the beam SNR task across multiple (3) people and 2 environments. The DNN classifier is trained mainly using gestures from the home environment and a few additional instances from the office environment. The classifier achieves an overall accuracy of 91.5% and per gesture accuracy above 75% across two environments. On the other hand, the DNN still achieves 100% accuracy on the CSI task, which is intuitive from Figures 11 and 13.

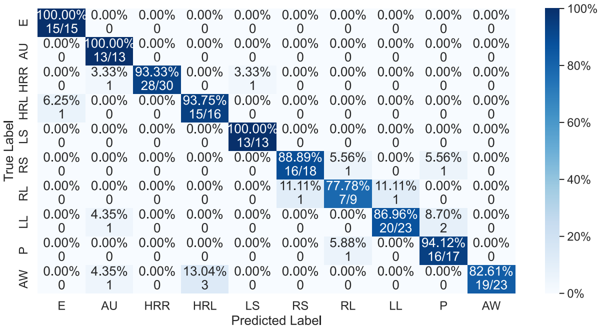

Fig. 14. Re-training beam SNR with gestures from the office environment.

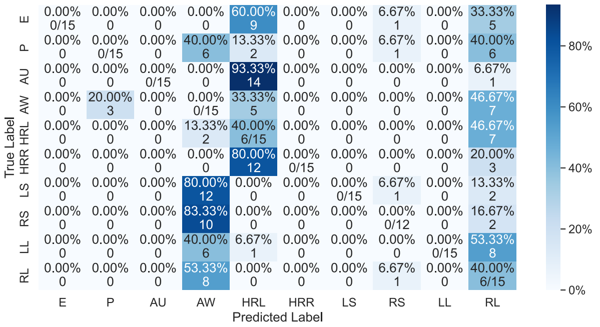

Fig. 15. Generalization of DNN on the CSI task: Train 0° orientation and test 90°.

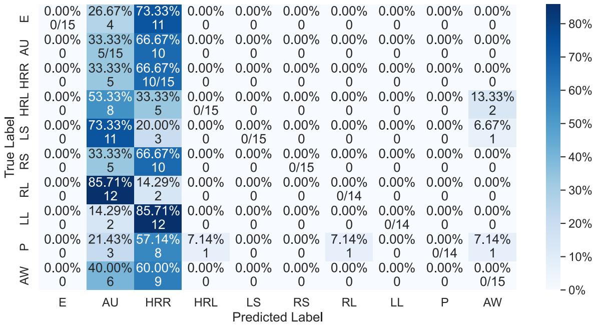

Fig. 16. Generalization of DNN on the beam SNR task: Train 0° orientation and test 90°.

Finally, we test the robustness of the DNN against the orientation of the gesture/pose. The DNN does not generalize across orientation, *i.e.,* a DNN trained in a home environment with gestures performed facing the router performs poorly on the dataset where gestures are performed at 90° orientation. The DNN achieves only 8% and 10% accuracy, as shown in Figures 15 and 16 for the CSI and beam SNR, respectively. Therefore, the DNN network needs additional re-training with the corresponding gestures in this case.

Figure 17 shows the final performance of the DNN for the CSI where the training set consists of examples of gestures performed at 90° orientation in addition to 0° examples from the office and home environment (experiments 1, 2, and 3). The classifier achieves an overall accuracy of 93.9% against 100% in the first two experiments. In contrast, Figure 18 shows the overall performance of the DNN on the beam SNR task. The classifier achieves an overall accuracy of 87% with a per gesture accuracy above 75%. In both cases, we see a drop in performance. Therefore, the orientation of the gesture impacts the performance of the DNN. This can be countered by collecting more data or working at the signal level, some pre-processing to tackle orientation. This is beyond the scope

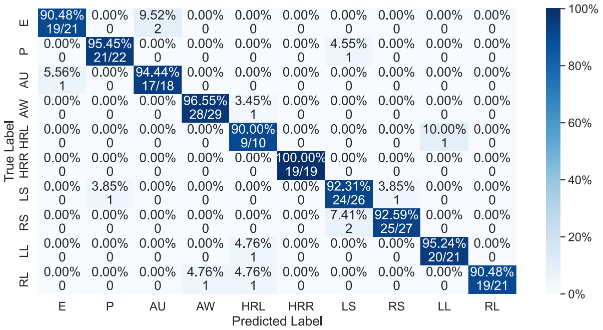

Fig. 17. Performance of the DNN on two environments with the effect of orientation on the CSI task.

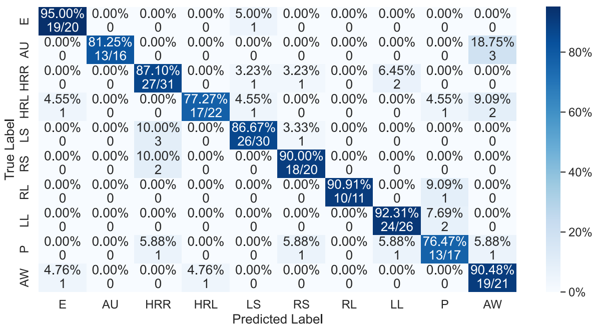

Fig. 18. Performance of the DNN on two environments with the effect of orientation on the beam SNR task.

of this paper.

VI. Conclusion and Future Work

In this work, we validated the performance of the beam SNR for gesture/pose recognition, leveraging a deep neural network (DNN). We showed that even with low sample beam SNRs and a limited dataset, the DNN achieves promising results under challenging gestures relevant to XR applications in two different environments. We achieved state-of-art accuracy of around 96.7% with beam SNRs in a single environment. We also compared beam SNR with CSI and conclude that a DNN trained on CSI achieves better performance and can generalize across environments and different people. We conclude that the environment impacts the beam SNRs significantly more than the CSI. However, concerning the orientation, there is little difference between the two tasks. Nevertheless, with minimal re-training, the DNN achieved very good results across multiple people and environments on the beam SNR task and did not lag much behind the CSI task. Concerning user orientation, the DNN performed poorly on both tasks and needs re-training with a small number of corresponding gestures. Overall, we conclude that mmWave access points can be used for Wi-Fi sensing with reliable accuracy for XR applications.

Our goal is to collect more data with the mmWave devices. Currently, we perform a predefined set of gestures. In our future experiments, we want to collect data with an actual Virtual Reality (VR) headset to make the gestures more natural and take any form. We will extend our data collection to more people, environments, and multi-people sensing with mmWave devices. Moreover, we also want to investigate recent explainable AI approaches to better understand how DNN is making decisions. Currently, CNNs or LSTMs are used for Wi-Fi sensing tasks. However, they are power-hungry, representing a bottleneck when deploying such models in real-world applications. Therefore, we are also working on energy-efficient neuromorphic spiking neural networks (SNNs) with significantly lower energy consumption than CNNs.

Acknowledgment

This research is partly funded by the FWO WaveVR project (Grant number: G034322N).

References

[1] Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu, "E-eyes: device-free location-oriented activity identification using fine-grained Wi-Fi signatures," in *Proceedings of the 20th annual international conference on Mobile computing and networking*, 2014, pp. 617–628.

[2] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu, "Understanding and modeling of Wi-Fi signal based human activity recognition," in *Proceedings of the 21st annual international conference on mobile computing and networking*, 2015, pp. 65–76.

[3] Sameera Palipana, David Rojas, Piyush Agrawal, and Dirk Pesch, "Falldefi: Ubiquitous fall detection using commodity Wi-Fi devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, pp. 1–25, 2018.

[4] Francesca Meneghello, Domenico Garlisi, Nicolò Dal Fabbro, Ilenia Tinnirello, and Michele Rossi, "Sharp: Environment and person independent activity recognition with commodity IEEE 802.11 access points," *IEEE Transactions on Mobile Computing*, 2022.

[5] Heba Abdelnasser, Moustafa Youssef, and Khaled A Harras, "Wigest: A ubiquitous Wi-Fi-based gesture recognition system," in *2015 IEEE conference on computer communications (INFOCOM)*. IEEE, 2015, pp. 1472–1480.

[6] Wenfeng He, Kaishun Wu, Yongpan Zou, and Zhong Ming, "Wig: Wi-Fi-based gesture recognition system," in *2015 24th International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 2015, pp. 1–7.

[7] Sheng Tan and Jie Yang, "Wifinger: Leveraging commodity Wi-Fi for fine-grained finger gesture recognition," in *Proceedings of the 17th ACM international symposium on mobile ad hoc networking and computing*, 2016, pp. 201–210.

[8] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung, "Signfi: Sign language recognition using Wi-Fi," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–21, 2018.

[9] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang, "Zero-effort cross-domain gesture recognition with Wi-Fi," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, 2019, pp. 313–325.

[10] Yiming Wang, Lingchao Guo, Zhaoming Lu, Xiangming Wen, Shuang Zhou, and Wanyu Meng, "From point to space: 3D moving human pose estimation using commodity Wi-Fi," *IEEE Communications Letters*, vol. 25, no. 7, pp. 2235–2239, 2021.

[11] Yili Ren, Zi Wang, Yichao Wang, Sheng Tan, Yingying Chen, and Jie Yang, "Gopose: 3D human pose estimation using Wi-Fi," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, pp. 1–25, 2022.

[12] Liangyi Gong, Wu Yang, Zimu Zhou, Dapeng Man, Haibin Cai, Xiancun Zhou, and Zheng Yang, "An adaptive wireless passive human detection via fine-grained physical layer information," *Ad Hoc Networks*, vol. 38, pp. 38–50, 2016.

[13] Kun Qian, Chenshu Wu, Zheng Yang, Yunhao Liu, Fugui He, and Tianzhang Xing, "Enabling contactless detection of moving humans with dynamic speeds using CSI," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 17, no. 2, pp. 1–18, 2018.

[14] Tianmeng Hang, Yue Zheng, Kun Qian, Chenshu Wu, Zheng Yang, Xiancun Zhou, Yunhao Liu, and Guilin Chen, "Wish: Wi-Fi-based real-time human detection," *Tsinghua Science and Technology*, vol. 24, no. 5, pp. 615–629, 2019.

[15] Kaishun Wu, Jiang Xiao, Youwen Yi, Dihu Chen, Xiaonan Luo, and Lionel M Ni, "CSI-based indoor localization," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 7, pp. 1300–1309, 2012.

[16] Anastasios Foliadis, Mario H Castañeda Garcia, Richard A Stirling-Gallacher, and Reiner S Thomä, "CSI-based localization with CNNs exploiting phase information," in *2021 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2021, pp. 1–6.

[17] Xuyu Wang, Lingjun Gao, Shiwen Mao, and Santosh Pandey, "CSI-based fingerprinting for indoor localization: A deep learning approach," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 1, pp. 763–776, 2016.

[18] Saandeep Depatla and Yasamin Mostofi, "Crowd counting through walls using Wi-Fi," in *2018 IEEE international conference on pervasive computing and communications (PerCom)*. IEEE, 2018, pp. 1–10.

[19] Heba Abdelnasser, Khaled A Harras, and Moustafa Youssef, "Ubibreathe: A ubiquitous non-invasive Wi-Fi-based breathing estimator," in *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2015, pp. 277–286.

[20] Yu Gu, Xiang Zhang, Zhi Liu, and Fuji Ren, "Wi-Fi-based real-time breathing and heart rate monitoring during sleep," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[21] Jin Zhang, Bo Wei, Wen Hu, and Salil S Kanhere, "Wi-Fi-ID: Human identification using Wi-Fi signal," in *2016 International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2016, pp. 75–82.

[22] Jin Zhang, Bo Wei, Fuxiang Wu, Limeng Dong, Wen Hu, Salil S Kanhere, Chengwen Luo, Shui Yu, and Jun Cheng, "Gate-ID: Wi-Fi-based human identification irrespective of walking directions in smart home," *IEEE Internet of Things Journal*, vol. 8, no. 9, pp. 7610–7624, 2020.

[23] Ding Wang, Zhiyi Zhou, Xingda Yu, and Yangjie Cao, "CSIID: Wi-Fi-based human identification via deep learning," in *2019 14th International Conference on Computer Science & Education (ICCSE)*. IEEE, 2019, pp. 326–330.

[24] Karly A Smith, Clément Csech, David Murdoch, and George Shaker, "Gesture recognition using mm-wave sensor for human-car interface," *IEEE sensors letters*, vol. 2, no. 2, pp. 1–4, 2018.

[25] Niklas Peinecke, Hans-Ullrich Doehler, and Bernd R Korn, "Phong-like lighting for mmw radar simulation," in *Millimetre Wave and Terahertz Sensors and Technology*. SPIE, 2008, vol. 7117, pp. 173–182.

[26] Carlos De Lima, Didier Belot, Rafael Berkvens, Andre Bourdoux, Davide Dardari, Maxime Guillaud, Minna Isomursu, Elena-Simona Lohan, Yang Miao, Andre Noll Barreto, et al., "Convergent communication, sensing and localization in 6G systems: An overview of technologies, opportunities and challenges," *IEEE Access*, vol. 9, pp. 26902–26925, 2021.

[27] Jaime Lien, Nicholas Gillian, M Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–19, 2016.

[28] Yuwei Ren, Jiuyuan Lu, Andrian Beletchi, Yin Huang, Ilia Karmanov, Daniel Fontijne, Chirag Patel, and Hao Xu, "Hand gesture recognition using 802.11ad mmwave sensor in the mobile device," in *2021 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*. IEEE, 2021, pp. 1–6.

[29] Jianyuan Yu, Pu Wang, Toshiaki Koike-Akino, Ye Wang, Philip V Orlik, and Haijian Sun, "Human pose and seat occupancy classification with commercial mmwave Wi-Fi," in *2020 IEEE Globecom Workshops (GC Wkshps*. IEEE, 2020, pp. 1–6.

[30] Cristina Perfecto, Mohammed S Elbamby, Jihong Park, Javier Del Ser, and Mehdi Bennis, "Mobile XR over 5G: A way forward with mmwaves and edge," *arXiv preprint arXiv:1905.04599*, 2019.

[31] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall, "Tool release: Gathering 802.11 n traces with channel state information," *ACM SIGCOMM computer communication review*, vol. 41, no. 1, pp. 53–53, 2011.

[32] Mohammed Abdulaziz Aide Al-qaness and Fangmin Li, "Wiger: Wi-Fi-based gesture recognition system," *ISPRS International Journal of Geo-Information*, vol. 5, no. 6, pp. 92, 2016.

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

[34] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios 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.

[35] Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Nurmi, and Zheng Wang, "Crosssense: Towards cross-site and large-scale Wi-Fi sensing," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018, pp. 305–320.

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

[37] Yili Ren, Zi Wang, Sheng Tan, Yingying Chen, and Jie Yang, "Winect: 3D human pose tracking for free-form activity using commodity Wi-Fi," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–29, 2021.

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

[39] Toshiaki Koike-Akino, Pu Wang, and Ye Wang, "Quantum transfer learning for Wi-Fi sensing," *arXiv preprint arXiv:2205.08590*, 2022.

[40] Matthias Schulz, Daniel Wegemer, and Matthias Hollick, "Nexmon: The c-based firmware patching framework. 2017," *URl: https://nexmon. org*, 2017.

[41] Yongsen Ma, Gang Zhou, and Shuangquan Wang, "Wi-Fi sensing with channel state information: A survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 3, pp. 1–36, 2019.

[42] Daniel Steinmetzer, Daniel Wegemer, and Matthias Hollick, "Talon tools: The framework for practical IEEE 802.11ad research," 2017.

[43] Daniel Steinmetzer, Daniel Wegemer, Matthias Schulz, Joerg Widmer, and Matthias Hollick, "Compressive millimeter-wave sector selection in off-the-shelf IEEE 802.11ad devices," in *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies*, 2017, pp. 414–425.

[44] Fei Wang, Jianwei Feng, Yinliang Zhao, Xiaobin Zhang, Shiyuan Zhang, and Jinsong Han, "Joint activity recognition and indoor localization with Wi-Fi fingerprints," *IEEE Access*, vol. 7, pp. 80058–80068, 2019.

[45] Linlin Guo, Lei Wang, Chuang Lin, Jialin Liu, Bingxian Lu, Jian Fang, Zhonghao Liu, Zeyang Shan, Jingwen Yang, and Silu Guo, "Wiar: A public dataset for Wi-Fi-based activity recognition," *IEEE Access*, vol. 7, pp. 154935–154945, 2019.

[46] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[47] Jörg Schäfer, Baldev Raj Barrsiwal, Muyassar Kokhkharova, Hannan Adil, and Jens Liebehenschel, "Human activity recognition using CSI information with nexmon," *Applied Sciences*, vol. 11, no. 19, pp. 8860, 2021.