Teaching RF to Sense without RF Training Measurements

HONG CAI*, University of California Santa Barbara
BELAL KORANY*, University of California Santa Barbara
CHITRA R. KARANAM*, University of California Santa Barbara
YASAMIN MOSTOFI, University of California Santa Barbara

In this paper, we propose a novel, generalizable, and scalable idea that eliminates the need for collecting Radio Frequency (RF) measurements, when training RF sensing systems for human-motion-related activities. Existing learning-based RF sensing systems require collecting massive RF training data, which depends heavily on the particular sensing setup/involved activities. Thus, new data needs to be collected when the setup/activities change, significantly limiting the practical deployment of RF sensing systems. On the other hand, recent years have seen a growing, massive number of online videos involving various human activities/motions. In this paper, we propose to translate such already-available online videos to instant simulated RF data for training any human-motion-based RF sensing system, in any given setup. To validate our proposed framework, we conduct a case study of gym activity classification, where CSI magnitude measurements of three WiFi links are used to classify a person's activity from 10 different physical exercises. We utilize YouTube gym activity videos and translate them to RF by simulating the WiFi signals that would have been measured if the person in the video was performing the activity near the transceivers. We then train a classifier on the simulated data, and extensively test it with real WiFi data of 10 subjects performing the activities in 3 areas. Our system achieves a classification accuracy of 86% on activity periods, each containing an average of 5.1 exercise repetitions, and 81% on individual repetitions of the exercises. This demonstrates that our approach can generate reliable RF training data from already-available videos, and can successfully train an RF sensing system without any real RF measurements. The proposed pipeline can also be used beyond training and for analysis and design of RF sensing systems, without the need for massive RF data collection.

CCS Concepts: • **Computing methodologies** → **Modeling and simulation**; **Machine learning**; • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Hardware** → *Wireless devices*;

Additional Key Words and Phrases: WiFi, RF Sensing, RF Simulation, RF Signal Processing, Machine Learning for RF Sensing, Human Activity Recognition

ACM Reference Format:
Hong Cai, Belal Korany, Chitra R. Karanam, and Yasamin Mostofi. 2020. Teaching RF to Sense without RF Training Measurements. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 120 (December 2020), 22 pages. https://doi.org/10.1145/3432224

*Joint first authors.

Authors' addresses: Hong Cai, hcai@ece.ucsb.edu, University of California Santa Barbara, Santa Barbara, CA; Belal Korany, belalkorany@ece.ucsb.edu, University of California Santa Barbara, Santa Barbara, CA; Chitra R. Karanam, ckaranam@ece.ucsb.edu, University of California Santa Barbara, Santa Barbara, CA; Yasamin Mostofi, ymostofi@ece.ucsb.edu, University of California Santa Barbara, Santa Barbara, CA.

2474-9567/2020/12-ART120 $15.00
https://doi.org/10.1145/3432224

1 INTRODUCTION

Recent years have witnessed a rapidly-growing presence of connected wireless devices in our everyday life, such as laptops, mobile phones, and smart speakers. Consequently, we are surrounded by the Radio Frequency (RF) signals emitted by these devices. As people move around in these environments, the RF signals are perturbed and thus implicitly capture information about the people. This has opened up new possibilities to passively, and non-intrusively, learn various characteristics of the people and their behaviors in the environment, using RF signals. Towards this end, researchers have been extensively exploring the various possibilities created by RF sensing in recent years, including person identification, gesture recognition, activity classification, and fall detection [17, 18, 24, 31, 32, 41].

In the RF sensing literature, most recent state-of-the-art systems utilize learning-based techniques, such as neural network [17], support vector machine [30], and deep learning [12], to enable a variety of sensing applications. Such methods, however, use a large number of RF measurements to train the system, which requires a laborious prior RF data collection process. In spite of the large-scale training measurements, the performance of these systems degrades considerably when operating with a new setup (e.g., new transceiver placement) or in an area that differs from the setup/area of the training phase [9, 31]. While there are a few recent studies on mitigating the environment dependency [12, 41], these methods still require extensive training measurements and are constrained by the training-phase system setup. Furthermore, for classification-related tasks, existing systems are not scalable, as they cannot be used for classes not seen during training, requiring additional RF training measurements for the new classes.

In this paper, we propose a novel, **generalizable and scalable** framework that enables human-motion-related RF sensing applications, **without the need to collect any RF training measurements**, thus allowing one to efficiently develop RF sensing systems for several different applications and with various configurations. Here is the underlying proposed idea. Recent years have seen a tremendous growth in the area of vision, resulting in many publicly-available videos of people involved in a variety of activities. In this paper, we propose to use such available videos for training RF sensing systems, thus eliminating the need to collect any RF training measurement (or video training data). This allows us to tap into the massive publicly-available online videos of people's motions and activities to generate the required RF training data for any motion-based RF sensing application. As such, our proposed framework makes it possible to train an RF sensing system with no RF training measurements. Furthermore, our proposed framework is flexible to accommodate changes in the task and RF sensing setup, e.g., new classes, different transceiver placements, and different frequencies of operation, since the RF training data can simply be re-generated from the videos according to the new specifications of the task and/or sensing setup. It can furthermore be used for analysis purposes, for instance to understand the amount of resources needed for a particular application, to understand the differentiability of different activities, or to understand the limitations of sensing with a certain setup configuration.

In order to validate our proposed methodology, we implement our pipeline for a realistic WiFi sensing application of gym activity classification. In this case study, the system is trained with the WiFi data generated from YouTube videos of people performing the gym activities. We then extensively test the trained system with real WiFi data of people performing the gym activities in different areas, and show that it classifies the activities with a high accuracy. To the best of our knowledge, this is the first time that a real-world RF sensing system is enabled with only video-based training data, and without any real RF measurements for training purposes. Next, we summarize the main contributions of this paper.

Statement of Contributions:

1. In this paper, we propose a novel idea: to train a human-motion-related RF sensing system without any training data, through leveraging the vast amount of online available videos. More specifically, we propose to translate the massive readily-available online videos of people's motions and activities to instant RF data and use them to train

RF sensing systems. This idea eliminates the labor-intensive process of collecting RF measurements for training, is **scalable** and is **applicable to any motion-based RF sensing system**. It further allows for easy re-training when new classes and/or a different RF sensing setup are given. Finally, it can enable new possibilities beyond system design and towards system analysis, in order to understand the fundamental limitations of a particular RF sensing system, as a function of the underlying activities, the amount of given resources, and the setup. Overall, researchers can utilize this idea to train RF sensing systems with no RF data collection.

2. Our pipeline consists of the following steps. In order to build a classifier pertaining to a number of motion-related activities, we first gather several online videos of different instances of the activities of interest. We then utilize a state-of-the-art vision-based human shape reconstruction algorithm to build a 3D mesh of the person in each video, as a function of time. Here we have to address a number of challenges. The person in an online video, for instance, could have been captured from any viewpoint, resulting in an unknown coordinate system, or could be doing a variety of different motion-related activities. Our proposed 3D mesh alignment technique via eigen-analysis then enables proper mesh extraction and positioning. We then simulate the corresponding RF signal that would have been measured if this person was in an RF-covered area, by modeling the interaction between the extracted human mesh and the electromagnetic waves propagating in the environment. In doing so, we take into account the needed sensing setup, e.g., the locations of the transceivers relative to the person and the frequency of operation. Once we translate all the videos to the RF domain, we perform time-frequency analysis on the generated RF data and extract key features from the corresponding RF spectrograms. We then use the extracted features to train a neural network to classify the underlying activities. This RF sensing system that is solely trained on available online videos is then ready for testing in any RF area/setting.

3. We demonstrate the efficacy of our proposed video-based training framework by implementing it for a realistic WiFi sensing application of gym activity classification using only WiFi CSI magnitude measurements of a small number of links. In this case study, we consider 10 different gym activities and generate a corresponding WiFi training dataset only from YouTube videos of people performing these activities. We then use our pipeline to train a WiFi-based gym activity classifier using only the video dataset. We extensively test the trained system in 3 real-world WiFi test areas, where 10 subjects are recruited to perform the activities. We achieve an accuracy of 86% in correctly classifying the gym activity when using a small period of the activity, which may contain a few repetitions of the same exercise (5.1 on average), and an accuracy of 81% when only considering one repetition. Overall, this demonstrates that our proposed idea can eliminate the labor-intensive collection of RF training measurements and enable training RF sensing systems with only already-available video data. It further shows the first demonstration of WiFi-based gym activity classification without any RF training data.

Remark 1. *The proposed video to RF pipeline can, in particular, be useful for times such as during the current COVID-19 pandemic, where it is hard to collect human-related RF measurements due to social distancing.*

2 RELATED WORK

In recent years, there has been a large body of work that uses RF signals for sensing human motion, activity, and behavior, in order to enable various useful applications, such as person identification, vital signs detection, fall detection, and activity recognition [17, 22, 33, 34, 42]. In terms of activity recognition, great progress has recently been made towards enabling device-free RF sensing. Several papers have utilized specialized hardware or radar for activity recognition. For instance, [3, 21, 24] use USRPs and/or radars to classify daily human actions. [18] uses an FMCW radar to capture the human pose. There has further been considerable interest in utilizing off-the-shelf WiFi devices to perform human activity recognition. For instance, CARM [31] utilizes the relationship between the WiFi signal variations and the speeds of human body parts for classifying a set of 8 activities. [35] achieves activity recognition through-walls on a set of 7 daily activities. [12, 28] utilize deep learning to classify various daily activities. In terms of classifying gym activities, such as push-up and jumping jack, there is a limited number

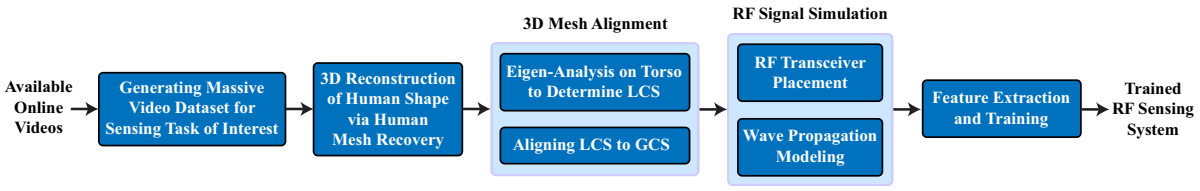

Fig. 1. Various steps involved in our proposed framework for training RF sensing systems solely based on video data. LCS stands for Local Coordinate System, while GCS stands for Global Coordinate System.

of work [9, 36, 39] that have used RF sensing systems to enable this application, for sets of 4, 9, and 10 physical exercises, respectively, by relying on line-of-sight crossing (i.e., blocking the direct path between the Tx and Rx).

All of these state-of-the-art methods on different aspects of activity recognition, however, need to collect a large amount of prior RF measurements for training purposes. Such data collection typically involves asking several people to perform all the designated actions in different areas. Furthermore, the collected training data is based on a specific experimental setup and a specific set of activities, which cannot be reused for other setups or applications with different activities. [18] has further collected real video measurements of the scene, simultaneous with RF training measurement collection, in order to annotate the collected RF data for the purpose of pose estimation. [8] has recently released their collected RF training dataset for WiFi-based action recognition. While releasing collected data could be useful for the research community, such a dataset is heavily dependent on the activities involved and is not generalizable to other activities/behaviors. Furthermore, the released data is heavily dependent on the sensing setup, such as the number of transceivers, their placement, and the frequency of operation. Thus, this dataset cannot be used for a sensing system with a different configuration. XModal-ID [17] considers the problem of WiFi-based person identification from a given video footage, thus providing identification across modalities. More specifically, it takes as input a video and a WiFi measurement of a walking person and determines if they belong to the same person by using gait as a unique identifier of a person. As such, it is on a different idea than this paper. In one of its steps, XModal-ID extracts the human mesh from a video of a walking person and simulates the corresponding RF signal. This successful translation is thus an inspiration for our RF simulation part. However, we should note that the overall idea and methodology of this paper is considerably different from XModal-ID. In addition, XModal-ID requires RF training data (and real video data) to train its pipeline, and thus, it does not eliminate the laborious effort of RF training measurement collection. Finally, its RF simulation step is specific to videos that are recorded from the side view of a person walking and cannot be used for existing online videos that are captured from other views and/or involve other activities, which are issues that we shall also address as part of our framework.

In summary, in this paper, we propose a general framework for training an RF sensing system, without the need to collect any RF measurements. Our proposed idea leverages the massive available online video data, pertaining to different human activities, motions, and behaviors, and shows how they can be translated to instant RF training data. As such, our proposed framework eliminates the need for RF measurement collection for training an RF sensing system. To the best of our knowledge, our proposed framework is the first to enable training an RF sensing system only with already-available video data, and without the need to collect any RF training data. The proposed idea is thus general and scalable. While we showcase its effectiveness in the context of gym activity classification, researchers can utilize this approach to enable other RF sensing applications, as part of future work.

3 VIDEO-BASED TRAINING FOR RF SENSING SYSTEMS

In this section, we describe our proposed general framework for translating the already-available video content to the RF domain, in order to generate RF data for training human-motion-based RF sensing systems. More specifically, given the video footage of a person performing an action (e.g., walking, running, physical exercises),

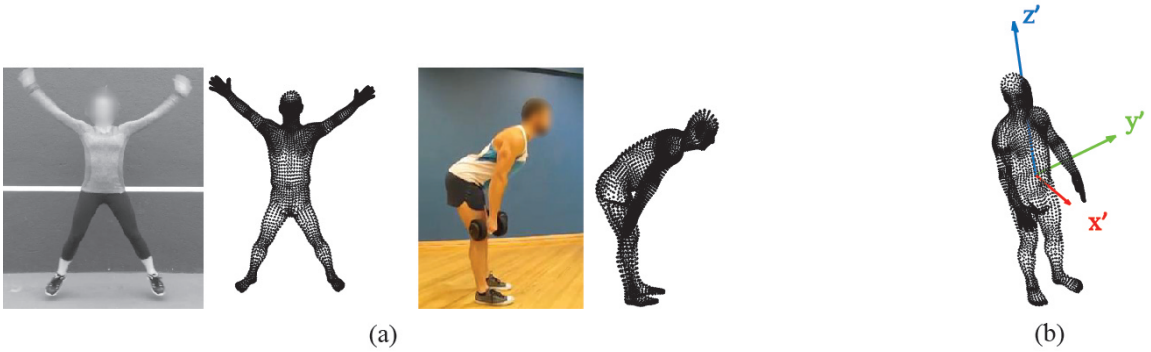

Fig. 2. (a) Sample reconstructed 3D human mesh from a snapshot of (left) a person doing jumping jacks and (right) a person doing stiff-leg deadlifts. (b) The Local Coordinate System (LCS), defined with respect to the human body.

our framework allows one to simulate the RF signal that would have been measured if this person performed the action in the vicinity of RF transceivers. As such, our proposed framework makes it possible to tap into the virtually unlimited publicly-available online videos to generate wireless data and construct a massive simulated RF dataset for training RF sensing systems for various applications.

Fig. 1 summarizes the key steps of our proposed framework. We first collect a massive video dataset pertaining to the RF sensing task of interest, using available online videos. For each video of a person doing an activity, our pipeline first extracts a 3D mesh of this person as a function of time. This extracted 3D human model is then transformed into a Global Coordinate System (GCS) that is independent of the camera view using eigen analysis. Then, given the desired RF sensing configuration (e.g., transceiver positions, frequency), our framework simulates the RF signals that would have been measured if the extracted human mesh was in the given RF sensing setup, via efficient wave propagation modeling. Through time-frequency analysis, we then extract key features from this simulated RF dataset and train an RF sensing system, which will then be deployed in a real wireless environment during operation. We next discuss each of these components in detail.

3.1 3D Reconstruction of Human Shape

The first step towards translating video content to the wireless domain is to build an accurate 3D shape of the person in a video frame. In order to do so, we utilize the recent advances in computer vision [14, 23, 26, 27]. For instance, [14] proposes a Human Mesh Recovery (HMR) algorithm that uses a convolutional encoder network and a regression module to infer information about the pose and the shape of a person from a single 2D image. Then, they utilize the Skinned Multi-Person Linear (SMPL) model [20] to translate the inferred body pose and shape information into a *human mesh*. An SMPL human mesh is a set of triangulated points in 3D space describing the body surface. More specifically, the extracted 3D human mesh is characterized by a dense set of mesh points (e.g., 6890 points) that describe, in 3D, the outer surface of the human body in the video. Fig. 2 (a) shows examples of our reconstructed 3D human mesh from a snapshot of a person doing jumping jacks and stiff-leg deadlifts, using the algorithm of [14]. In the first step of our approach, we then extract the 3D human shape from the video, using the state-of-the-art mesh reconstruction algorithms in computer vision. It is noteworthy that such a human mesh recovery algorithm has recently been used in XModal-ID [17] for walking people, in order to provide cross-modal identification.

3.2 3D Mesh Alignment via Eigen-analysis

The goal of our proposed framework is to simulate the RF signals that would have been measured by one or more RF transceivers if the person in the video was in their vicinity, given any RF *sensing setup*. By sensing setup,

we mean the relative location and orientation of the person with respect to the RF transceivers, as well as the relative locations among the transceivers. We thus need to correctly place the transceivers with respect to the extracted human mesh in the simulation environment, in order to generate a given sensing setup, which requires translating the mesh to a global coordinate system.

In the reconstructed 3D human mesh from the output of a vision algorithm (e.g., [14]), the coordinates of the 3D points are calculated based on the camera view, which is estimated from the given video frame. In other words, the reconstructed mesh may reside in different coordinate systems across different videos, when the videos are shot from different views. Therefore, it is essential to transform the 3D mesh into a Global Coordinate System (GCS) that is invariant to the camera view. Such a transformation allows us to arbitrarily choose the orientation and the position of the human mesh in the target GCS, as well as place the RF transceivers in the GCS as needed for the sensing setup. In order to enable this transformation, however, we need to know some information about the orientation of the body parts. We next show how to achieve this through an eigen-analysis on the mesh points of the torso.

Let the set of 3D points of the extracted human mesh at time t be $\mathcal{M}'(t) = \{\mathrm{p}'_m(t), m \in \{1, \ldots, M\}\}$, where $\mathrm{p}'_m(t) \in \mathbb{R}^3$ is the 3D location of the m-th point at time t and M is the total number of points in the mesh. These points are given in a coordinate system where the x-y plane is parallel to the camera plane. We define a Local Coordinate System (LCS) with respect to the human body, where the x-axis points to the front of the person, the y-axis points to the person's left, and the z-axis points upward, as shown in Fig. 2 (b). The axes of the LCS are denoted by x', y', and z', respectively, and are determined based on the set of mesh points $\mathcal{M}'$ as follows. Let $\mathrm{H} = [\mathrm{p}'_{s_1}, \mathrm{p}'_{s_2}, \ldots, \mathrm{p}'_{s_{M_T}}]$ be a $3 \times M_T$ matrix of the 3D locations of all the mesh points belonging to the torso in the original coordinate system, where $s_1, \ldots, s_{M_T}$ are the indices of the torso points among the M total mesh points and M_T is the total number of torso points.[1] Since the anterior-posterior (i.e., front-back) is the smallest dimension of the human torso, x' is the eigenvector of $\mathrm{HH}^\intercal$ that corresponds to the smallest eigenvalue, where $^\intercal$ is the transpose operator. Similarly, z' is the eigenvector of $\mathrm{HH}^\intercal$ that corresponds to the largest eigenvalue, since the inferior-superior (i.e., bottom-top) is the largest dimension of the human torso.

Once we have determined the axes x', y', and z' of the LCS corresponding to the human mesh of a video frame, the human mesh can be rotated such that the person faces any desired direction in the GCS. For instance, consider the case where the person is required to face the positive x-axis in the GCS in the simulation, we multiply all the original mesh points by the rotation matrix $\mathrm{R} = [\mathrm{x}', \mathrm{y}', \mathrm{z}']$, i.e., $\mathrm{p}_m(t) = \mathrm{R}^\intercal \mathrm{p}'_m(t)$. It is straightforward to show that after this operation, the LCS of the rotated mesh points $\mathrm{p}_m(t)$ aligns with the x, y, and z axes of the GCS. The mesh points can also be easily translated in the new GCS to any arbitrary location. For example, if the x-y plane in the GCS is assumed to be the floor and the person is in a standing position, we can translate the mesh points such that the feet points lie on the x-y plane.

In general, depending on the configuration of the RF sensing setup, such as the transceiver positions relative to the person, the human mesh can be put into any desired orientation and location in the GCS. Concrete examples will be provided in our case study of Sec. 4.

3.3 RF Signal Simulation

Let $\mathcal{M}(t) = \{\mathrm{p}_m(t), m \in \{1, \ldots, M\}\}$, where $\mathrm{p}_m \in \mathbb{R}^3$ is the 3D location of the m-th point of the human mesh in the GCS, where the values of $\mathrm{p}_m(t)$ were calculated such that the human model has a specific location and orientation in the GCS, as described previously. Let p_T and p_R denote the locations of the RF transmitter (Tx) and receiver (Rx) in the GCS, respectively, in the sensing setup of interest. We then simulate the RF signal that would have been measured from reflections off of the extracted human mesh in this setup.

[1] The indices of the mesh points belonging to any specific body part are fixed and known for all meshes generated by a human mesh recovery algorithm.

In general, the received electric field at the Rx, $E(\mathrm{p}_R)$, due to a transmission from the Tx, is the solution to the volume integral equation [5]:

$$E(\mathrm{p}_R) = \mathrm{g}(\mathrm{p}_R, \mathrm{p}_T) + \iiint_{\mathrm{D}} \mathrm{g}(\mathrm{p}_R, \mathrm{p}) O(\mathrm{p}) E(\mathrm{p}) d\mathrm{p}, \tag{1}$$

where D is the workspace where the transceivers and the person are located, $\mathrm{g}(\mathrm{p}_1, \mathrm{p}_2) = \frac{\exp(j\frac{2\pi}{\lambda}\|\mathrm{p}_1-\mathrm{p}_2\|)}{4\pi\|\mathrm{p}_1-\mathrm{p}_2\|}$ is the Green's function from point p_2 to point p_1, where $\|.\|$ denotes the Euclidean distance and λ is the wavelength of the wireless signal. $O(\mathrm{p})$ is a parameter that captures the electric/magnetic properties of what resides at position p in the area. In the electromagnetics literature, several methods have been proposed for finding the solution to Eq. 1, such as the Method of Moments (MoM) [7] and the Finite Element Method (FEM) [13]. However, they are very computationally intensive. Thus, efficient linearizing approximations to Eq. 1 have been proposed, such as the Born approximation [4]. Based on our extensive studies (reported in detail in Sec. 4.4.1), Born approximation well approximates the details of the received signal with the accuracy needed for training based on the simulated RF data, while being computationally very efficient. In Born approximation, we have the following for the received signal:

$$E(\mathrm{p}_R) = \mathrm{g}(\mathrm{p}_R, \mathrm{p}_T) + \iiint_{\mathrm{D}} \mathrm{g}(\mathrm{p}_R, \mathrm{p}) O(\mathrm{p}) \mathrm{g}(\mathrm{p}, \mathrm{p}_T) d\mathrm{p}, \tag{2}$$

where the total electric field at point p, $E(\mathrm{p})$, inside the integral of Eq. 1 is approximated by the Green's function from the Tx to point p. This means that the Born approximation treats each point in space as an independent scatterer, and does not take into account the higher-order scattering effects.

In summary, given a video of a person engaged in some activity, we first extract a dense set of 3D mesh points describing the outer human surface. We then transform the reconstructed 3D human mesh into a GCS, which can be put into any arbitrary desired location and orientation in the GCS. Then, based on the given sensing setup, we determine the Tx and Rx positions in the GCS, and simulate the RF signal that would have been measured using the approximated Born wave model of Eq. 2. More advanced wave models (e.g., Eq. 1) can certainly be used as part of the proposed pipeline if more details are needed for a particular application, at the cost of higher computational complexity. Furthermore, given a different RF sensing setup, we can easily re-run the simulation to obtain the corresponding RF signal, by changing the transceiver locations or the orientation and location of the already-aligned human mesh in the GCS, according to the new setting. As we shall see in Sec. 4, the proposed pipeline can generate realistic RF data for the purpose of training an RF sensing system.

3.4 Feature Extraction and Training

Once the RF signals are simulated based on the desired scenario, they can be used to train a machine learning algorithm for a given RF sensing application. For traditional machine learning algorithms, such as support vector machine and neural network, one can extract several features from the simulated RF signals for training the system. Since our proposed framework enables the generation of massive RF training data, it is also possible to utilize our framework to generate sufficient data for training deep learning algorithms.

Overall, our proposed scalable and general framework enables training RF sensing systems without the need for collecting any real RF training data, and by translating the vast available video data to the RF domain. Given the generated RF training data, one can then apply any machine learning algorithm to train the RF sensing system. In the next section, we then showcase the possibilities created by the proposed framework, in the context of RF sensing for gym activity recognition.

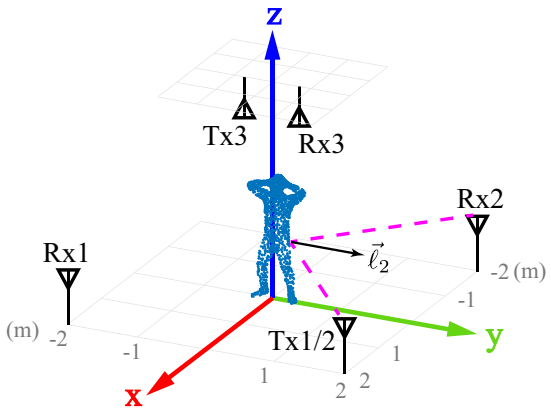

Fig. 3. The sensing setup used for our case study of gym activity classification. The 3 WiFi links capture the velocity components of different body parts along the 3 dimensions.

4 CASE STUDY: GYM ACTIVITY CLASSIFICATION

In this section, we demonstrate the efficacy of our proposed framework with a real-world application of WiFi-based activity recognition. More specifically, we consider gym activity classification, the objective of which is to identify the performed activity from a set of several different physical exercises, such as push-up and sit-up, using only WiFi CSI magnitude measurements of a small number of links. RF-based gym activity classification is a challenging problem due to the complex movements involving different parts of the body, which has only been explored sparsely in the literature [9, 36, 39]. However, all such existing work on gym activity classification require a significant effort in collecting massive wireless training measurements for the set of activities that they want to classify, using the same RF-sensing setup that will be used during the operation phase.

In contrast to these existing papers, we show how to train a WiFi-based gym activity classifier without the need for collecting any wireless training measurements. As discussed in Sec. 3, our proposed framework enables the translation of the video content of a human activity into the wireless domain, which allows us to utilize the available online videos to create an instant RF training dataset. In this section, we then discuss our WiFi sensing setup for gym activity classification, as well as how we implement the different steps of our proposed framework for this real-world application.

4.1 Sensing Setup

In this case study, the WiFi sensing system is tasked with capturing the characteristics of different gym activities to perform classification. As such, we consider a sensing setup that is capable of measuring the motion profile of each activity, which can then serve as the signature for classification.

Consider the setup of Fig. 3, with the coordinate system as marked, where the person is located at the origin, facing the positive x direction, and the positive z-axis is pointing upward. Our WiFi sensing system consists of 3 links, each with a pair of transmitter (Tx) and receiver (Rx). Link 1 is placed in front of the person and parallel to the y-axis, Link 2 is placed to the left of the person and parallel to the x-axis, while Link 3 is placed above the person and parallel to the y-axis, as shown in Fig. 3. Links 1 and 2 share the same Tx, which we denote by Tx1/2. As the person performs the activity, the WiFi signal emitted by the Tx bounces off of the person's body, as well as the static objects in the environment, and is received by the Rx, for each link. The baseband WiFi signal

received by the Rx of the i-th link can be approximated as follows:

$$r_i(t) \approx r_i^{\circ} + \sum_s r_i^s + \sum_n \underbrace{\alpha_n e^{j\frac{4\pi}{\lambda}\langle \vec{\mathrm{v}}_n(t)\cdot\vec{\ell}_i\rangle t}}_{\substack{\text{reflected signal off}\\ \text{the n-th body part}}}, \tag{3}$$

where r_i° is the direct signal from the Tx to the Rx of the i-th link, r_i^s is the signal of s-th static path (path reflected off of a static object in the environment), α_n is the amplitude of the reflected path off of the n-th body part, $\vec{\mathrm{v}}_n(t)$ is the 3D velocity vector of the n-th body part at time t, $\vec{\ell}_i$ is a unit vector bisecting the angle between the two lines connecting the n-th body part to the Tx and Rx of the i-th link, $\langle * \cdot * \rangle$ is the inner product of the vector arguments, and λ is the wavelength of the RF signal. We are interested in gym activity recognition using only WiFi CSI magnitude measurements. In practice, the direct path from the Tx to the Rx is stronger than all reflected paths (i.e. $|r_i^{\circ}| \gg |r_i^s|$, $|r_i^{\circ}| \gg \alpha_n$) due to their longer lengths and the reflection losses. Hence, the squared signal magnitude can be written as [15],

$$\begin{aligned} |r_i(t)|^2 &= P + \sum_n 2|r_i^{\circ}|\alpha_n \cos\left(\frac{4\pi}{\lambda}\langle\vec{\mathrm{v}}_n(t)\cdot\vec{\ell}_i\rangle t - \angle r_i^{\circ}\right) + \sum_n\sum_s 2\alpha_n|r_i^s|\cos\left(\frac{4\pi}{\lambda}\langle\vec{\mathrm{v}}_n(t)\cdot\vec{\ell}_i\rangle t - \angle r_i^s\right) \\ &\quad + \sum_n\sum_{n'>n} 2\alpha_n\alpha_{n'}\cos\left(\frac{4\pi}{\lambda}\langle(\vec{\mathrm{v}}_n(t) - \vec{\mathrm{v}}_{n'}(t))\cdot\vec{\ell}_i\rangle t\right) \\ &\approx P + \sum_n 2|r_i^{\circ}|\alpha_n\cos\left(\frac{4\pi}{\lambda}\langle\vec{\mathrm{v}}_n(t)\cdot\vec{\ell}_i\rangle t - \angle r_i^{\circ}\right), \end{aligned} \tag{4}$$

where $P = |r_i^{\circ}|^2 + \sum_s\sum_{s'}|r_i^s||r_i^{s'}|e^{j\left(\angle r_i^s - \angle r_i^{s'}\right)} + \sum_s 2|r_i^s||r_i^{\circ}|\cos\left(\angle r_i^s - \angle r_i^{\circ}\right) + \sum_n \alpha_n^2$ is the DC component of $|r_i(t)|^2$, $\angle$ denotes the phase of the signal, and the approximation in the last line of Eq. 4 is due to the fact that the direct path is stronger than the reflected ones. It can be seen from Eq. 4 that the static multipath in the environment affects the DC component of the received signal, which does not carry any motion information (i.e., information on $\vec{\mathrm{v}}_n(t)$), and can be easily subtracted in practice.

Given the sensing setup of Fig. 3, we can see that $\vec{\ell}_1$, $\vec{\ell}_2$, and $\vec{\ell}_3$ can be approximated by $[1,0,0]^\intercal$, $[0,1,0]^\intercal$, and $[0,0,1]^\intercal$, respectively. For instance, Fig. 3 shows $\vec{\ell}_2$ in our setup, which, as can be seen, is approximately a unit vector along y. $\langle\vec{\mathrm{v}}_n\cdot\vec{\ell}_i\rangle$, $\forall i = 1,2,3$, are then the 3 components of the velocity vector $\vec{\mathrm{v}}_n$ in the 3D space along the three directions of the Cartesian coordinate axes. Then, based on Eq. 4, the frequency content of the received signals at the 3 links will directly capture the velocity components of different body parts along the x, y, and z directions, respectively. As such, the received signals at the 3 links contain key information on the person's 3D motion profile, which will be very useful for the classification task.[2]

We next demonstrate how to translate relevant online video data to WiFi training data using our proposed framework, for the gym activity classification task. We start by describing our implementation of the various steps shown in Fig. 1, tailored for the gym activity recognition and for the sensing setup of Sec. 4.1.

4.2 Training with Zero RF Training Data

In this case study, the gym activity classification includes 10 different physical exercises, such as jumping jack and push-up, as shown in Fig. 4. They are representative of a variety of typical workouts that involve the movements of different body parts.

[2]Other configurations of the links can be translated to the setup of Fig. 3 by projecting their measured motion information along the 3 axes in Fig. 3, as we shall discuss in Sec. 5.

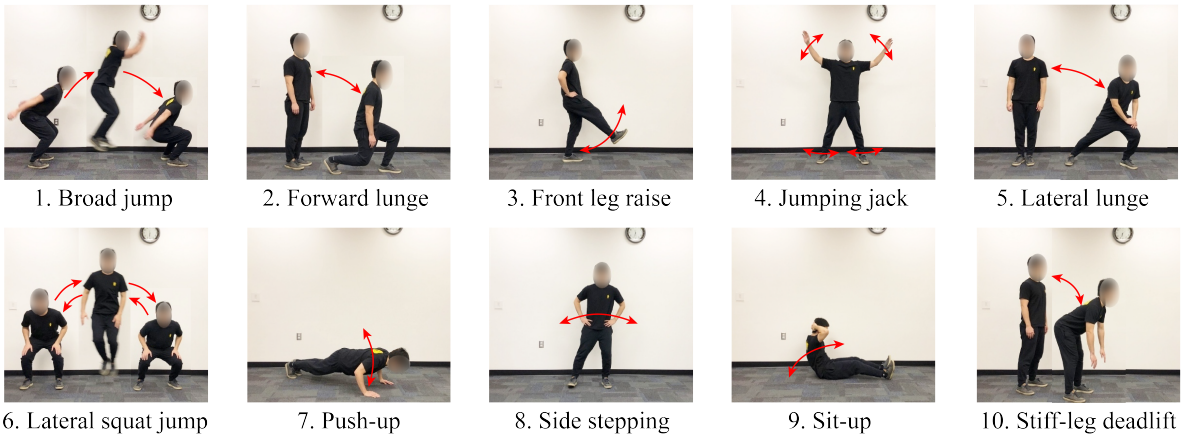

Fig. 4. The set of 10 activities considered in our gym activity classification case study. In our study, the activities are performed with only body weight, i.e., without any equipment (e.g., resistance band, dumbbell). These images show the movements involved in each activity, as indicated by the red arrows. See the color pdf for optimal viewing.

We download the training videos for these activities from YouTube.[3] For each video, we manually identify and extract the overall chunk in which the person is actually performing the corresponding activity (to remove chunks where, for instance, the person is just talking). Although this is done manually in the current study, it is possible to use recent computer vision techniques to temporally localize such activity periods automatically [1, 38]. In order to best facilitate the 3D mesh extraction, we do not use videos in which there is blockage of the exercising person by the surrounding objects. This is not a restrictive requirement since most online videos of gym activities provide unobstructed views of the person.[4] Our training video dataset contains a total of 61 videos of gym activities, each of which containing an average of 8.39 sec of relevant activity content.

4.2.1 Activity Repetition Segmentation. In our WiFi-based gym activity classification system, a repetition is taken as the atomic unit to describe the gym activities. More specifically, a repetition is defined as one complete movement cycle of a gym activity, such as one push-up or one jumping jack.

Given a video of a person performing a gym activity for multiple repetitions (e.g., doing 10 jumping jacks), we want to segment the time duration for each individual repetition. In order to do so, we first use Mask R-CNN [11] to extract the bounding box of the person for each frame. As the person performs an activity for multiple repetitions, the shape of the bounding box as a function of time captures this periodicity and can be used to calculate the time duration for each repetition. For instance, in a video of a person doing jumping jacks, the height and width of the bounding box varies periodically due to the arm and leg movements. We then segment the individual jumping jacks based on the autocorrelation function of the aspect ratio of the bounding box, which is changing with time.

4.2.2 3D Human Mesh Extraction and Alignment. For each video frame of the activity, we use the HMR algorithm of [14] to extract the 3D mesh of the person, which consists of a large number of 3D points describing the outer surface of the person in the image. We then use the alignment framework proposed in Sec. 3.2 to align the extracted 3D mesh of the person to the 3D Global Coordinate System (GCS) of the setup described in Sec. 4.1.

[3]Here are the links to a few sample training videos for interested readers: https://youtu.be/96zJo3nlmHI (broad jump), https://youtu.be/UpyDdQjBTa0 (forward lunge), https://youtu.be/33UV3Jl8wEk (jumping jack), and https://youtu.be/FvJS_MSN4Lo (lateral lunge).
[4]While it is ideal to use videos without any occlusion, the vision algorithms are still able to reconstruct the 3D human mesh if the occluded part is small, e.g., [14, 40].

While we can, in principle, align any given frame using the method of Sec. 3.2, some frames are easier to align as they would require the extraction of the least amount of information from the image, pertaining to the positions of different body parts. We next show our approach to find such a frame for each activity video, in the context of the stiff-leg deadlift exercise.

Consider the stiff-leg deadlift exercise, in which a person slowly lowers their upper body from the standing position to a bend-forward position, and then quickly rises back to the standing position, while keeping their legs straight (see Fig. 4). Given any random frame, the angle between the person's torso and legs needs to be estimated from the frame, in order to determine the correct corresponding orientation in the GCS, and consequently calculate the rotation matrix. While this knowledge can be estimated from each frame, if we use the frame where the person is fully standing, we do not need to extract any additional knowledge in order to build the rotation matrix, which can then be applied to all the frames of the video. In order to find a frame where the person is in a fully standing position during the stiff-leg deadlift exercise, we utilize the Mask-RCNN algorithm [11] to estimate a bounding box around the person in each frame, which can then automatically identify the frame of the person fully standing as the one with the tallest bounding box. According to our sensing setup of Fig. 3, the LCS axes $\mathrm{x}', \mathrm{y}', \mathrm{z}'$ of the person in the frame with fully standing position should align with the major $\mathrm{x}, \mathrm{y}, \mathrm{z}$ of the GCS. We can then easily calculate the rotation matrix R from the LCS of this frame as $\mathrm{R} = [\mathrm{x}', \mathrm{y}', \mathrm{z}']$, and use this matrix to align the meshes of all the frames of this video into the correct orientation in the GCS. Since the feet are static in this activity, the mesh should be translated in the GCS such that the average of the mesh points of the feet is at the origin.

As an additional example, consider the sit-up exercise. The fully recumbent position is the easiest to use for alignment. More specifically, given the sensing setup of Sec. 4.1, when the person is in the full recumbent position, x', y', and z' of the LCS should align with $+\mathrm{z}$, $+\mathrm{y}$, and $-\mathrm{x}$ axes of the GCS, respectively. Hence, the rotation matrix R can be estimated from the LCS of the video frame in which the person is fully lying down as: $\mathrm{R} = [-\mathrm{z}', \mathrm{y}', \mathrm{x}']$.

As such, we then use the common-sense knowledge of each activity to provide a label for a key frame that our automated algorithm needs to look for, in order to efficiently align the meshes of all the video frames of that activity to the GCS.

4.2.3 WiFi Signal Simulation. After the mesh alignment, the 3D mesh of the person doing the activity is placed and oriented in the GCS of Fig. 3 in a simulation environment. We further place the WiFi transceivers in the locations described in Sec. 4.1 in our simulation environment, where the Tx-Rx separation distances for Links 1 and 2 are 4 m each, and the Tx-Rx distance for Link 3 is 0.6 m. The antennas of Links 1 and 2 are placed at 0.75 m above the floor, and the antennas for Link 3 are placed at 2.75 m above the floor (which is a typical height of room ceilings). As the human mesh moves over time, we simulate the received WiFi signals for all the links as a function of time, by utilizing the Born electromagnetic approximation (Eq. 2). More specifically, given the locations of the M 3D mesh points in the GCS at any time instant (any video frame), we simulate the received WiFi signal as follows,

$$r_{\mathrm{sim}}(\mathrm{p}_R; t) = \left| \mathrm{g}(\mathrm{p}_R, \mathrm{p}_T) + \sum_{m=1}^{M} A_m G_m \, \mathrm{g}\Big(\mathrm{p}_R, \mathrm{p}_m(t)\Big) \, \mathrm{g}\Big(\mathrm{p}_m(t), \mathrm{p}_T\Big) \right|^2, \tag{5}$$

where p_R and p_T are the locations of the Rx and Tx, respectively, $\mathrm{p}_m(t)$ is the location of the m-th mesh point at time t, A_m is the reflection coefficient of the m-th mesh point, and G_m is a scaling parameter that captures the quasi-specular reflection nature of the human body and depends on the normal direction to the body at the m-th mesh point. Since the clothing has a negligible impact on the reflection coefficient, we model human body as a homogeneous reflector (constant A_m). Furthermore, since we are only interested in the motion-related part of the received signal (see Eq. 4), we do not need to calculate the exact value of the received signal and a scaled version will suffice to model human motion. As such, we use a uniform constant reflection coefficient of 1 over the whole body.

Remark 2. *The reflected signals off of the static objects in a real environment, e.g. walls and furniture, contribute to the DC term of Eq. 4, which carries no information about the human motion, and can be easily removed in the operation phase, as we shall see in Sec. 4.3.1. Hence, there is no need to consider the static multipath in the simulation environment and we only need to consider the mesh points of the moving human body in the received signal calculation in Eq. 5.*

When using the HMR algorithm of [14], particularly for video frames where one of the person's arms is occluded from camera view, we have observed some artifacts in the estimation of the arm poses, which may result in abnormal arm movements that the person in the video did not perform. We also noticed that due to its small surface area, as compared to the other body parts, the arms have little contribution to the received WiFi signal [6]. For these reasons, we do not consider the mesh points of the arms in the simulation platform, and only model the interaction between the electromagnetic waves and the rest of the human body.

Remark 3. *At higher frequencies (e.g., 60 GHz) or for some other applications, the impact of arms on the received signal could be higher. In such a case, one can then include the mesh points of the arms in the simulation to capture the impact of the arms on the received signal. If one needs to include the arms in the modeling, other more recent vision algorithms that better estimate the arms can be used [40]. Alternatively, instead of YouTube videos, one can use existing online motion capture data which directly provide high-quality 3D human models.*

4.2.4 Feature Extraction and Classification. For each gym activity, the body parts of the person would produce a different velocity profile, in terms of the speed and the motion direction. Such a velocity profile, i.e., the information about the speed and motion direction of different body parts, can be used as a signature for each gym activity. For instance, when doing push-ups, the up-down motion of the person's body parts (e.g., head, torso) results in moderate speeds in the $\pm z$ direction of the GCS. Meanwhile, push-up produces nearly no speeds in the x and y directions. As another example, consider the broad jump, in which the person jumps forward towards Link 1. The motion of the body parts produces very high speeds in the $+x$ direction, moderate speeds in the $\pm z$ direction, and negligible speeds in the y direction.

As discussed in Sec. 4.1, the instantaneous frequencies of the baseband received signal carry the information on the velocity components of different body parts. Thus, we utilize the frequency content of the received signals at the 3 links to construct informative features that can describe the person's motion characteristics during the gym activity. Towards this goal, we carry out time-frequency analysis of the simulated wireless signals, in order to extract features and train a classifier, as we discuss in the following part.

Time-Frequency Analysis: We estimate the frequency content of the signal via performing time-frequency analysis, a common method in harmonic analysis, which has also been used in RF sensing [17, 30, 31]. More specifically, given a received signal in the time domain, we utilize Short-Time Fourier Transform (STFT) with windows of size 0.4 sec and a window overlap of 0.35 sec to generate a *spectrogram*, which contains the frequency content of the signal (in the range of $[1, 100]$ Hz) as a function of time. In order to enhance the quality of the spectrogram, we carry out a denoising process, as follows. We first zero out all the spectrogram values that are below a noise floor of 0.01. Then, we binarize the spectrogram with a threshold of 0.01 and extract all the connected components from the 2D binary plot. Regions that correspond to very small components are zeroed out, since they are less likely to be the result of continuous human movements.

Fig. 5 (b) shows a sample spectrogram of the simulated signal at Link 3, for one repetition of stiff-leg deadlift extracted from the video of Fig. 5 (a). For instance, when the person moves into a bend-forward position as well as when she moves back to the initial position, the vertical speed of her body is clearly captured by the spectrogram of Link 3, in the first half (from 0 to 1 sec) and the second half (from 1 to 2 sec) of the repetition, respectively. In particular, in the video, the motion of going back to the standing position is faster than that of bending down. This is captured by the spectrogram where the frequency components are higher in the second half of the repetition,

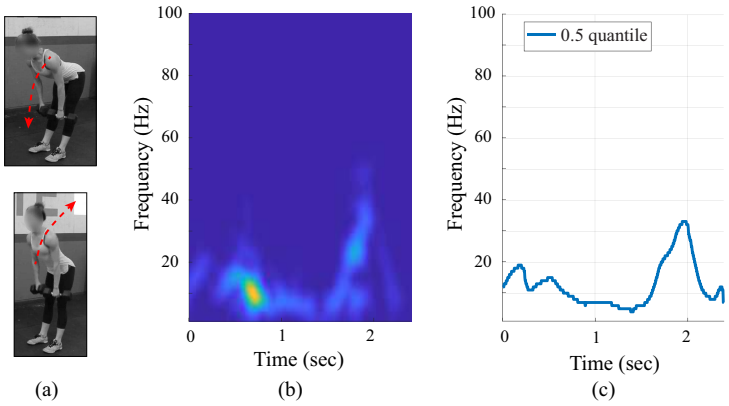

Fig. 5. (a) Snapshots from a video of a person performing a repetition of the stiff-leg deadlift exercise, (b) the corresponding spectrogram of the simulated WiFi signal of Link 3 capturing the motion pattern of the person, and (c) the 0.5 quantile curve of the spectrogram in (b). See the color pdf to best view this figure.

as compared to the first half. Note that the spectrogram (non-DC part) captures the frequency components of the motion and does not depend on the actual received signal strength as long as it is above the noise floor.

Features: Given the simulated signals of one repetition from a gym activity video, we first generate the corresponding spectrograms for the 3 links. We then extract several informative features from the spectrogram of a repetition in order to train a classifier.[5] Our proposed features are mainly based on the quantiles of the spectrograms as a function of time. More specifically, the q quantile of a spectrogram $S(f,t)$ is given as follows, as a function of time:

$$Q(t;q) = \min\left\{ f_q : \frac{\sum_{i=1}^{f_q} S(i,t)}{\sum_{i=1}^{f_{\max}} S(i,t)} \geq q \right\}, \tag{6}$$

where $Q(t;q)$ is the q quantile of the spectrogram as a function of time and $f_{\max}$ is the upper bound of the frequency in the spectrogram (i.e., 100 Hz in our case).

The quantiles as a function of time capture the temporal variations of the spectrogram which capture the time-varying speeds of the body parts, while staying robust to noise. In this study, we use the 0.5 and 0.7 quantiles for each spectrogram, which capture the median speed and the higher-speed components of the motion over time, respectively. Fig. 5 (c) shows the 0.5 quantile for the sample spectrogram of Fig. 5 (b).

Consider one repetition of an activity. We then calculate the histograms of the 0.5 quantile and the 0.7 quantile, respectively, within the repetition, for each spectrogram of each link. Each histogram is a 5-dimensional vector that contains the respective numbers of points with quantile value in the following intervals: $[1, 10]$ Hz, $(10, 20]$ Hz, $(20, 30]$ Hz, $(30, 40]$ Hz, and $(40, 100]$ Hz. These histograms efficiently capture the distributions of the quantile values in each spectrogram of the repetition. In order to capture any possible temporal asymmetry within one repetition of an activity, we calculate the difference between the maximum value of the quantile values in the first and the second halves of the repetition, and use a binary number to indicate whether the difference is larger than 10 Hz, for each quantile curves of Link 3. Finally, the time duration of the repetition is used as the last feature. This amounts to a total of 33 features for each repetition of an activity. Overall, these features capture various time and frequency attributes of the motion, which are useful for classifying the activities.

[5]For a video that contains multiple repetitions of the same gym activity, we first temporally segment the video to extract each repetition (see Sec. 4.2.1) and treat each repetition as an individual training data point.

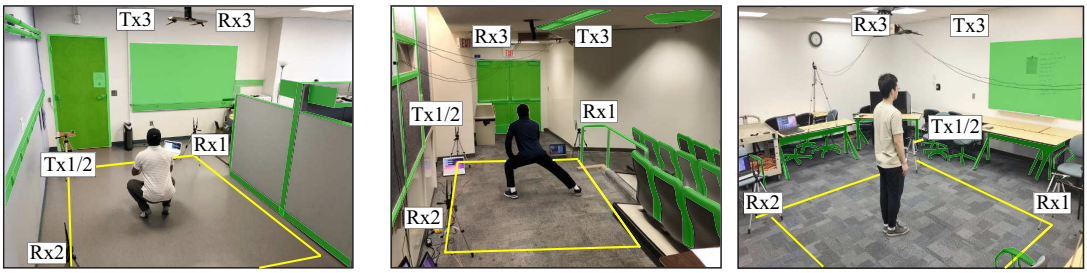

Fig. 6. Our experimental setup in 3 different areas to test our gym activity classification system. Link 1 consists of Tx1/2 and Rx1, Link 2 consists of Tx1/2 and Rx2, and Link 3 consists of Tx3 and Rx3. Area 1 is in a lab, Area 2 is in the back of a classroom, and Area 3 is in a conference room. As can be seen, the areas are cluttered with a variety of objects. For instance, all metallic objects, which can be strong reflectors, are highlighted in green. See the color pdf to best view this figure.

Since in the operation phase, the person may not exactly stand at the center (see Fig. 3), we further perturb the extracted meshes as follows, in order to augment the simulated dataset. More specifically, to generate a perturbed dataset, we draw two numbers uniformly distributed in $[-0.1, 0.1]$ m, which we then use to shift the x and y positions of the 3D human mesh, respectively, in the GCS. We perturb each mesh 10 times in this manner. Overall, we have a total of 1878 simulated feature vectors to be used for training the classifier. As the number of repetitions differs for each activity class, we apply oversampling [2] to balance the training data.

Training a Classifier: We then train a linear classifier using the feature vectors of our simulated gym activity RF dataset. In the operation phase, the trained classifier then takes as input the features of one individual repetition from a measured RF signal and outputs a predicted probability distribution over the 10 activity classes via the softmax operation. The class that corresponds to the highest predicted probability is taken as the classification decision for the input data sample. Given an activity period that possibly contains more than one repetition of the same activity, we can also fuse the predictions of the individual repetitions, in order to achieve a more accurate overall classification. More specifically, we can average the predicted probability distributions of all the repetitions within the same activity period. The activity class corresponding to the highest predicted probability in the aggregated distribution then serves as the classification decision for the activity period.

4.3 Test Experiments with Real WiFi

We have conducted a number of test experiments to collect real WiFi measurements and evaluate the performance of our gym activity classification system that is trained only with online video data. In this section, we then discuss our WiFi experimental setup, and the test data collection and processing.

4.3.1 Experimental Setup and Data Processing. The test experiments are conducted in a total of 3 test areas, which are shown in Fig. 6. These test areas represent several real-world environments with a variety of area sizes, geometry, and clutter. More specifically, Area 1 is located in a lab, Area 2 is located in the back of a classroom, while Area 3 is located in a conference room. Moreover, each test area contains several metallic objects of various sizes (marked with green), e.g., white board, desks, and chairs, which makes the test areas similar to the metal-rich environment in a gym. Clutter, however, does not impact the performance of our approach as the impact of static objects appear in the DC term and is removed, as discussed earlier.

Each Tx/Rx in the test area consists of $N_T = 2$ and $N_R = 2$ antennas connected to a laptop with Intel 5300 Network Interface Card. The shared transmitter for Link 1 and 2 transmits 400 WiFi packets per second on WiFi channel 36 ($f_c = 5.18$ GHz), while that of Link 3 transmits WiFi packets with the same rate on WiFi channel 44 ($f_c = 5.22$ GHz). We use CSItool [10] on each of the receivers to log the squared magnitude of the Channel

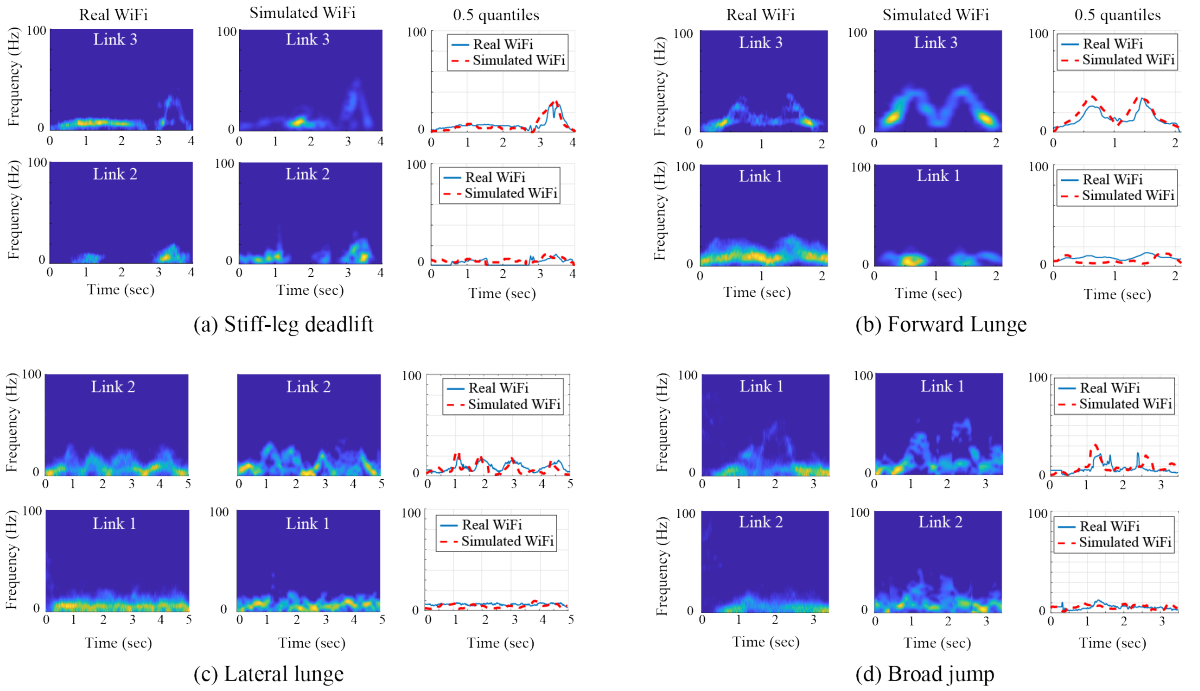

Fig. 7. Comparison between the real and simulated WiFi spectrograms on two links for four exercises: (a) Stiff-leg deadlift, (b) Forward Lunge, (c) Lateral lunge, and (d) Broad jump. For each exercise, left figure is the real WiFi spectrogram, middle figure is simulated WiFi spectrogram, and right figure is the extracted 0.5-quantiles from the two spectrograms. It can be seen that the real and simulated spectrograms are visually similar, and the 0.5-quantiles confirm their similarity. Note that for each exercise, we show the two links that have the best (top) and the worst (bottom) matches between the real and simulated WiFi spectrograms.

State Information (CSI) of the received packets. Each receiver logs a total of $N_T \times N_R \times 30$ subcarriers = 120 data streams, which we first denoise using Principal Component Analysis as described in [31]. More specifically, we use the first 10 principal components of the data streams and generate the received signal's spectrogram using the method of [17], where both STFT and Hermite functions are used to generate high-quality spectrograms. Note that the static multipath from the static objects in the environment appears at DC in the spectrogram, and can easily be removed by subtracting the mean of the signal before generating the spectrogram. We then denoise the spectrograms using the same denoising scheme as described in Sec. 4.2.4. However, here we adaptively estimate the noise floor of the spectrogram as the 99th percentile of the spectrogram values above 70 Hz. We assume that the frequency range above 70 Hz has no informative reflections from the human body, since it corresponds to speeds above 2 m/s.

4.3.2 WiFi Test Data Collection. For the WiFi test experiments, we have recruited a total of 10 subjects to participate in our test experiments, including 8 males and 2 females. In each area, each subject is asked to perform the 10 gym activities. For each activity, we collect the WiFi measurements of each subject for 45 seconds, to which we refer as an activity period. During each activity period, the subject performs multiple repetitions of this activity. We then temporally segment the WiFi measurement of each activity period to extract the time intervals of the individual repetitions, based on the brief resting periods between two consecutive repetitions.

Overall, we have a total of 1543 repetitions for the 10 gym activities, or equivalently, a total of 300 activity periods (100 activity periods per area), from the 10 subjects in the 3 areas. The activity periods each contain an average of 5.1 repetitions. More specifically, we have 523 individual repetitions in area 1, 517 in area 2, and 503 in area 3.

4.4 Performance Evaluation

In this section, we evaluate the performance of our proposed approach for training the WiFi gym activity classifier using only video data and no RF data. We first analyze the similarity between the simulated and real WiFi signals, in order to validate our generated training data. Then, we extensively evaluate the classification performance of our trained WiFi sensing system with real WiFi test data.

4.4.1 Similarity between Simulated and Real Data. In order to present a proper assessment of the similarity between the simulated WiFi signal and the real one, we collect the WiFi measurements of a person performing two activities (stiff-leg deadlift and forward lunge), while recording a video of the scene at the same time. We then analyze the similarity of the simulated WiFi signal and the real one via spectrogram analysis, since a spectrogram can effectively capture the motion of different gym activities, as discussed in Sec. 4.2.4. Note that the video recording and WiFi measurements of this experiment are collected solely for the purpose of showing the similarity between the simulated data and the real one, and neither of them was used in the training set or in the test set of our system.

Fig. 7 shows the comparison between the real and simulated WiFi spectrograms for four sample exercises. For each activity, we show the two links that have the best (top) and the worst (bottom) matches between the real and simulated WiFi spectrograms. Fig. 7 (a) shows the spectrograms of the measured WiFi data on Link 2 and Link 3 for one repetition of the stiff-leg deadlift exercise (left column), as well as the video-based simulated WiFi data on these 2 links for the same exercise (middle column). It can be seen that the spectrograms based on simulating from the video properly capture the motion patterns, and match the WiFi spectrograms well. The figure also shows the 0.5 quantile curves (one of the features we shall use) of both the real WiFi spectrogram and the simulated one (right column), showing a good match between the two. Similarly, a good match can be seen between the real WiFi spectrograms and the simulated ones of the forward lunge, lateral lunge, and broad jump activities. The average cosine similarity between the 0.5 quantile curves of the simulated and real spectrograms of all four activities is 0.88, while two identical curves have a cosine similarity of 1. Finally, the spectrograms of these exercises reveal the unique patterns of each exercise in terms of the frequency content of the spectrograms, or, equivalently, the speed profile of the person's body.

Overall, Fig. 7 shows the power of our proposed approach in generating simulated RF data that closely resemble the real one, and highlights its potential in eliminating the need for the collection of real RF measurements when training RF sensing systems.

4.4.2 Classification Performance. In this part, we evaluate the performance of our WiFi-based gym activity classifier, which is trained with only video data. We first present the results for the case where the classifier is tested with individual activity repetitions. In other words, multiple repetitions of the same activity by the same person are treated as independent test cases in this setting. We then evaluate the classifier for the case where it jointly uses all the repetitions done by the same person during an activity period to classify his/her activity (by fusing their corresponding decisions). As expected, the second case would perform better since the data of a few repetitions is used for classification. The first case, however, is important as it establishes a lower bound on the performance for the case when the person only does one repetition of an exercise. As we shall see, we can still classify the activities well, even with only one repetition.

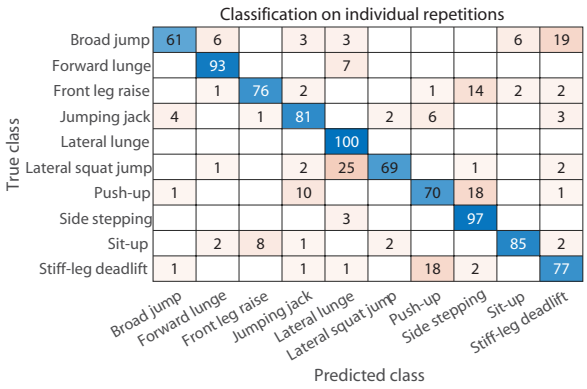

Fig. 8. Confusion matrix of classifying the 10 gym activities with WiFi, based on individual repetitions of the activities.

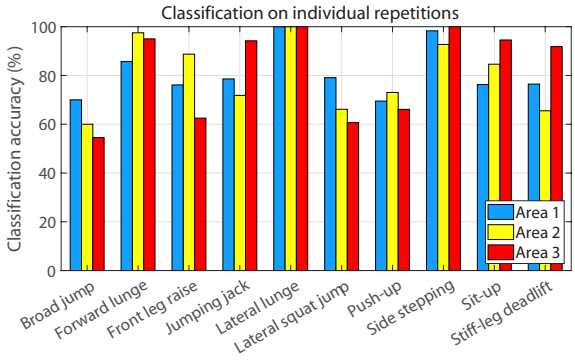

Fig. 9. Classification accuracy for each activity class in each test area, when using individual repetitions.

Classification on Individual Repetitions: In this setting, our classifier achieves an average classification accuracy of 81%, over all the test areas (a random selection would have resulted in 10% accuracy). Fig. 8 shows the confusion matrix that corresponds to the classification performance of our trained system on individual repetitions. The diagonal entries indicate the classification accuracy for each activity (in %) and the off-diagonal entries indicate the percentages of misclassifications to that corresponding class. Overall, it can be seen that our system can classify all the activities pretty well. In particular, activities such as forward lunge, lateral lunge, and side stepping are recognized very well, with classification accuracies above 90%. On the other hand, some other activities, such as lateral squat jump, are classified with a lower accuracy due to the inherent similarity with another activity. For instance, lateral squat jump is very similar to lateral lunge, as they both require lateral and vertical motion of the body.

Next, we show the classification accuracy as a function of the areas in Fig. 9. It can be seen that the respective accuracies in the three areas are very similar to each other. This indicates that our video-trained WiFi sensing system is not sensitive to the multipath effects from static objects in real WiFi environments, due to the fact that we utilize motion-driven features that capture only the person's speed profile, as discussed in Sec. 4.2.4.

Classification on Activity Periods: We next show the performance when a small number of repetitions of an activity period are used to classify that activity. During the test, each person performs each activity for 45 seconds in each area. Thus, depending on the speed of the person, there may be more than one repetition in an activity period (the average is 5.1 repetitions per activity period). For such a case, our classifier fuses the predictions of the individual repetitions in order to improve the classification quality, as discussed in Sec. 4.2.4. Fig. 10 shows the confusion matrix for this case. The overall average classification accuracy improves to 86%, as compared to the accuracy of 81% on individual repetitions. In particular, jumping jack and side stepping are now classified correctly 100% of the time, as compared to 81% and 97% in the repetition-based setting. This case also has a similar performance across the 3 areas.

Classification Error Analysis: In this part, we perform an in-depth analysis on the classification errors. Fig. 11 shows the spectrograms for a sample activity pair that is confusing for the classifier (push-up and side stepping), as indicated in the confusion matrices in Figs. 8 and 10. More specifically, push-ups are classified as side stepping 23% of the time according to Fig. 10. It can be seen that the frequency distribution in each link is very similar for the two activities, due to the inherent motion similarity between the two, which is the main source of classification error. For instance, since neither of them involve highly-dynamic motion, the corresponding frequencies captured in all the links are low for both activities, with similar spectrogram patterns.

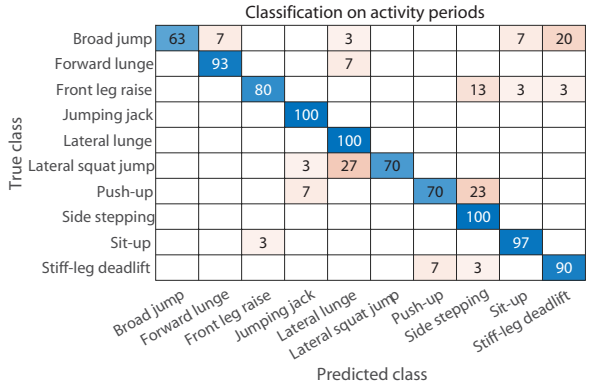

Fig. 10. Confusion matrix of classifying the 10 gym activities with WiFi, based on activity periods containing an average of 5.1 repetitions each.

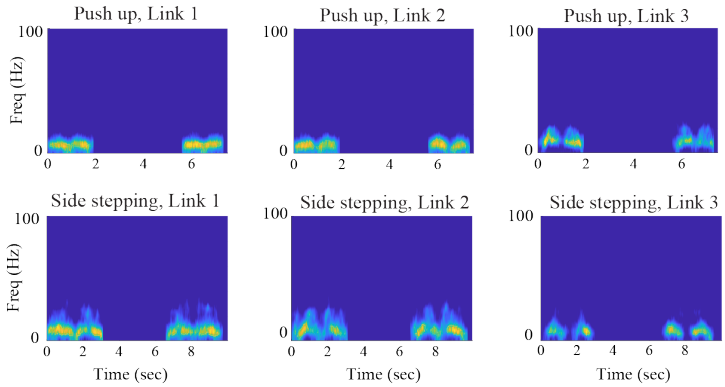

Fig. 11. WiFi spectrograms of two activities: push up and side stepping, which are the very confusing for the classifier (push-ups are classified as side stepping 23% of the time as can be seen in Fig. 10). Two repetitions of the exercise are shown in each spectrogram. It can be seen that the frequency distribution is similar for both activities, which makes it more challenging for the classifier to differentiate them.

4.4.3 Robustness to Metallic Objects in the Environment. Our test areas of Fig. 6 contain various metallic objects (comparable to human heights) around the person to resemble the gym environment. In our proposed system, we utilize non-DC parts of the spectrograms to capture the motion information, which are insensitive to the static scatterers in the environment which appear at DC, even if they are highly reflective. In order to further test the robustness of our pipeline to highly reflective clutter, we have created an even more metal-rich environment by including additional large highly-reflective objects (e.g., multiple metallic drawer cabinets and highly-reflective shielding material sheets) in one of our test areas, as shown in Fig. 12. One subject then performs different exercises in both the original setting of this area (top row) and the new metal-heavy setting (bottom row). Fig. 12 also shows the spectrograms of the received WiFi signals in both settings, for four sample exercises. It can be seen that the spectrograms of the received signals are almost identical, with or without the large metallic objects in the area. This result validates our signal model of Eq. 4 and indicates that our proposed pipeline is not affected by the static objects in the environment, even if they are highly reflective.

Overall, our results show that we have, for the first time, successfully trained an RF sensing system, without collecting any prior RF training measurements. Moreover, in terms of RF-based gym activity classification, our

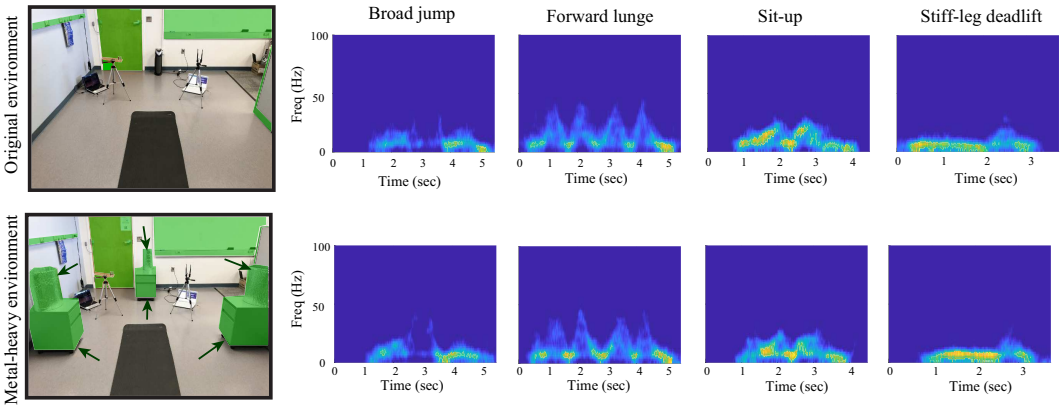

Fig. 12. Spectrograms of the collected WiFi data on Link 1 for a person performing 4 sample exercises in both the original setting (top row) and a metal-heavy setting (bottom row) of Area 1. Metallic objects are highlighted in green. The arrows point to the added metallic objects in the metal-heavy setting of the bottom figure.

proposed approach has not only enabled it with no prior RF training measurements, but has also enabled the first reflection-based system, whereas all the existing methods rely on the person to cross/block the line-of-sight path (i.e., the direct path between the Tx and Rx) while performing the exercises.

5 DISCUSSION AND FUTURE WORK

In this section, we discuss a few more aspects related to our proposed framework and the case study.

Generalization of the Video-Based Training Approach to Other Applications: In this paper, we proposed a general approach that can train RF sensing systems without any RF training data, and by using the vast available online videos. While we showcased the performance of this approach in the context of gym activity recognition, showing how we can achieve a high accuracy with zero RF training data, the proposed methodology is applicable to many other RF sensing applications, scenarios, and setups. As part of the future work, we envision that this approach can be used to train other RF sensing systems to recognize other activities, gestures, and in general other situations that involve motion of body parts. It can further be used for analysis purposes, for instance to understand the optimal RF setup/amount of needed resources for a particular application, to understand the differentiability of different activities, or to understand the limitations of sensing with a certain setup or at a particular frequency, all without the need to collect any RF data. Overall, the proposed approach is scalable and general, and can thus enable new work in the area of RF sensing.

Further Discussions on the Sensing Setup: In the considered sensing setup for our gym activity classification study (Fig. 3), we assumed that the person is at the center of the coordinate system and facing the positive x direction. In order to set the coordinate system of the sensing setup in that manner, the location and orientation of the person are assumed to be known. This is a realistic assumption since there is a great body of work on localization and tracking with RF signals, e.g., [19, 25, 29], that can be utilized to first estimate the location and orientation of the person.

Furthermore, we assumed that the 3 WiFi links are placed such that ℓ_1 is parallel to the x-axis, ℓ_2 is parallel to the y-axis, and ℓ_3 is parallel to the z-axis (see Fig. 3). As such, in this configuration, each link directly captures the information about one of the three-dimensional components of the velocity vectors of different body parts (as discussed in Sec. 4.1). For any general transceiver locations that are not similar to Fig. 3, the motion information across the three dimensions can become coupled in the measurements. Then, a simple linear system equation can

be solved to directly extract the motion components across the three dimensions, similar to what is done in [41]. Once these are extracted, our trained pipeline based on the configuration of Fig. 3 can be used for classification. In summary, the configuration of Fig. 3 can be used as a base since it directly measures the motion across the three dimensions while other configurations can be translated to it.

Sensing Multiple People: In the gym activity classification study, we assumed that there is only one person doing the exercise, with little movements from other people nearby (other people were present but not moving much). In the case where there are other people simultaneously moving nearby (e.g., performing exercises), the received signal will contain the motion information of the person of interest as well as those nearby. In such scenarios, one can then use multiple antennas at each transceiver to create a small antenna array and separate the impact of multiple people on the received signal by beamforming towards each person [16]. The reflected signals off of different people can also be separated in other domains, e.g., Time-of-flight, Angle-of-Departure [37].

6 CONCLUSIONS

In this paper, we proposed a new and generalizable framework that allows for successfully training RF sensing systems only with already-available video data, and without any real RF data, thus eliminating the traditional labor-intensive phase of collecting real RF training measurements. More specifically, our proposed approach taps into the vast number of available online videos of different human activities/motions, translates them into instant simulated RF data, extracts relevant time-frequency features, and trains a neural network pipeline. Our approach is general and scalable to any motion-based human activity and any given setup. In order to validate our proposed framework, we carried out a case study of gym activity classification using WiFi transceivers. We utilize YouTube videos of the corresponding gym activities, construct a simulated RF dataset, extract key features via time-frequency analysis, and train a classifier, thus using no real RF training measurement. After training, the classifier was then extensively tested with real WiFi measurements of 10 subjects performing the 10 gym activities in 3 different test areas. Overall, our system achieved a classification accuracy of 86% when tested on a small activity period that contains an average of 5.1 repetitions, and 81% when tested on individual repetitions of activities. This demonstrates that the proposed approach can successfully train an RF sensing system with already-available video data, and without any real RF measurements.

ACKNOWLEDGMENTS

The authors would like to thank all the participants in our experiments. The authors would also like to thank the editors and reviewers for their valuable comments and helpful suggestions. This work was funded in part by ONR award N00014-20-1-2779 and in part by NSF NeTS award 1816931.

REFERENCES

[1] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar. 2018. Rethinking the Faster R-CNN architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[2] N. V. Chawla. 2009. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*. 875–886.

[3] Q. Chen, B. Tan, K. Chetty, and K. Woodbridge. 2016. Activity recognition based on micro-Doppler signature with in-home Wi-Fi. In *Proceedings of the IEEE International Conference on e-Health Networking, Applications and Services*.

[4] W. C. Chew. 1995. *Waves and fields in inhomogeneous media*. IEEE Press.

[5] S. Depatla, L. Buckland, and Y. Mostofi. 2015. X-ray vision with only WiFi power measurements using rytov wave models. *IEEE Transactions on Vehicular Technology* 64, 4 (2015), 1376–1387.

[6] J. L. Geisheimer, E. F. Greneker III, and W. S. Marshall. 2002. High-resolution Doppler model of the human gait. In *Radar Sensor Technology and Data Visualization*.

[7] W. C. Gibson. 2014. *The method of moments in electromagnetics*. Chapman and Hall/CRC.

[8] L. Guo, L. Wang, C. Lin, J. Liu, B. Lu, J. Fang, Z. Liu, Z. Shan, J. Yang, and S. Guo. 2019. Wiar: A public dataset for WiFi-based activity recognition. *IEEE Access* 7 (2019), 154935–154945.

[9] X. Guo, J. Liu, C. Shi, H. Liu, Y. Chen, and M. C. Chuah. 2018. Device-free personalized fitness assistant using WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 165.

[10] D. Halperin, W. Hu, A. Sheth, and D. Wetherall. 2011. Tool release: Gathering 802.11n traces with channel state information. *ACM SIGCOMM Computer Communication Review* 41, 1 (2011), 53–53.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*.

[12] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas, and W. Xu. 2018. Towards environment independent device free human activity recognition. In *Proceedings of the ACM International Conference on Mobile Computing and Networking*.

[13] J.-M. Jin. 2015. *The finite element method in electromagnetics*. John Wiley & Sons.

[14] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[15] C. R. Karanam, B. Korany, and Y. Mostofi. 2018. Magnitude-based angle-of-arrival estimation, localization, and target tracking. In *Proceedings of the ACM/IEEE International Conference on Information Processing in Sensor Networks*.

[16] C. R. Karanam, B. Korany, and Y. Mostofi. 2019. Tracking from one side: Multi-person passive tracking with WiFi magnitude measurements. In *Proceedings of the International Conference on Information Processing in Sensor Networks*.

[17] B. Korany, C. R. Karanam, H. Cai, and Y. Mostofi. 2019. XModal-ID: Using WiFi for through-wall person identification from candidate video footage. In *Proceedings of the ACM International Conference on Mobile Computing and Networking*.

[18] T. Li, L. Fan, M. Zhao, Y. Liu, and D. Katabi. 2019. Making the invisible visible: Action recognition through walls and occlusions. In *Proceedings of the IEEE International Conference on Computer Vision*.

[19] X. Li, D. Zhang, Q. Lv, J. Xiong, S. Li, Y. Zhang, and H. Mei. 2017. IndoTrack: Device-free indoor human tracking with commodity Wi-Fi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 72.

[20] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. 2015. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics* 34, 6 (2015), 248.

[21] X. Ma, R. Zhao, X. Liu, H. Kuang, and M. A. A. Al-qaness. 2019. Classification of human motions using micro-Doppler radar in the environments with micro-motion interference. *Sensors* 19, 11 (2019), 2598.

[22] Y. Ma, G. Zhou, and S. Wang. 2019. WiFi sensing with channel state information: A survey. *Comput. Surveys* 52, 3 (2019), 46.

[23] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele. 2018. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *Proceedings of the International Conference on 3D Vision*.

[24] Q. Pu, S. Gupta, S. Gollakota, and S. Patel. 2013. Whole-home gesture recognition using wireless signals. In *Proceedings of the ACM International Conference on Mobile Computing and Networking*.

[25] K. Qian, C. Wu, Y. Zhang, G. Zhang, Z. Yang, and Y. Liu. 2018. Widar2.0: Passive human tracking with a single Wi-Fi link. In *Proceedings of the 1International Conference on Mobile Systems, Applications, and Services*.

[26] Y. Sun, Y. Ye, W. Liu, W. Gao, Y. Fu, and T. Mei. 2019. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE International Conference on Computer Vision*.

[27] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. 2018. Bodynet: Volumetric inference of 3D human body shapes. In *Proceedings of the European Conference on Computer Vision*.

[28] F. Wang, W. Gong, J. Liu, and K. Wu. 2018. Channel selective activity recognition with WiFi: A deep learning approach exploring wideband information. *IEEE Transactions on Network Science and Engineering* (2018).

[29] J. Wang, H. Jiang, J. Xiong, K. Jamieson, X. Chen, D. Fang, and B. Xie. 2016. LiFS: Low human-effort, device-free localization with fine-grained subcarrier information. In *Proceedings of the ACM International Conference on Mobile Computing and Networking*.

[30] W. Wang, A. X. Liu, and M. Shahzad. 2016. Gait recognition using WiFi signals. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*.

[31] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu. 2015. Understanding and modeling of WiFi signal based human activity recognition. In *Proceedings of the ACM International Conference on Mobile Computing and Networking*.

[32] Y. Wang, K. Wu, and L. M. Ni. 2016. WiFall: Device-free fall detection by wireless networks. *IEEE Transactions on Mobile Computing* 16, 2 (2016), 581–594.

[33] Z. Wang, B. Guo, Z. Yu, and X. Zhou. 2018. Wi-Fi CSI-based behavior recognition: From signals and actions to activities. *IEEE Communications Magazine* 56, 5 (2018), 109–115.

[34] D. Wu, D. Zhang, C. Xu, H. Wang, and X. Li. 2017. Device-free WiFi human sensing: From pattern-based to model-based approaches. *IEEE Communications Magazine* 55, 10 (2017), 91–97.

[35] X. Wu, Z. Chu, P. Yang, C. Xiang, X. Zheng, and W. Huang. 2018. TW-See: Human activity recognition through the wall with commodity Wi-Fi devices. *IEEE Transactions on Vehicular Technology* 68, 1 (2018), 306–319.

[36] F. Xiao, J. Chen, X. Xie, L. Gui, J. Sun, and R. Wang. 2018. SEARE: A system for exercise activity recognition and quality evaluation based on green sensing. *IEEE Transactions on Emerging Topics in Computing* (2018).

[37] Y. Xie, J. Xiong, M. Li, and K. Jamieson. 2019. mD-Track: Leveraging multi-dimensionality for passive indoor Wi-Fi tracking. In *Proceedings of the International Conference on Mobile Computing and Networking*.
[38] S. Yeung, O. Russakovsky, G. Mori, and Li Fei-Fei. 2016. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[39] F. Zhang, K. Niu, J. Xiong, B. Jin, T. Gu, Y. Jiang, and D. Zhang. 2019. Towards a diffraction-based sensing approach on human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 33.
[40] T. Zhang, B. Huang, and Y. Wang. 2020. Object-occluded human shape and pose estimation from a single color image. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.
[41] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang. 2019. Zero-effort cross-domain gesture recognition with Wi-Fi. In *Proceedings of the International Conference on Mobile Systems, Applications, and Services*.
[42] Y. Zou, W. Liu, K. Wu, and L. M. Ni. 2017. Wi-Fi radar: Recognizing human behavior with commodity Wi-Fi. *IEEE Communications Magazine* 55, 10 (2017), 105–111.