Wireless Sensing with Machine Learning: Through-Wall Vision & Contactless Health Monitoring

Mingmin Zhao

Wireless Sensing with Machine Learning: Through-Wall Vision & Contactless Health Monitoring

by

Mingmin Zhao

M.S. in Electrical Engineering and Computer Science, Massachusetts Institute of Technology (2017)
B.S. in Computer Science, Peking University (2015)

Submitted to the
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

February 2022

Author . . . .
Department of Electrical Engineering and Computer Science
October 25, 2021

Certified by . . . .
Dina Katabi
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Wireless Sensing with Machine Learning: Through-Wall Vision & Contactless Health Monitoring

by

Mingmin Zhao

Submitted to the Department of Electrical Engineering and Computer Science
on October 25, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

There is significant interest in technologies that can sense people and monitor their health with minimal overhead. Existing solutions typically require people to wear different sensors and devices on their bodies. This thesis demonstrates how we can use wireless signals and machine learning to sense people without any physical contact with their bodies. We develop novel radio sensors that sit in the background like a Wi-Fi router. Our sensors however analyze the surrounding radio signals using novel machine learning algorithms to monitor people's movements and activities, assess their vital signs, learn their sleep and sleep stages, and recognize their emotions. Since wireless sensors traverse walls, our sensors can deliver all of these functions through walls and occlusions.

The key challenge in delivering the above contributions is that radio signals interact with people and the environment in complex ways, resulting in an underdetermined mapping that varies across time and space. To address this problem, this dissertation adopts a data-driven approach and develops custom machine learning models that operate on radio signals. Developing such models requires technical innovations to address unique challenges due to the specularity of radio signals in the frequencies of interest, multipath reflections in indoor environments, high data rates and computation complexity, and the lack of training data and the difficulty in annotating radio signals. Our work addresses these challenges and enables two new capabilities: through-wall tracking of the human pose and contactless health monitoring.

Thesis Supervisor:	Dina Katabi
Title:	Professor of Electrical Engineering and Computer Science

Previously Published Material

Chapter 3 revises a previous publication [1]: Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-Wall Human Pose Estimation Using Radio Signals. CVPR 2018.

Chapter 4 revises a previous publication [2]: Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. RF-based 3D Skeletons. ACM SIGCOMM 2018.

Chapter 5 revises a previous publication [3]: Mingmin Zhao, Yingcheng Liu, Aniruddh Raghu, Tianhong Li, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-Wall Human Mesh Recovery Using Radio Signals. ICCV 2019.

Chapter 6 revises a previous publication [4]: Mingmin Zhao, Kreshnik Hoti, Hao Wang, Aniruddh Raghu, and Dina Katabi. Assessment of Medication Self-Administration Using Artificial Intelligence. Nature Medicine, 27(4), 2021.

Chapter 7 revises a previous publication [5]: Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi Jaakkola, and Matt Bianchi. Learning Sleep Stages from Radio Signals: A Conditional Adversarial Architecture. ICML 2017.

Chapter 8 revises a previous publication [6]: Mingmin Zhao, Fadel Adib, and Dina Katabi. Emotion Recognition using Wireless Signals. ACM MobiCom 2016.

Acknowledgments

The research presented in this thesis would not have been possible without the help and support from many people to whom I owe gratitude.

The first person I want to thank is my advisor Dina Katabi, who has given me this opportunity to work with her at MIT and has always been there for me during this journey. Dina's passion for research is truly an inspiration to everyone around her. For the past years, she has spent countless hours with me, discussing my research, helping me with my papers and talks, and shaping me as an independent researcher. Dina has encouraged and inspired me to become a better researcher. I could not wish for a better advisor, and I hope that I can do the same things for my students as Dina has done for me.

I am also profoundly grateful to my thesis committee and letter writers: Fadel Adib, Tommi Jaakkola, Antonio Torralba, Ranveer Chandra, and Brian Kim. They have been great sources of inspiration and insights for me and have been incredibly supportive during my Ph.D. journey and the faculty application process. I worked closely with Fadel Adib during my first year at MIT. He has since then become a role model I always look up to and learn from. I am also fortunate to collaborate with Tommi Jaakkola on the sleep staging project, which resulted in my first publication that is outside the traditional research areas of our group. Tommi helped me gain confidence in conducting multi-disciplinary research, which brought about a large body of research in this thesis. I am really grateful for Antonio Torralba, who collaborated with me on through-wall-vision systems. His enthusiasm for cool research ideas is truly inspirational to me and has made me always look for more exciting ways to fuse radio and visual data. Ranveer Chandra mentored me at Microsoft Research during my summer internship. It has been exhilarating to see how Ranveer transforms research ideas into products solving some of the most challenging problems facing our society. Brian Kim collaborated with me on a novel use of wireless sensing technologies to help patients suffering from atopic dermatitis. Working with Brian helped me realize that nothing could be better than seeing my research helping patients.

At MIT, I was lucky to work closely with a large group of smart minds: Chen-Yu Hsu, Shichao Yue, Zach Kabelac, Hao Wang, Hao He, Rumen Hristov, Chengzhi Mao, Guang-He Lee, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Yingcheng Liu, Lijie Fan, Aniruddh Raghu, Kreshnik Hoti, Sarah Gurev, and Michail Ouroutzoglou. We have spent many sleepless nights collaborating on various projects, and they owe much credit for the research in this thesis.

I have also been fortunate to be part the NETMIT research group together with: Lixin Shi, Haitham Hassanieh, Swarun Kumar, Omid Abari, Dinesh Bharadia, Ezz Hamed, Hariharan Rahul, Anubhav Jain, Guo Zhang, Yuzhe Yang, Yingcong Chen, Yuan Yuan, Kevin Zha, Peng Cao, Henry Davidge, and Mary McDavitt. They have provided invaluable feedback on my paper submissions and talks, and made lab a fun place to spend time in.

I am extremely thankful to the people who enabled me to get to MIT in the first place. I am very grateful to Fan Ye and Yizhou Wang, who guided me into research and mentored my undergraduate research thesis. I have also been fortunate to work with Ruipeng Gao, Tao Ye, Jiaxu Zhu, Kaigui Bian, Ming Zhang, Guojie Luo, Tao Wang, and Xiaoming Li at Peking University, as well as Tai-Sing Lee at Carnegie Mellon University. I owe deep gratitude to everyone who believed in me and prepared me for the next steps in my journey.

I'm blessed with a large group of friends who have made my life so much fun: Zhiwei Ding, Jinhu Dou, Jiaqi Liu, Chuang Gan, Lisan Liao, Ximeng Qiu, Bing Gou, Tianyu Zhao, Yi Wang, Yunfei Ma, Hongzi Mao, Yimeng Zhang, Chu Liu, Ruizhi Liao, Tete Xiao, Tao Du, Chao Liang, Xiao Chen, Chenren Xu, Ziao Lin, Sicong Shan, Bicheng Han, Jiaxin Chen, Youjie Li, Sagar Jha, Feng Wei, and Yang Liu.

Last but not least, words cannot describe my gratitude to my family: my parents, parents-in-law, grandparents, sisters, brothers, uncles, and aunts. I couldn't have achieved this without your endless support and love. My wife, Yining, and our kids (dogs), Moko and Moki; with you I feel at home, wherever we are. This thesis is dedicated to you.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

Today, wireless systems analyze radio frequency (RF) signals with carefully-crafted signal processing algorithms. These algorithms are, however, limited by human understanding of how signals propagate and interact with the environment. This thesis explores the customization of machine learning to interpret wireless signals, enabling entirely new applications and services. RF signals have the potential of being a uniquely powerful sensing modality; they propagate in space, traverse walls and obstacles, reflect off people, and get modulated by human movements, respiration, and even heartbeats. If we can interpret such RF reflections, we can sense people through walls and occlusions, and learn much information about their health and wellbeing without any physical contact.

Developing sensing technologies that infer people's movements and physiological signals from the surrounding radio waves is not simple. While the information may be encoded in the radio signal, such coding is complex, unknown a priori, varies across time and space, and depends on the specific characteristics of the objects in the environment, their shapes and material. Further, in most cases, the signal is weak and marred in noise, and the system is underdetermined (i.e., the radio receiver does not have enough resolution to separate the RF signals that bounce off different objects in the environment).

Traditional methods for RF-based sensing, such as RADAR and RF-based localization (see Chapter 2 for details), cannot effectively address these challenges. They mainly rely on signal processing and closed-form equations that do not capture the complexity and uncertainty faced in practical scenarios. For example, much of the work in this space assumes that radio signals reflect off either point objects or rigid surfaces [7, 8, 9, 10, 11, 12, 13, 14, 15]. Yet, the human body is not a single point in space. It is a large deformable object that changes its shape and orientation as the person moves.

In this dissertation, we adopt a data-driven approach, where we integrate the knowledge of RF signal processing and propagation properties into novel machine learning models that analyze radio signals to track people, infer their movements, and monitor their health and vital signs through walls and occlusions. We produce smart radio sensors that enable two types of capabilities: 1) detailed through-wall human sensing, and 2) touchless health monitoring.

■ 1.1 Though-Wall Human Sensing

For many years, humans have dreamed of X-ray vision and explored the concept in comic books and sci-fi movies. Yet, our eyes can only sense the visible light, which does not traverse walls or occlusions. Radio signals, on the other hand, traverse such obstacles. It is natural to wonder whether we can leverage radio waves to see though walls.

The concept of imaging using wireless reflections has roots in RADAR and SONAR technologies. Such technologies transmit a wireless signal and analyze its reflections to extract an image of the reflecting object. Yet, adapting such technologies to seeing people through walls is difficult. Indoor environments are full of reflective objects, including people, furniture, walls, floor, ceilings, etc. Radio waves bounce off these objects many times, and combine both constructively and destructively before they reach the receiver, making it difficult to track the incoming rays back to the reflecting object. Hence, until recently, seeing through-wall technologies have been limited to detecting the presence and approximate location of people behind a wall [7, 8, 9, 10, 11, 12, 13].

This dissertation introduces the first RF systems that can sense a full and dynamic human body through walls. Our systems can infer the 3D skeletons of multiple people. The skeletons are dynamic; their movements follow the movements and actions of the people behind the wall as they sit, stand, or walk. Further, their body shapes and sizes (e.g., tall vs. short, thin vs. heavy) match the bodies of the actual people. Underlying our systems are new specialized neural networks that operate over radio signals, and that we have gradually refined to enable more capabilities and better accuracy from one system to

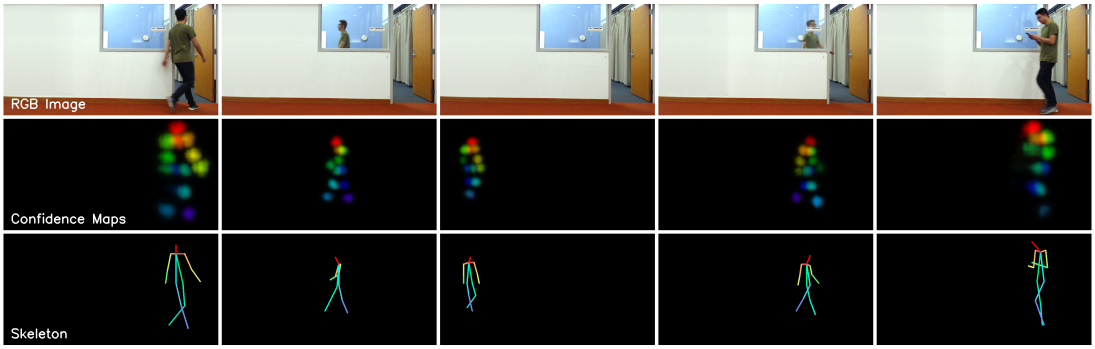


Figure 1-1: **Through-Wall 2D Human Pose Estimation.** RF-Pose tracks 2D human pose as the person enters the room and even when he is fully occluded behind the wall. **Top:** Images captured by a camera colocated with our radio sensor, and presented here for visual reference. **Middle:** Keypoint confidence maps extracted from RF signals *alone*, without any visual input. **Bottom:** Skeleton parsed from keypoint confidence maps showing that we can use RF signals to estimate the human pose even in the presence of full occlusion. Full video is available at: `https://www.youtube.com/watch?v=HgDdaMy8KNE`.

the next, as we describe below.

1.1.1 Through-Wall 2D Human Pose Estimation

Estimating the human pose is an important task with applications in activity recognition, gaming, etc. The problem is defined as generating 2D skeletal representations of the joints on the arms and legs, and keypoints on the torso and head. Prior to our work, research in computer vision has developed neural network models to extract the human pose from images and videos [16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29]. But camera-based systems do not work in the presence of occlusions or poor lighting conditions. Prior work in RF-based sensing has also considered this problem [7, 8, 9, 10, 11, 12, 13]. However, prior RF systems detected people as blobs, and none was capable of estimating the human pose or localizing its keypoints.

We introduced RF-Pose, a neural network system that parses radio signals and extracts accurate 2D human poses, even when people are occluded or behind a wall. We also showed that the pose learned from RF signals extracts identifying features of the people and their motion style. A classifier that uses the extracted poses was able to identify 100 individuals with an average accuracy of 83%, even when they were behind a wall. Figure 1-1 shows an example output of RF-Pose tracking a person as he enters a room, becomes partially visible through a window, and then walks behind the wall.

To develop machine learning models that operate on radio signals, we had to address multiple challenges that evolved with the complexity of the system and the task. The first challenge that we faced was: *how do we label radio signals?* We had to develop a new dataset for training and testing our neural network models. However, unlike images and audio datasets, radio signals cannot be interpreted by humans. Hence, we could not generate labelled data by asking workers to label the person in a radio sample. To address this challenge, we used cross-modal supervision. During training, we attached a web camera to our wireless sensor, and synchronized the wireless and visual streams. We extracted pose information from the visual stream and used it as a supervisory signal for the wireless stream. Once the system was trained, it used only the radio signal as input. The result was a system capable of estimating human pose using wireless signals only, without requiring human annotation as supervision. This cross-modal supervision has become a standard method for generating large-scale RF datasets and training RF-based neural models [30, 31, 32, 33, 34, 35, 36].

The second challenge stems from the *specularity of radio signals* at frequencies that traverse walls. RF specularity is a physical phenomenon that occurs when the wavelength is larger than the roughness of the surface. In this case, the object acts like a reflector - i.e., a mirror - as opposed to a scatterer. The wavelength of our radio is about 5cm and hence humans act as reflectors. Depending on the orientation of the surface of each limb, the signal may be reflected towards our sensor or away from it. Thus, in contrast to cameras where any snapshot shows all unoccluded keypoints, in radio systems, a single snapshot has information about a subset of the limbs and misses limbs and body parts whose orientation at that time deflects the signal away from the sensor. To address this challenge we had to design our model to work across time and space to generate snapshots of 2D poses from radio signals.

Addressing these challenges enabled the first system that can track the 2D human pose as it moves behind walls and occlusions, as shown in Figure 1-1.

■ 1.1.2 Through-Wall 3D Human Pose Estimation

Unlike images and videos, RF signals carry information about depth. As a radio wave travels in space, its phase changes with distance. This information can be retrieved to measure the depth of the object that reflected the radio signal. Thus, a natural next step

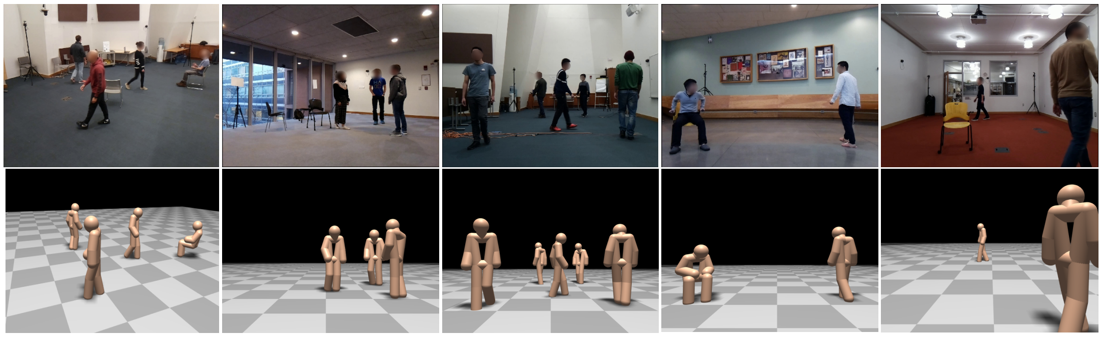


Figure 1-2: **Through-Wall 3D Human Pose Estimation.** RF-Pose3D estimates 3D human skeletons for multiple people and in different environments. It captures depth information in addition to the 2D human poses. **Top:** Images captured by cameras for visual reference. **Bottom:** Predicted 3D human poses with RF signals only. Full video is available at: `https://www.youtube.com/watch?v=XCEgyQKLaJ0`.

was to use radio signals to extract 3D human poses through walls and occlusions.

Predicting 3D human poses from radio signals led us to a new important challenge: *the computational complexity of learning from radio signals.* Since the depth information is in the phase of the signal, we had to approach RF signals as complex-valued tensors. We developed a novel CNN that leverages the properties of radio waves to decompose 4D CNN to 3D convolutions over 2D planes and the time axis. This method allowed us to maintain spatiotemporal relationships between human keypoints, yet operate on individual views of the signal over time, which reduced the computational complexity and allowed us to used common neural network platforms.

The resulting system RF-Pose3D tracks each keypoint on the human body with an average error of 4.2 cm, 4.0 cm, and 4.9 cm along the X, Y, and Z axes, respectively. Figure 1-2 shows a few example outputs of RF-Pose3D. It maintains this accuracy even in the presence of multiple people and in new environments not seen in the training set.

■ 1.1.3 Through-Wall Human Mesh Recovery

Next, we were interested in investigating whether radio signals can capture the shape of the human body (e.g., thin vs. heavy), not just a stick figure of it. We developed RF-Avatar, a neural network model that estimates 3D mesh representation of human bodies from radio signals. Our predicted meshes are dynamic and can smoothly track the movements of the corresponding people while distinguishing their body shapes, as shown in Figure 1-3.

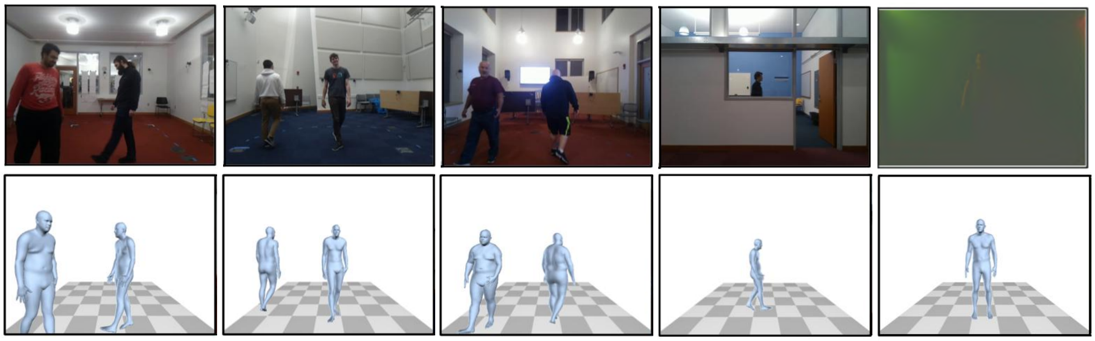


Figure 1-3: **Through-Wall Human Mesh Recovery.** RF-Avatar uses RF signals to estimate dynamic 3D human meshes consisting of 6890 vertices, even in the presence of occlusions and bad lighting conditions. 3D human meshes characterizing both pose and shape could enable new applications in gaming, fitness, and healthcare. **Top:** Images captured by cameras for visual reference. **Bottom:** Predicted human meshes from RF-Avatar using RF signals only. The figure shows that RF-Avatar works with different body shapes and sizes, and in the presence of partial and full occlusion.

The key new challenge we faced in inferring body meshes from radio signals is that this is a *highly under-constrained problem*. Unlike cameras which have millions of photodiode detectors and can obtain detailed spatial resolutions, a radio device at best has tens of antennas and a very limited spatial resolution. Our model deals with this challenge using: 1) a combination of strong and weak supervision, 2) a multi-headed self-attention mechanism that attends differently to temporal information in the radio signal, and 3) an adversarially trained temporal discriminator that imposes a prior on the dynamics of human motion. Our results show that RF-Avatar accurately recovers dynamic 3D meshes in the presence of occlusions, baggy clothes, bad lighting conditions, and even walls.

Collectively, RF-Pose, RF-Pose3D and RF-Avatar have transformed the state-of-the-art research on seeing through wall – from simple systems that rely on signal processing to detect people and track their position as a blob in space, to data-driven neural-network systems that capture rich information about dynamic human bodies through walls and occlusions.

■ 1.2 Contactless Health Monitoring

RF signals not only capture our activities and body shapes, but they also change with our physiological signals. Every small movement that we take leaves an imprint on the surrounding radio waves. Our breathing, the pulsing of our blood, even the twitching of our eye muscles as we dream during sleep are all encoded on radio waves that bounce off our bodies. As part of this dissertation, we examined novel systems that can extract health-related information from radio waves, and leverage them for health monitoring.

Remote in-home monitoring of people's vital signs, sleep, medication adherence, and emotional health is critical for the future of healthcare. It is motivated by skyrocketing costs, a limited access to healthcare in rural and disadvantaged socio-economic communities, and an aging population that often lives alone and is increasingly vulnerable due the COVID pandemic. However, existing remote monitoring solutions fail the very old, very sick, and people who have cognitive difficulties. Existing solutions typically require these groups to interact with advanced technology, including wearing sensors on one's body, and self-measuring and self-reporting their physiological signals. These tasks can be difficult for old sick people, who may have memory or cognition problems [37, 38].

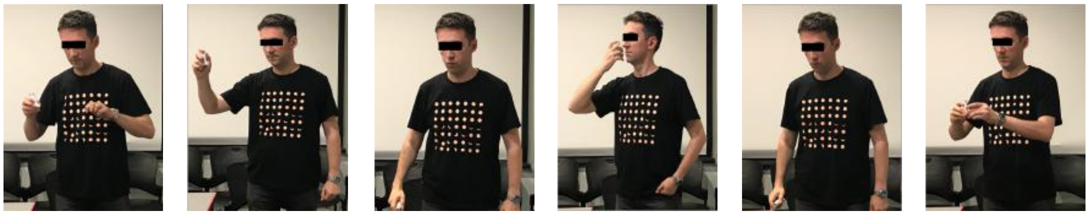
In this dissertation, we propose a system that passively monitors health at home while the residents go about their normal lives. The design is based on a wireless sensor that looks like a Wi-Fi router. It transmits very low-power radios signals (1000 times lower than standard Wi-Fi). The radio signals bounce off nearby people, and reflect off their bodies after being modulated by their movements and physiological signals. Our sensor analyses such radio reflections using novel algorithms and neural network models. It monitors people's vital signs and emotional status, tracks their sleep and sleep stages, and assesses medication administration. It does so without wearable sensors or body contact.

Below we describe the three components of our in-home health monitoring solution.

■ 1.2.1 Assessment of Medication Administration using Radio Signals

Poor medication adherence is a major healthcare problem, contributing to 10% of hospitalizations, 125,000 deaths per year, and up to $290 billion in annual cost in the United States alone [39, 40]. Medication errors are particularly common when medication delivery involves devices such as inhalers or insulin pens [41, 42, 43, 44]. Given our success in tracking human skeletons using radio signals, we were interested in checking whether one can use radio signals to detect when a patient uses their inhaler or insulin pen, and whether they use the device properly.

The task of assessing medication administration introduces new challenges beyond



Step 1: Pick up inhaler

Step 2: Shake inhaler

Step 3: Exhale before use

Step 4: Inhale dose & hold breatl

Step 5: Exhale

Step 6: Put down inhaler

Figure 1-4: **Key Constituent Steps of Inhaler Device Self-Administration.** We partitioned a medication self-administration event into key constituent steps based on recommendations pertaining to insulin pen and inhaler device administration. This allows us to build sample-efficient model to detect such events and to assess patient's administration techniques.

those introduced by human pose estimation. It requires *detection of a composite activity from a series of events.* Specifically, proper medication administration requires the patient to follow a sequence of steps for each medication device [45, 46]. For example, Figure 1-4 shows the recommended steps a patient should follow when using an inhaler: 1) shake the inhaler, 2) fully exhale, 3) inhale a dose, 4) hold breath for 10 seconds, 5) exhale, then 6) put down the inhaler [45]. Similarly, the proper use of an insulin pen requires following certain steps [46]. Ensuring the patient follows the proper steps is essential; otherwise the patient may take the medication but fails to deliver the drug or obtain the proper dose.

While there is a significant amount of work on activity detection in videos, our approach differs from past work in two ways. First we detect activities from radio signals. Second, we are interested in a sequence of actions/steps that together deliver the activity of interest. Past methods for understanding human activities typically focus on short and simple actions, e.g., drinking, dancing, waving hands, etc. [47, 48, 49]. Such methods face challenges when applied to composite human activities consisting of multiple actions or steps. Specifically, the intra-class variation of composite activities increases drastically considering the variation within individual steps, as well as the varied pause between steps. As a result, a prohibitively large number of samples are required to train a model that could capture the intra-class variation within the composite activities. It is even more challenging if the composite activities involve patients (e.g., medication self-administration) due to limited amount of health-related data.

To design a sample-efficient model, we used neural discriminative models for recognizing key constituent steps for medication self-administration, and Bayesian inference to model how those steps construct the full composite activities. Findings from our study

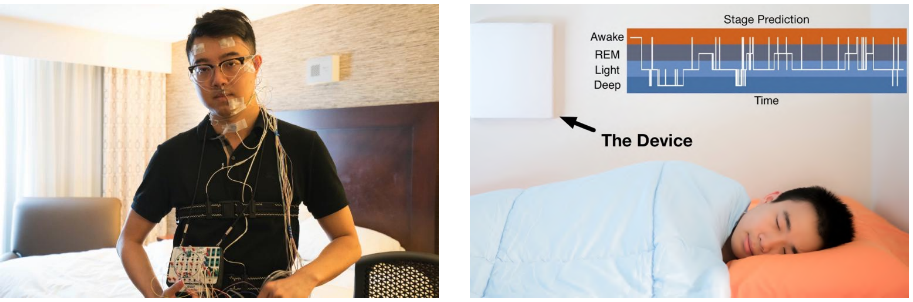


Figure 1-5: **Sleep Stages Monitoring. Left:** Existing methods require patients to sleep in hospitals wearing various sensors. **Right:** RF-Sleep enables accurate monitoring of sleep stages, while the person sleeps in their own bed without wearing any sensors. Video is available at: https://www.youtube.com/watch?v=ltcjly-CYkI

demonstrated that our approach can automatically detect when patients use their inhalers (area under the curve (AUC) = 0.992) or insulin pens (AUC = 0.967), and assess whether they follow the appropriate steps for using these devices (AUC = 0.952).

■ 1.2.2 Contactless Monitoring of Sleep Stages

Sleep plays a vital role in an individual's health and wellbeing. Sleep progresses in cycles that involve different sleep stages: Awake, Light sleep, Deep sleep and REM (Rapid Eye Movement). Different stages are associated with different physiological functions. For example, REM is the stage in which we dream and is essential for emotional health. Memory consolidation happens during deep sleep and is related to diseases like Alzheimer's. Monitoring sleep stages typically requires a person to spend the night in a hospital or sleep lab, sleeping with EEG electrodes and other sensors on their bodies. A sleep technician manually analyses the resulting EEG signals and labels every 30-second episode with a sleep stage.

We have developed a novel neural network model that takes as input the radio signals that bounce off a person's body while asleep, and outputs for each 30-second episode the person's sleep stage. We compared our model to an EEG-based FDA-approved sleep stage monitor and showed that it achieves high accuracy, comparable to the consistency between two sleep technicians analyzing the same EEG signals to predict sleep stages.

The key challenge in predicting sleep stages from radio signals is that *radio waves carry much extraneous information that is irrelevant to the task of interest*. They have information about body shape and posture in addition to information about other people and objects

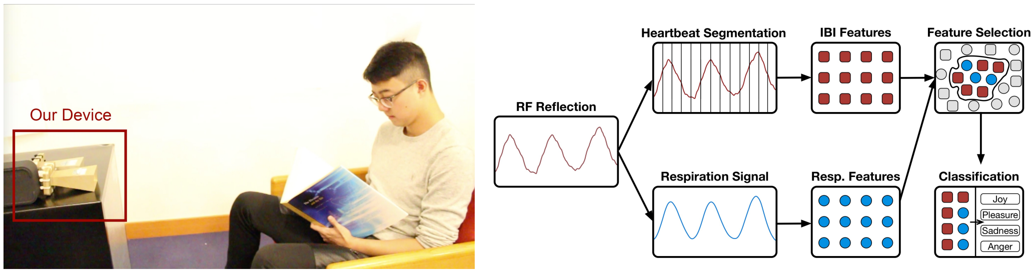


Figure 1-6: **Contactless Monitoring of Emotions and Vital Signs.** EQ-Radio extracts breathing and heart rate variability from RF signals with millisecond-level accuracy, enabling contactless emotion recognition. Full video is available at: https://www.youtube.com/watch?v=nmcDnEhZTJM

in the environment. This information can overwhelm the model and prevent it from generalizing to new people and new homes. To address this problem we introduced a new predictive model based on conditional adversarial architecture, and analytically proved that our model in equilibrium converges to eliminate all extraneous information, while maintaining all information relevant to the predictive task.

■ 1.2.3 Contactless Monitoring of Emotions and Vital Signs

As part of this dissertation, we developed a new technology that can infer a person's emotions from RF signals reflected off their body. As shown in Figure 1-6, EQ-Radio transmits an RF signal and analyzes its reflections off a person's body to extract their breathing and heart rate variability, which it then processes via a machine learning model to infer the person's emotional state (happy, sad, angry, etc.). The operation of EQ-Radio intrinsically depends on extremely accurate estimation of the length of each heart beat [50, 51, 52]. For example, excitement causes our heart to beat faster, while sadness causes our heartbeats to become more monotonic (i.e., less heart rate variability). Yet to capture these features, our system needs to estimate the length of each heart beat to within a few milliseconds. In contrast to ECG signals which have a known pattern, the shape of a heartbeat in RF reflections is unknown and varies depending on the person's body and exact posture with respect to the radio. Thus we faced a new challenge: *extracting a weak and unknown pattern from noisy radio signals.*

To address this challenge, we formulated the problem as a joint optimization, where we iterate between two sub-problems. The first sub-problem learns a template of the shape of the heartbeat in RF signals given a particular beat segmentation, while the second finds the segmentation that maximizes resemblance to the learned template. We keep iterating between the two sub-problems until we converge to the best beat template and the optimal segmentation that maximizes resemblance to the template. Our experiments show that EQ-Radio is on par with state-of-the-art ECG-based emotion recognition systems, which require on-body sensors [53]. Specifically the accuracy of emotion classification is 72.3% in EQ-Radio and 73.2% in the ECG-based system [53].

1.3 Summary of Contributions

Below we summarize the contributions of this dissertation.

- This dissertation delivers the first systems that can sense human poses and meshes through walls and occlusions. While seeing through obstacles with radio signals has been a research topic for three decades, none of the prior systems can extract the human pose and its keypoints from radio signals. Our systems were enabled through novel neural network designs that address the challenges faced when operating neural networks on RF signals. The techniques we have developed to deal with RF specularity, high computation complexity, and highly under-constrained sensing problems, as well as the cross-modal training scheme, have been used by researchers to address other RF-based sensing tasks [33, 34, 35, 36, 30, 31, 32, 54, 55, 56, 57, 58, 59, 60]. Beyond its contributions to wireless sensing, this research also advances the state of computer vision by enabling pose estimation in the presence of occlusions and bad lighting conditions.
- This dissertation introduces the first solution for assessing medication self-administration at home using radio signals. This solution can accurately detect when patients use their inhalers or insulin pens. It is also the first solution to automatically detect whether patients follow the proper technique for using these medication delivery devices (e.g., shake the inhaler and prime the insulin pen before use).
- This dissertation develops RF-Sleep, a system that can infer sleep stages from radio signals with accuracy comparable to FDA-approved EEG sleep monitors. It introduces a new machine learning method for domain adaptation based on conditional domain adversarial training.
- This dissertation presents EQ-Radio, the first system that demonstrates the feasibility of emotion recognition using RF signals. It also introduces a new algorithm for extract-

ing individual heartbeats from RF reflections with millisecond-level inter-beat-interval accuracy.

Finally, we note that all the studies involving human subjects in this dissertation have been reviewed and approved by the Institutional Review Board (IRB) of the Massachusetts Institute of Technology.

Chapter 2
Related Work

The research presented in this thesis is naturally related to wireless and RADAR systems (Section 2.1), as well as machine learning and computer vision (Section 2.2). We also summarize other mechanisms that have been used for human and health sensing.

2.1 Wireless and RADAR systems

RF signals reflect off the human body and are modulated by body motion. Past work leverages this phenomenon to sense human motion: it transmits an RF signal and analyzes its reflections to track user locations [8], gestures [7, 61, 62, 63, 64, 65], activities [66, 67], and vital signs [68, 69, 70]. Past proposals also differ in the transmitted RF signals: Doppler RADAR [69, 70], frequency-modulated continuous wave (FMCW) [68], multi-antenna FMCW [8, 9, 71] and WiFi [7, 61]. Among these techniques, multi-antenna FMCW has the advantage of measuring the signal from different 3D voxels in the environment. Thus, multi-antenna FMCW is more robust for capturing body movements and extracting vital signs for multiple users simultaneously; hence, we use FMCW signals for the systems in this thesis.

Through-wall vision: Our work on through-wall vision is related to research on localizing people and tracking their motion using wireless signals. The literature can be classified into two categories. The first category operates at very high frequencies (e.g., millimeter wave or terahertz) [72]. These can accurately image the surface of the human body (as in airport security scanners), but do not penetrate walls and furniture. The second category uses lower frequencies, around a few GHz, and hence can track people through walls and occlusions. Such through-wall tracking systems can be divided into: device-based and device-free. Device-based tracking systems localize people using the signal generated by some wireless device they carry. For example, one can track a person using the WiFi signal from their cellphone [73, 74, 75]. Since the tracking is performed on the device not the person, one can track different body parts by attaching different radio devices to each of them. On the other hand, device-free wireless tracking systems do not require the tracked person to wear sensors on their body. They work by analyzing the radio signal reflected off the person's body. However, device-free systems typically have low spatial resolution and cannot localize multiple body parts simultaneously. Different papers either localize the whole body [8, 10], monitor the person's walking speed [14, 15], track the chest motion to extract breathing and heartbeats [68, 6, 5], or track the arm motion to identify a particular gesture [61, 76]. The closest to our work is a system called RF-Capture which creates a coarse description of the human body behind a wall by collapsing multiple body parts detected at different points in time [71]. None of the past work however is capable of estimating the human pose or simultaneously localizing its various keypoints.

Contactless health monitoring: Our work on contactless health monitoring is related to prior arworkt that uses RF signals to extract a person's breathing rate and average heart rate [69, 70, 77, 78, 79, 69, 70, 77, 78, 79, 80, 81, 68]. In contrast to this past work, which captures the chest movements due to breathing and heartbeats and derives breathing and heart rates, the systems presented in this thesis either capture vital signs with finer granularity (i.e., average heart rate vs. beat-to-beat intervals with millisecond-level accuracy) or extract semantic information based on these measurements (e.g., sleep stages, medication adherence). More specifically, prior research that aims to segment RF reflections into individual heart beats either cannot achieve sufficient accuracy for emotion recognition [82, 83, 84] or requires the monitored subjects to hold their breath [85]. In Chapter 8, we describe a heartbeat segmentation algorithm that builds on this past literature yet recovers heartbeats with a mean accuracy of 3.2 milliseconds, hence achieving an order of magnitude reduction in errors in comparison to past techniques. This high accuracy enables us to deliver the first emotion recognition system that relies purely on wireless signals.

■ 2.2 Machine Learning and Computer Vision

Human pose estimation: Inferring the human pose from images is a known problem in the computer vision literature. The problem comes in two flavors: 2D and 3D. 2D pose estimation has achieved remarkable success recently [86, 87, 20, 21, 88, 22, 23]. This is due to the availability of large-scale datasets of annotated 2D human poses, and the introduction of deep neural network models. In contrast, advances in 3D human pose estimation remain limited due to the difficulty and ambiguity of recovering 3D information from 2D images.

In terms of methods, human pose estimation in images generally falls into two main categories: top-down and bottom-up methods. Top-down methods [23, 86, 89, 87] first detect each person in the image, and then apply a single-person pose estimator to each person to extract keypoints. Bottom-up methods [22, 20, 21], on the other hand, first detect all keypoints in the image, then use post-processing to associate the keypoints belonging to the same person. We build on this literature and adopt both bottom-up (Chapter 3) and top-down (Chapter 4) approaches for pose estimation, but differ in that we learn poses from RF signals. While some prior papers use sensors other than conventional cameras, such as RGB-D sensors [90] and VICON [91], unlike RF signals, those data inputs still suffer from occlusions by walls and other opaque structures.

Our work builds on human pose estimation in computer vision in three ways. First, we similarly use deep neural networks to address this problem. Second, we leverage a vision system called OpenPose [22] to extract 2D skeletons from images. We further integrate this module in our camera system, which combines such 2D skeletons across 12 cameras to create 3D skeletons that can be used as training examples for our network. Third, our module that zooms in on the RF signal from a particular individual and separates it from the signals from other individuals is inspired by object detection in computer vision, specifically systems like R-CNN, Fast R-CNN and Faster R-CNN [92, 93, 94] which use deep neural models to generate a bounding box around objects of interest in an image (e.g., a dog).

Our work, however, is fundamentally different from all past work in computer vision. We infer 3D poses from RF signals, which is intrinsically different from extracting 3D poses from images due to basic differences between the two data types. In particular, images have high spatial resolution whereas RF signals have low spatial resolution, even when using multi-antenna systems. Second, the human body scatters visible light, but acts as a reflector for the RF bands of interest (frequencies around few GHz) [95]. Hence, only signals that fall close to the normal on the body surface are reflected back towards the radio source. As a result, at any time, only a few body parts are visible to the radio [71]. Furthermore, our neural network model differs from past work in vision, and is the first to propose 4D CNN decomposition. Even our dataset is different. Existing datasets for inferring 3D poses from images are limited to one environment or one person (e.g., Human3.6M [96]). In contrast, our dataset spans multiple environments and our scenes include multiple people. This allows our system to learn to generalize to new environments which are unseen during training.

Capturing human shapes: There are broadly two methods used to capture body shape in prior work. In scanning-based methods, several images of a subject are obtained, typically in a canonical pose, and then optimization-based methods are used to recover the SCAPE or SMPL parameters representing the subject's shape. The authors of [97, 98, 99, 100, 101] used scanning approaches, incorporating silhouette information and correspondence cues to fit a SCAPE or SMPL model. However, scanning-based methods have the inherent limitation that they can be easily affected by clothing, so they only work well when subjects are in form-fitting clothes. They are also limited to indoor settings and do not properly capture motion dynamics. Thus, many recent works, including ours, use scanning methods only to provide supervision to learning-based methods.

In learning-based methods, models are trained to predict parameters of a shape model (e.g., SMPL). Such methods are challenging due to the lack of 3D human mesh dataset. Despite this, there has been significant success in this area. Bogo et al. [102] proposed a two-stage process to firstly predict joint locations and then fit SMPL parameters from a 2D image. Lassner et al. [103] developed on this approach, incorporating a semi-automatic annotation scheme to improve scalability. More recent work [104, 105] captured 3D meshes from 2D images using adversarial loss, and Kanazawa et al. [106] learned dynamic 3D meshes using videos as an additional data source. In Chapter 5, we adopt a learning-based approach, building on the above literature, and expanding it to deal with scenarios with occlusions and bad lighting.

Domain adversarial training: Our work on domain adversarial training that ensures our

RF sensing models to generalize across environments and individuals (Chapter 7) is also related to learning invariant representations in deep adversarial networks. Adversarial networks were introduced to effectively train complex generative models of images [107, 108, 109] where the adversary (discriminator) was introduced so as to match generated samples with observed ones. The broader approach has since been adopted as the training paradigm across a number of other tasks as well, from learning representations for semi-supervised learning [110], and modeling dynamic evolution [111, 112] to inverse maps for inference [113, 114], and many others. Substantial work has also gone into improving the stability of adversarial training [115, 116, 117]. On a technical level, this work is most related to adversarial architectures for domain adaptation [118, 119, 120, 121]. Yet, there are key differences between our approach and the above references, beyond the main application of sleep staging that we introduce. First, our goal is to remove conditional dependencies rather than making the representation domain independent. Thus, unlike the above references which do not involve conditioning in the adversary, our adversary takes the representation but is also conditioned on the predicted label distribution. Second, our game theoretic setup controls the information flow differently, ensuring that only the representation encoder is modified based on the adversary performance. Specifically, the predicted distribution over stages is strategically decoupled from the adversary (conditioning is uni-directional). Third, we show that this new conditioning guarantees an equilibrium solution that fully preserves the ability to predict sleep staging while removing, conditionally, extraneous information specific to the individuals or measurement conditions. Guarantees of this kind are particularly important for healthcare data where the measurements are noisy with a variety of dependencies that need to be controlled.

■ 2.3 Other sensing mechanisms

Smart sensors for assessing medication self-administration: Past solutions for assessing medication self-administration (MSA) at home attach sensors to medication devices to monitor MSA [122, 123, 124]. However, these solutions can impose a new burden on the patient, as they require the patient to regularly charge or replace their battery and bring the devices in the vicinity of a smartphone so they can upload their data. Although such solutions can detect dose release, they lack information on whether the patient followed the proper MSA technique to ensure adequate dose delivery – that is, the sensor captures the actuation and movements of the medication device itself but cannot capture the patient's actions and their sequence, which are crucial for correct MSA. To our knowledge, our work (Chapter 6) is the first to introduce an automated solution for assessing an individual's MSA technique and whether it follows the proper steps. Being able to assess MSA techniques is essential because failures to follow the proper techniques are common and have been associated with high non-adherence levels and subsequent poor disease outcomes [125, 126, 127, 128].

Polysomnography for sleep staging: The gold standard in sleep staging is based on Polysomnography (PSG) conducted overnight in a hospital or sleep lab. The subject has to sleep while wearing multiple sensors including an EEG monitor, an EMG monitor, an EOG monitor, nasal probes, etc. A sleep technologist visually inspects the output of the sensors and assigns to each 30-second window a stage label [129]. A few past proposals have tried to automate the process and reduce the number of sensors. These solutions can be classified into four categories according to their source signal: EEG-based, Cardiorespiratory-based, Actigraphy-based, or RF-based. Table 7-1 summarizes the state-of-the-art performance in each category. The table shows both the classification accuracy and the Cohen's Kappa coefficient, κ. The most accurate methods rely on EEG signals [130, 131, 132, 133]. However, EEG monitors are also the most intrusive because they require the subject to sleep with a skullcap or a head-band equipped with multiple electrodes, which is uncomfortable and can cause headaches and skin irritation. The second category requires the subject to wear a chest-band and analyzes the resulting cardiorespiratory signals. It is more comfortable than the prior method but also less accurate [134, 135]. The third approach is based on actigraphy; it leverages accelerometers in FitBit or smart phones [136, 137] to monitor body movements and infer sleep quality. Yet, motion is known to be a poor metric for measuring sleep stages and quality [138]. The last approach relies on RF signals reflected off the subject body during her sleep. It allows the subject to sleep comfortably without any on-body sensors. Yet past approaches in this category have the worst performance in comparison to other solutions. Our approach to sleep monitoring builds on the above literature but delivers new contributions. In comparison to methods that use sources other than RF signals, our work enables accurate monitoring of sleep stages while allowing the subject to sleep comfortably in her own bed without sensors on her body.

Audiovisual and physiological techniques for Emotion Recognition: Recent years have witnessed a growing interest in systems capable of inferring user emotions and reacting to them [139, 140]. Such systems can be used for designing and testing games, movies, advertisement, online content, and human-computer interfaces [141, 142]. Existing approaches for extracting emotion-related signals fall under two categories: audiovisual techniques and physiological techniques. Audiovisual techniques rely on facial expressions, speech, and gestures [143, 144]. The advantage of these approaches is that they do not require users to wear any sensors on their bodies. However, because they rely on outwardly expressed states, they tend to miss subtle emotions and can be easily controlled or suppressed [145]. Moreover, vision-based techniques require the user to face a camera in order for them to operate correctly. On the other hand, physiological measurements, such as ECG and EEG signals, are more robust because they are controlled by involuntary activations of the autonomic nervous system (ANS) [146]. However, existing sensors that can extract these signals require physical contact with a person's body, and hence interfere with the user experience and affect her emotional state. In contrast, our approach can capture physiological signals without requiring the user to wear any sensors by relying purely on wireless signals reflected off her/his body.

Chapter 3

Through-Wall 2D Human Pose Estimation using Radio Signals

Estimating the human pose is an important task in computer vision with applications in surveillance, activity recognition, gaming, etc. The problem is defined as generating 2D skeletal representations of the joints on the arms and legs, and keypoints on the torso and head. It has recently witnessed major advances and significant performance improvements [16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29]. However, as in any camera-based recognition task, occlusion remains a fundamental challenge. Past work deals with occlusion by hallucinating the occluded body parts based on the visible ones. Yet, since the human body is deformable, such hallucinations are prone to errors. Further, this approach becomes infeasible when the person is fully occluded, behind a wall or in a different room.

This section presents a fundamentally different approach to deal with occlusions in pose estimation, and potentially other visual recognition tasks. While visible light is easily blocked by walls and opaque objects, radio frequency (RF) signals in the WiFi range can traverse such occlusions. Further, they reflect off the human body, providing an opportunity to track people through walls. Recent advances in wireless systems have leveraged those properties to detect people [8] and track their walking speed through occlusions [15]. Past systems however are quite coarse: they either track only one limb at any time [8, 71], or generate a static and coarse description of the body, where body-parts observed at different time are collapsed into one frame [71]. Use of wireless signals to produce a detailed

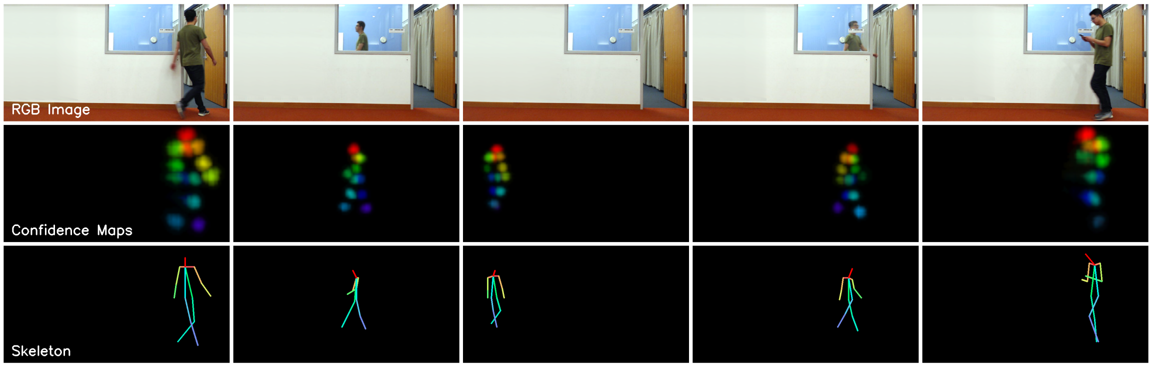


Figure 3-1: **An example output of RF-Pose.** The figure shows a test example with a single person. It demonstrates that our system tracks the pose as the person enters the room and even when he is fully occluded behind the wall. **Top:** Images captured by a camera colocated with the radio sensor, and presented here for visual reference. **Middle:** Keypoint confidence maps extracted from RF signals *alone*, without any visual input. **Bottom:** Skeleton parsed from keypoint confidence maps showing that we can use RF signals to estimate the human pose even in the presence of full occlusion.

and accurate description of the pose, similar to that achieved by a state-of-the-art computer vision system, has remained intractable.

In this section, we introduce RF-Pose, a neural network system that parses wireless signals and extracts accurate 2D human poses, even when the people are occluded or behind a wall. RF-Pose transmits a low power wireless signal (1000 times lower power than WiFi) and observes its reflections from the environment. Using only the radio reflections as input, it estimates the human skeleton. Fig. 3-1 shows an example output of RF-Pose tracking a person as he enters a room, becomes partially visible through a window, and then walks behind the wall. The RGB images in the top row show the sequence of events and the occlusions the person goes through; the middle row shows the confidence maps of the human keypoints extracted by RF-Pose; and the third row shows the resulting skeletons. Note how our pose estimator tracks the person even when he is fully occluded behind a wall. While this example shows a single person, RF-Pose works with multiple people in the scene, just as a state-of-art vision system would.

The design and training of our network present different challenges from vision-based pose estimation. In particular, there is no labeled data for this task. It is also infeasible for humans to annotate radio signals with keypoints. To address this problem, we use cross-modal supervision. During training, we attach a web camera to our wireless sensor, and synchronize the the wireless and visual streams. We extract pose information from the visual stream and use it as a supervisory signal for the wireless stream. Once the system is trained, it only uses the radio signal as input. The result is a system that is capable of estimating human pose using wireless signals only, without requiring human annotation as supervision. Interestingly, the RF-based model learns to perform pose estimation even when the people are fully occluded or in a different room. It does so despite it has never seen such examples during training.

Beyond cross-modal supervision, the design of RF-Pose accounts for the intrinsic features of RF signals including low spatial resolution, specularity of the human body at RF frequencies that traverse walls, and differences in representation and perspective between RF signals and the supervisory visual stream.

We train and test RF-Pose using data collected in public environments around our campus. The dataset has hundreds of different people performing diverse indoor activities: walking, sitting, taking stairs, waiting for elevators, opening doors, talking to friends, etc. We test and train on different environments to ensure the network generalizes to new scenes. We manually label 2000 RGB images and use them to test both the vision system and RF-Pose. The results show that on visible scenes, RF-Pose has an average precision (AP) of 62.4 whereas the vision-based system used to train it has an AP of 68.8. For through-wall scenes, RF-Pose has an AP of 58.1 whereas the vision-based system fails completely.

We also show that the skeleton learned from RF signals extracts identifying features of the people and their style of moving. We run an experiment where we have 100 people perform free walking, and train a vanilla-CNN classifier to identify each person using a 2-second clip of the RF-based skeleton. By simply observing how the RF-based skeleton moves, the classifier can identify the person with an accuracy over 83% in both visible and through wall scenarios.

■ 3.1 RF Signals Acquisition and Properties

Our RF-based pose estimation relies on transmitting a low power RF signal and receiving its reflections. To separate RF reflections from different objects, it is common to use techniques like FMCW (Frequency Modulated Continuous Wave) and antenna arrays [71]. FMCW separates RF reflections based on the distance of the reflecting object, whereas antenna arrays separate reflections based on their spatial direction. In this section, we introduce a radio similar to [71], which generates an FMCW signal and has two antenna arrays: vertical and horizontal (other radios are also available [147, 148]). Thus, our input data takes the form of two-dimensional heatmaps, one for each of the horizontal and vertical antenna arrays. As shown in Fig. 3-2, the horizontal heatmap is a projection of the signal reflections on a plane parallel to the ground, whereas the vertical heatmap is a projection of the reflected signals on a plane perpendicular to the ground (red refers to large values while blue refers to small values). Note that since RF signals are complex numbers, each pixel in this map has a real and imaginary components. Our radio generates 30 pairs of heatmaps per second.

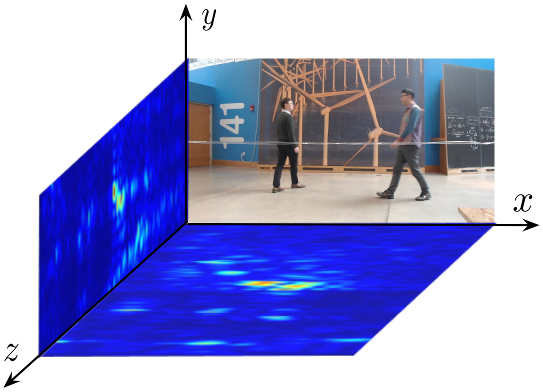


Figure 3-2: **RF heatmaps and an RGB image recorded at the same time.**

It is important to note that RF signals have intrinsically different properties than visual data, i.e., camera pixels.

- First, RF signals in the frequencies that traverse walls have low spatial resolution, much lower than vision data. The resolution is typically tens of centimeters [8, 148, 71], and is defined by the bandwidth of the FMCW signal and the aperture of the antenna array. In particular, our radio has a depth resolution about 10 cm, and its antenna arrays have vertical and horizontal angular resolution of 15 degrees.
- Second, the human body is specular in the frequency range that traverse walls [95]. RF specularity is a physical phenomenon that occurs when the wavelength is larger than the roughness of the surface. In this case, the object acts like a reflector - i.e., a mirror - as opposed to a scatterer. The wavelength of our radio is about 5cm and hence humans act as reflectors. Depending on the orientation of the surface of each limb, the signal may be reflected towards our sensor or away from it. Thus, in contrast to camera systems

where any snapshot shows all unoccluded key-points, in radio systems, a single snapshot has information about a subset of the limbs and misses limbs and body parts whose orientation at that time deflects the signal away from the sensor.

- Third, the wireless data has a different representation (complex numbers) and different perspectives (horizontal and vertical projections) from a camera.

The above properties have implications for pose estimation, and need to be taken into account in designing a neural network to extract poses from RF signals.

3.2 Method

Our model, illustrated in Fig. 3-3, follows a teacher-student design. The top pipeline in the figure shows the teacher network, which provides cross-modal supervision; the bottom pipeline shows the student network, which performs RF-based pose estimation.

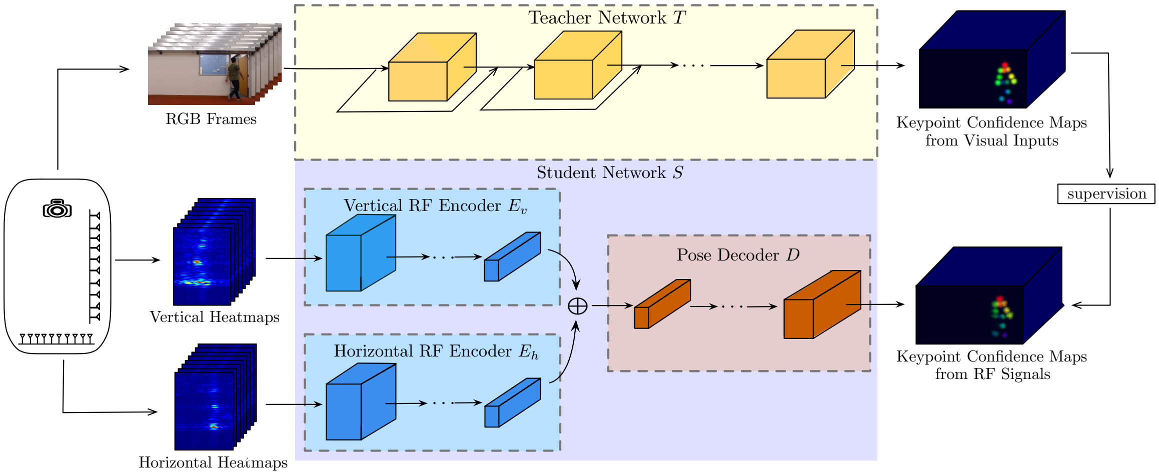

Teacher Network T
RGB Frames
Keypoint Confidence Maps from Visual Inputs
Student Network S
supervision
Vertical RF Encoder E_v
Pose Decoder D
Vertical Heatmaps
Horizontal RF Encoder E_h
Keypoint Confidence Maps from RF Signals
Horizontal Heatmaps

Figure 3-3: **Our teacher-student network model used in RF-Pose.** The upper pipeline provides training supervision, whereas the bottom pipeline learns to extract human pose using only RF heatmaps.

3.2.1 Cross-Modal Supervision

One challenge of estimating human pose from RF signals is the the lack of labelled data. Annotating human pose by looking at RF signals (e.g., Fig. 3-2) is almost impossible. We address this challenge by leveraging the presence of well established vision models that are trained to predict human pose in images [149, 150].

We design a cross-modal teacher-student network that transfers the visual knowledge of human pose using synchronized images and RF signals as a bridge. Consider a synchronized pair of image and RF signals $(\mathbf{I}, \mathbf{R})$, where $\mathbf{R}$ denotes the combination of the vertical and horizontal heatmaps, and $\mathbf{I}$ the corresponding image in Fig. 3-2. The teacher network $\mathbf{T}(\cdot)$ takes the images $\mathbf{I}$ as input and predicts keypoint confidence maps as $\mathbf{T}(\mathbf{I})$. These predicted maps $\mathbf{T}(\mathbf{I})$ provide cross-modal supervision for the student network $\mathbf{S}(\cdot)$, which learns to predict keypoint confidence maps from the RF signals. In this work, we adopt the 2D pose estimation network in [22] as the teacher network. The student network learns to predict 14 keypoint confidence maps corresponding to the following anatomical parts of the human body: head, neck, shoulders, elbows, wrists, hips, knees and ankles.

The training objective of the student network $\mathbf{S}(\cdot)$ is to minimize the difference between its prediction $\mathbf{S}(\mathbf{R})$ and the teacher network's prediction $\mathbf{T}(\mathbf{I})$:

$$\min_{\mathbf{S}} \sum_{(\mathbf{I},\mathbf{R})} L(\mathbf{T}(\mathbf{I}), \mathbf{S}(\mathbf{R})) \tag{3.1}$$

We define the loss as the summation of binary cross entropy loss for each pixel in the confidence maps:

$$L(\mathbf{T}, \mathbf{S}) = -\sum_{c} \sum_{i,j} \mathbf{S}_{ij}^{c} \log \mathbf{T}_{ij}^{c} + (1 - \mathbf{S}_{ij}^{c}) \log (1 - \mathbf{T}_{ij}^{c}),$$

where $\mathbf{T}_{ij}^{c}$ and $\mathbf{S}_{ij}^{c}$ are the confidence scores for the (i, j)-th pixel on the confidence map c.

■ 3.2.2 Keypoint Detection from RF Signals

The design of our student network has to take into account the properties of RF signals. As mentioned earlier, the human body is specular in the RF range of interest. Hence, we cannot estimate the human pose from a single RF frame (a single pair of horizontal and vertical heatmaps) because the frame may be missing certain limbs tough they are not occluded. Further, RF signals have low spatial resolution. Hence, it will be difficult to pinpoint the location of a keypoint using a single RF frame. To deal with these issues, we make the network learn to aggregate information from multiple snapshots of RF heatmaps so that it can capture different limbs and model the dynamics of body movement. Thus, instead of taking a single frame as input, we make the network look at sequences of frames. For each sequence, the network outputs keypoint confidence maps as the number of frames in the input – i.e., while the network looks at a clip of multiple RF frames at a time, it still outputs a pose estimate for every frame in the input.

We also want the network to be invariant to translations in both space and time so that it can generalize from visible scenes to through-wall scenarios. Therefore, we use spatio-temporal convolutions [151, 152, 48] as basic building blocks for the student networks.

Finally, the student network needs to transform the information from the views of RF heatmaps to the view of the camera in the teacher network (see Fig. 3-2). To do so, the model has to first learn a representation of the information in the RF signal that is not encoded in original spatial space, then decode that representation into keypoints in the view of the camera. Thus, as shown in Fig. 3-3, our student network has: 1) two RF encoding networks $E_h(\cdot)$ and $E_v(\cdot)$ for horizontal and vertical heatmap streams, and 2) a pose decoding network $D(\cdot)$ that takes a channel-wise concatenation of horizontal and vertical RF encodings as input and predicts keypoint confidence maps. The RF encoding networks uses strided convolutional networks to remove spatial dimensions [153, 111] in order to summarize information from the original views. The pose decoding network then uses fractionally strided convolutional networks to decode keypoints in the camera's view.

3.2.3 Implementation and Training

RF encoding network. Each encoding network takes 100 frames (3.3 seconds) of RF heatmap as input. The RF encoding network uses 10 layers of $9 \times 5 \times 5$ spatio-temporal convolutions with $1 \times 2 \times 2$ strides on spatial dimensions every other layer. We use batch normalization [154] followed by the ReLU activation functions after every layer.

Pose decoding network. We combine spatio-temporal convolutions with fractionally strided convolutions to decode the pose. The decoding network has 4 layers of $3 \times 6 \times 6$ with fractionally stride of $1 \times \frac{1}{2} \times \frac{1}{2}$, except the last layer has one of $1 \times \frac{1}{4} \times \frac{1}{4}$. We use Parametric ReLu [155] after each layer, except for the output layer, where we use sigmoid.

Training Details. We represent a complex-valued RF heatmap by two real-valued channels that store the real and imaginary parts. We use a batch size of 24. Our networks are implemented in PyTorch.

■ 3.2.4 Keypoint Association

The student network generates confidence maps for all keypoints of all people in the scene. We map the keypoints to skeletons as follows. We first perform non-maximum suppression on the keypoint confidence maps to obtain discrete peaks of keypoint candidates. To associate keypoints of different persons, we use the relaxation method proposed by Cao et al. [22] and we use Euclidean distance for the weight of two candidates. Note that we perform association on a frame-by-frame basis based on the learned keypoint confidence maps. More advanced association methods are possible, but outside the scope of this thesis.

■ 3.3 Dataset

We collected synchronized wireless and vision data. We attached a web camera to our RF sensor and synchronized the images and the RF data with an average synchronization error of 7 milliseconds.

We conducted more than 50 hours of data collection experiments from 50 different environments (see Fig. 3-4), including different buildings around our campus. The environments span offices, cafeteria, lecture and seminar rooms, stairs, and walking corridors. People performed natural everyday activities without any interference from our side. Their activities include walking, jogging, sitting, reading, using mobile phones and laptops, eating, etc. Our data includes hundreds of different people of varying ages. The maximum and average number of people in a single frame are 14 and 1.64, respectively. A data frame can also be empty, i.e., it does not include any person. Partial occlusions, where parts of the human body are hidden due to furniture and building amenities, are also present. Legs and arms are the most occluded parts.

To evaluate the performance of our model on through-wall scenes, we build a mobile camera system that has 8 cameras to provide ground truth when the people are fully occluded. After calibrating the camera system, we construct 3D poses of people and project them on the view of the camera colocated with RF sensor. The maximum and average number of people in each frame in the through-wall testing set are 3 and 1.41, respectively. This through-wall data was *only for testing* and was not used to train the model.

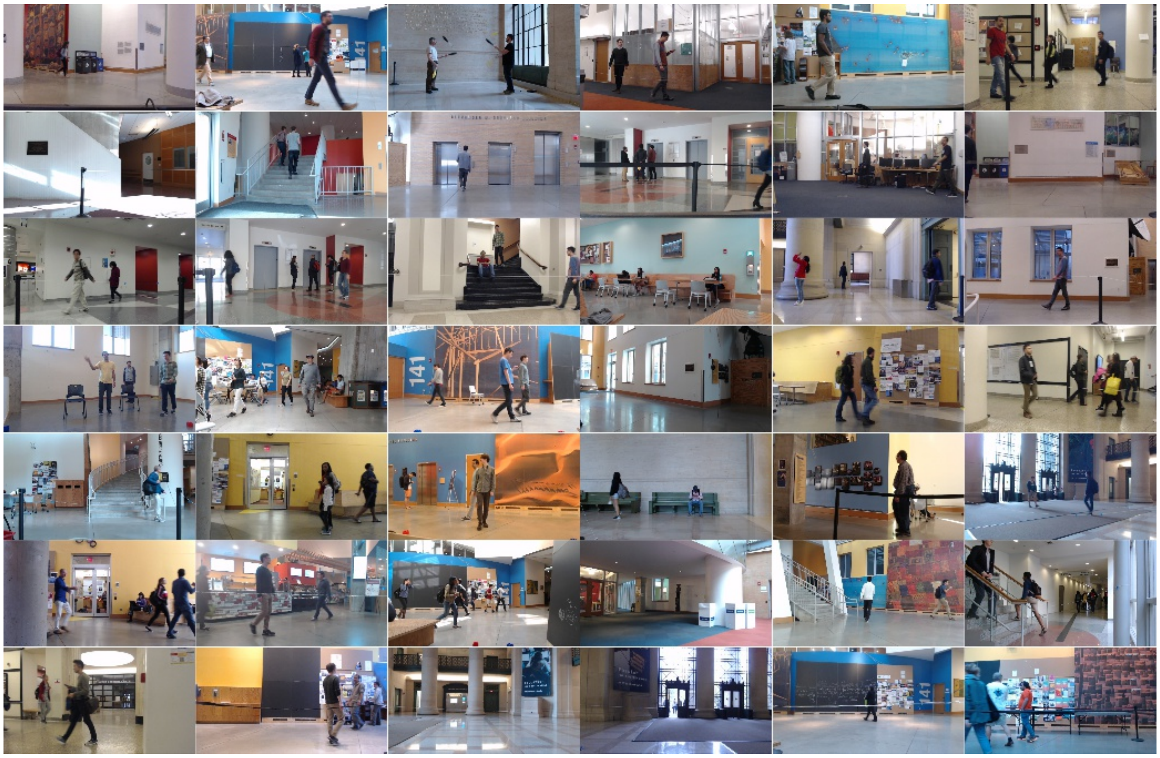


Figure 3-4: **Different environments in the dataset.**

3.4 Experiments

RF-Pose is trained with 70% of the data from visible scenes, and tested with the remaining 30% of the data from visible scenes and all the data from through-wall scenarios. We make sure that the training data and test data are from different environments.

3.4.1 Setup

Evaluation Metrics: Motivated by the COCO keypoints evaluation [149] and as is common in past work [22, 89, 23], we evaluate the performance of our model using the average precision over different object keypoint similarity (OKS). We also report AP^{50} and AP^{75}, which denote the average precision when OKS is 0.5 and 0.75, and are treated as loose and strict match of human pose, respectively. We also report AP, which is the mean average precision over 10 different OKS thresholds ranging from 0.5 to 0.95.

Baseline: For visible and partially occluded scenes, we compare RF-Pose with OpenPose [22], a state-of-the-art vision-based model, that also acts as the teacher network.

Ground Truth: For visible scenes, we manually annotate human poses using the images

Methods	Visible scenes			Through-walls		
	AP	AP^{50}	AP^{75}	**AP**	AP^{50}	AP^{75}
RF-Pose	62.4	**93.3**	70.7	**58.1**	**85.0**	**66.1**
OpenPose[22]	**68.8**	77.8	**72.6**	-	-	-

Table 3-1: **Average precision in visible and through-wall scenarios.**

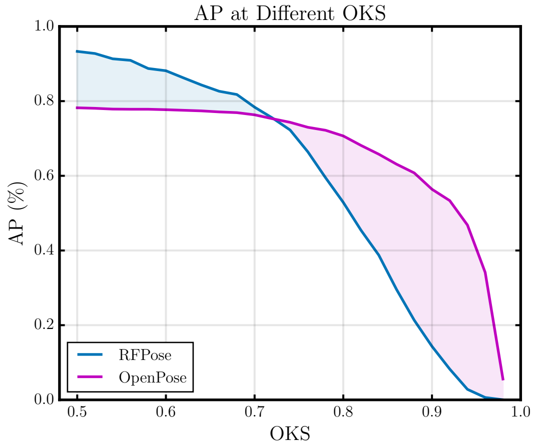


Figure 3-5: **Average precision at different OKS values.**

captured by the camera colocated with our RF sensor. For through-wall scenarios where the colocated camera cannot see people in the other room, we use the eight-camera system described in 3.3 to provide ground truth. We annotate the images captured by all eight cameras to build 3D human poses and project them on the view of the camera colocated with the radio. We annotate 1000 randomly sampled images from the visible-scene test set and another 1000 examples from the through-wall data.

■ 3.4.2 Multi-Person Pose Estimation Results

We compare human poses obtained via RF signals with the corresponding poses obtained using vision data. Table 3-1 shows the performance of RF-Pose and the baseline when tested on both visible scenes and through-wall scenarios. The table shows that, when tested on visible scenes, RF-Pose is almost as good as the vision-based OpenPose that was used to train it. Further, when tested on through-wall scenarios, RF-Pose can achieve good pose estimation while the vision-based baseline completely fail due to occlusion.

The performance of RF-Pose on through-wall scenarios can be surprising because the system did not see such examples during training. However, from the perspective of radio signals, a wall simply attenuates the signal power, but maintains the signal structure. Since our model is space invariant, it is able to identify a person behind a wall as similar to the

Methods	Hea	Nec	Sho	Elb	Wri	Hip	Kne	Ank
RF-Pose	**75.5**	**68.2**	62.2	56.1	51.9	**74.2**	63.4	54.7
OpenPose[22]	73.0	67.1	**70.8**	**64.5**	**61.5**	71.4	**68.4**	**68.3**

Table 3-2: **Average precision of different keypoints in visible scenes.**

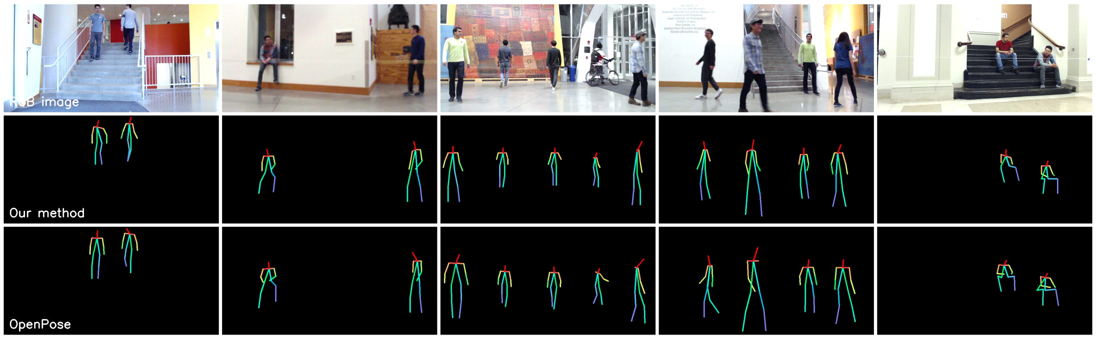
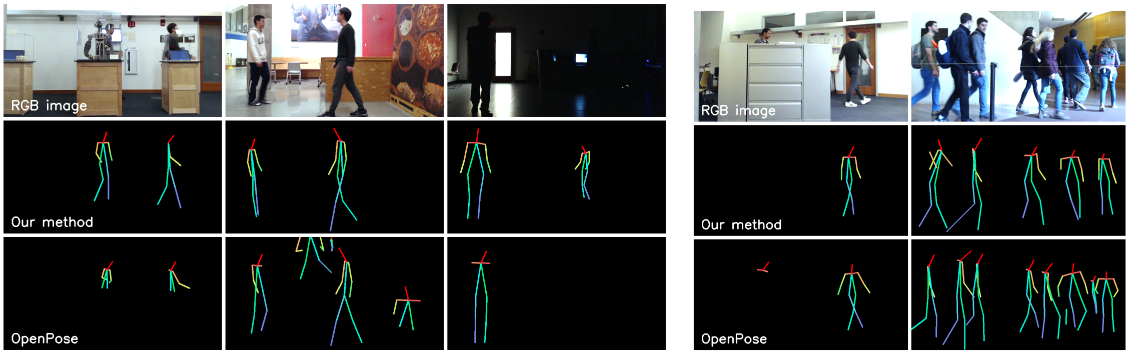


Figure 3-6: **Pose estimation with different activities and environments. First row:** Images captured by a web camera (shown as a visual reference). **Second row:** Pose estimation by our model using RF signals *only* and without any visual input. **Third row:** Pose estimation using OpenPose based on images from the first row.

examples it has seen in the space in front of a wall.

An interesting aspect in Table 3-1 is that RF-Pose outperforms OpenPose for AP^{50}, and becomes worse at AP^{75}. To further explore this aspect, we plot in Fig. 3-5 the average precision as a function of OKS values. The figure shows that at low OKS values (< 0.7), our model outperforms the vision baseline. This is because RF-Pose predicts less false alarm than the vision-based solution, which can generate fictitious skeletons if the scene has a poster of a person, or a human reflection in a glass window or mirror. In contrast, at high OKS values (> 0.75), the performance of RF-Pose degrades fast, and becomes worse than vision-based approaches. This is due to the intrinsic low spatial resolution of RF signals which prevents them from pinpointing the exact location of the keypoints. The ability of RF-Pose to exactly locate the keypoints is further hampered by imperfect synchronization between the RF heatmaps and the ground truth images.

Next, we zoom in on the various keypoints and compare their performance. Table 3-2 shows the average precision of RF-Pose and the baseline in localizing different body parts including head, right and left shoulders, elbows, wrists, hips, knees, and ankles. The results indicate that RF signals are highly accurate at localizing the head and torso (neck and hips) but less accurate in localizing limbs. This is expected because the amount of RF reflections depends on the size of the body part. Thus, RF-Pose is better at capturing the head and torso, which have large reflective areas and relatively slow motion in comparison



(a) Failure examples of OpenPose due to occlusioin, posters, and bad lighting.

(b) Failure examples of ours due to metal and crowd.

Figure 3-7: **Common failure examples. First row:** Images captured by a web camera (shown as a visual reference). **Second row:** Pose estimation by our model using RF signals *only* and without any visual input. **Third row:** Pose estimation using OpenPose based on images from the first row.

to the limbs. As for why RF-Pose outperforms OpenPose on some of the keypoints, this is due to the RF-based model operating over a clip of a few seconds, whereas the OpenPose baseline operates on individual images.

Finally, we show a few test skeletons to provide a qualitative perspective. Fig. 3-6 shows sample RF-based skeletons from our test dataset, and compares them to the corresponding RBG images and OpenPose skeletons. The figure demonstrates RF-Pose performs well in different environments with different people doing a variety of everyday activities. Fig. 3-7 illustrates the difference in errors between RF-Pose and vision-based solutions. It shows that the errors in vision-based systems are typically due to partial occlusions, bad lighting [1], or confusing a poster or wall-picture as a person. In contrast, errors in RF-Pose happen when a person is occluded by a metallic structure (e.g., a metallic cabinet in Fig. 3-7(b)) which blocks RF signals, or when people are too close and hence the low resolution RF signal fails to track all of them.

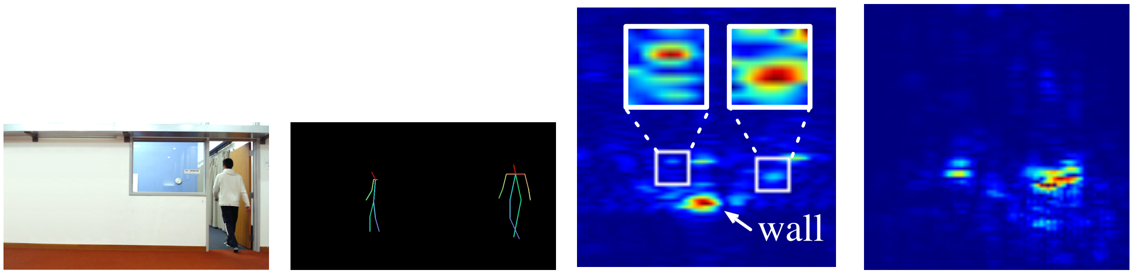
■ 3.4.3 Model Analysis

We use guided back-propagation [156] to visualize the gradient with respect to the input RF signal, and leverage the information to provide insight into our model.

Which part of the RF heatmap does RF-Pose focus on? Fig. 3-8 presents an example where one person is walking in front of the wall while another person is hidden behind

[1] Images with bad lighting are excluded during training and testing.

it. Fig. 3-8(c) shows the raw horizontal heatmap. The two large boxes are the rescaled versions of the smaller boxes and zoom in on the two people in the figure. The red patch indicated by the marker is the wall, and the other patches are multipath effects and other objects. The gradient in Fig. 3-8(d) shows that RF-Pose has learned to focus its attention on the two people in the scene and ignore the wall, other objects, and multipath.



(a) RGB image (b) Parsed poses (c) Horizontal Heatmap (d) Gradients

Figure 3-8: **Attention of the model across space.**

How does RF-Pose deal with specularity? Due to the specularity of the human body, some body parts may not reflect much RF signals towards our sensor, and hence may be de-emphasized or missing in some heatmaps, even though they are not occluded. RF-Pose deals with this issue by taking as input a sequences of RF frames (i.e., a video clip RF heatmaps). To show the benefit of processing sequences of RF frames, we sum up the input gradient in all pixels in the heatmaps to obtain activation per RF frame. We then plot in Fig. 3-9 the activation as a function of time to visualize the contribution of each frame to the estimation of various keypoints. The figure shows: that the activations of the right knee (RKnee) and right ankle (RAnkle) are highly correlated, and have peaks at time t_1 and t_2 when the person is taking a step with her right leg. In contrast, her left wrist (LWrist) gets activated after she raises her forearm at t_3, whereas her left elbow (LElbow) remains silent until t_4 when she raises her backarm.

Fig. 3-9 shows that, for a single output frame, different RF frames in the input sequence contribute differently to the output keypoints. This emphasizes the need for using a sequence of RF frames at the input. But how many frames should one use? Table 3-3 compares the model's performance for different sequence length at the input. The average precision is poor when the input uses only 6 RF frames and increases as the sequence length increases.

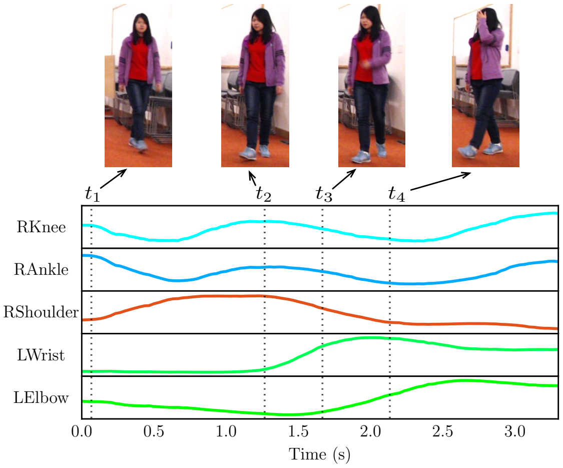


Figure 3-9: **Activation of different keypoints over time.**

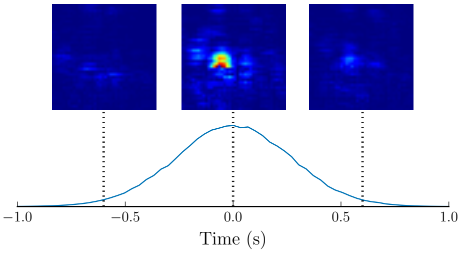


Figure 3-10: **Contribution of the neighbor to the current frame.**

But how much temporal information does RF-Pose need? Given a particular output frame, i, we compute the contributions of each of the input frames to it as a function of their time difference from i. To do so, we back-propagate the loss of a single frame w.r.t. to the RF heatmaps before it and after it, and sum up the spatial dimensions. Fig. 3-10 shows the results, suggesting that RF-Pose leverages RF heatmaps up to 1 second away to estimate the current pose.

# RF frames	AP
6	30.8
20	50.8
50	59.1
100	**62.4**

Table 3-3: **Average precision of pose estimation trained on varying lengths of input frames.**

3.4.4 Identification Using RF-Based Skeleton

We would like to show that the skeleton generated by RF-Pose captures personalized features of the individuals in the scene, and can be used by various recognition tasks. Thus, we experiment with using the RF-based skeleton for person identification.

We conduct person identification experiment with 100 people in two settings: visible environment, where the subject and RF device are in the same room, and through-wall environment, where the RF device captures the person's reflections through a wall. In each setting, every person walks naturally and randomly inside the area covered by our RF device, and we collect 8 and 2 minutes data separately for training and testing. The skeleton heatmaps are extracted by the model trained on our pose estimation dataset, which never overlaps with the identification dataset. For each setting, we train a 10-layer vanilla CNN to identify people based on 50 consecutive frames of skeleton heatmaps.

Method	Visible scenes		Through-walls	
	Top1	Top3	Top1	Top3
RF-Pose	83.4	96.1	84.4	96.3

Table 3-4: **Top1 and top3 identification percent accuracy in visible and through-wall settings.**

Table 3-4 shows that RF-based skeleton identification can reach 83.4% top1 accuracy in visible scenes. Interestingly, even when a wall blocks the device and our pose extractor never sees these people and such environments during training, the extracted skeletons can still achieve 84.4% top1 accuracy, showing its robustness and generalizability regardless of the wall. As for top3 accuracy, we achieve more than 96% in both settings, demonstrating that the extracted skeleton can preserve most of the discriminative information for identification even though the pose extractor is never trained or fine-tuned on the identification task.

3.5 Scope & Limitations

RF-Pose leverages RF signals to infer the human pose through occlusions. However, RF signals and the solution that we present herein have some limitations: First, the human body is opaque at the frequencies of interest – i.e., frequencies that traverse walls. Hence, inter-person occlusion is a limitation of the current system. Second, the operating distance of a radio is dependent on its transmission power. The radio we use in our system works up to 40 feet. Finally, we have demonstrated that our extracted pose captures identifying features of the human body. However, our identification experiments consider only one activity: walking. Exploring more sophisticated models and identifying people in the wild while performing daily activities other than walking is left for future work.

■ 3.6 Conclusion

Occlusion is a fundamental problem in human pose estimation and many other vision tasks. Instead of hallucinating missing body parts based on visible ones, we demonstrate a solution that leverages radio signals to accurately track the 2D human pose through walls and obstructions. We believe this work opens up exciting research opportunities to transfer visual knowledge about people and environments to RF signals, providing a new sensing modality that is intrinsically different from visible light and can augment vision systems with powerful capabilities.

Chapter 4
Through-Wall 3D Human Pose Estimation using Radio Signals

The past decade has witnessed much progress in using RF signals to localize people and track their motion. Novel algorithms have led to accurate localization within tens of centimeters [73, 74]. Advanced sensing technologies have enabled people tracking based on the RF signals that bounce off their bodies, even when they do not carry any wireless transmitters [8, 10, 157]. Various papers have developed classifiers that use RF reflections to detect actions like falling, walking, sitting, etc. [61, 76, 158]. This literature shows that RF signals carry an impressive amount of information about people and their movements. But, how rich a description of people can one extract from the surrounding radio signals?

In this section, we demonstrate the potential of extracting rich and detailed information about people using the radio signals that bounce off their body. Instead of simply returning a person's location, we present RF-Pose3D, a new system that can use the RF signals in the

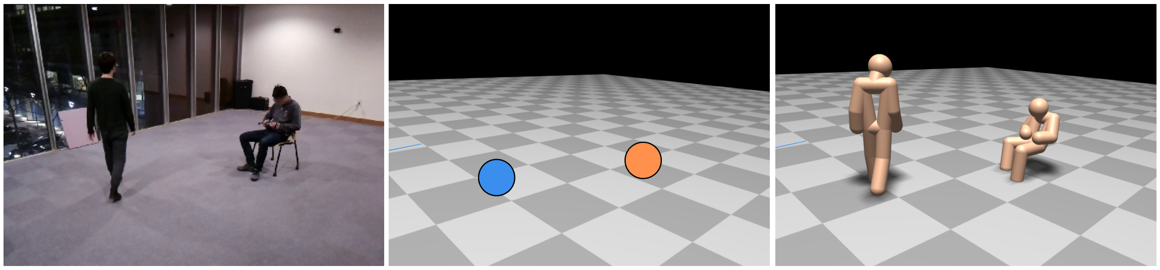


Figure 4-1: **Example output of RF-Pose3D.** Left: RGB image. Middle: RF-based localization results. Right: 3D skeletons from our system.

environment to extract full 3D skeletons of people including the head, arms, shoulders, hip, legs, etc. Further, the extracted skeletons are dynamic, i.e., they move and act like the original people in the scene. Fig. 4-1 presents the output of our system, and compares it against RF-based localization. The figure on the left shows a scene with two people. The figure in the middle illustrates the output of today's RF-based localization systems. The figure on the right shows the output of our system, which not only localizes the people, but also provides their detailed 3D skeletons and reveals their exact posture. Further, if the persons in Fig. 4-1(a) move, the skeletons in Fig. 4-1(c) would move accordingly.

Such 3D skeletons have applications in gaming where they can extend systems like Kinect to work across occlusions. They may be used by law enforcement personnel to assess a hostage scenario, leveraging the ability of RF signals to traverse walls. They also have applications in healthcare, where they can track motion disorders such as involuntary movements (i.e., dyskinesia) in Parkinson's patients.

Designing a system that maps RF signals to 3D skeletons is a highly complex task. The system must model the relationship between the observed radio waves and the human body, as well as the constraints on the location and movement of different body parts. To deal with such complexity we resort to deep neural networks. Our aim is to leverage recent success of convolutional neural network (CNN), which has demonstrated a major leap in abstracting the human pose in images and videos [22, 23, 20].

Our neural network takes as input the RF signal captured by an FMCW radio similar to the radio used in past work on localization [8]. The network operates on sliding time windows of 3 seconds. It produces a continuous 3D video of the skeletons in the scene, where for each skeleton it tracks the 3D location of 14 keypoints: head, neck, shoulders, elbows, wrists, hip, knees, and feet.

The design of RF-Pose3D is structured around three components that together provide an architecture for using deep learning for RF-sensing. Each component serves a particular function as we describe below.

(1) Sensing the 3D Skeleton: This component takes the RF signals that bounce off someone's body, and leverages deep CNN to infer the person's 3D skeleton. There is a key challenge, however, in adapting CNNs to RF data. The RF signal that we deal with is a 4 dimensional function of space and time. Thus, our CNN needs to apply 4D convolutions. But common deep learning platforms (e.g., PyTorch, TensorFlow) do not support

4D CNNs. They are targeted to images or videos, and hence support only up to 3D convolutions. More fundamentally, the computational and I/O resources required by 4D CNNs are excessive and limit scaling to complex tasks like 3D skeleton estimation.

To address this challenge, we leverage the properties of RF signals to decompose 4D convolutions into a combination of 3D convolutions performed on two planes and the time axis. We also decompose CNN training and inference to operate on those two planes. We analytically prove that our decomposition is valid and equivalent to performing 4D convolutions at each layer of the neural network. This approach not only addresses the dimensional difference between RF data and existing deep learning tools, but also reduces the complexity of the model and speed up training by orders of magnitude.

(2) Scaling to Multiple People: Most environments have multiple people. To estimate the 3D skeletons of all individuals in the scene, we need a component that separates the signals from each individual so that it can be processed independently to infer his or her skeleton. The most straightforward approach to this task would run past localization algorithms, locate each person in the scene, and zoom in on signals from that location. The drawbacks of such approach are: 1) localization errors will lead to errors in skeleton estimation, and 2) multipath effects can create fictitious people. To avoid these problems, we design this component as a deep neural network that directly learns to detect people and zoom in on them. However, instead of zooming in on people in the physical space, the network first transforms the RF signal into an abstract domain that condenses the relevant information, then separates the information pertaining to different individuals in the abstract domain. This allows the network to avoid being fooled by fictitious people that appear due to multipath, or random reflections from objects in the environment.

(3) Training: Once the network is setup, it needs training data –i.e., it needs many labeled examples where each example is a short clip (3-second) of received RF signals and a 3D video of the skeletons and their keypoints as functions of time. How do we obtain such labeled examples?

We leverage past work in computer vision which, given an image of people, identifies the pixels that correspond to their keypoints [22]. To transform such 2D skeletons to 3D skeletons, we develop a coordinated system of 12 cameras. We collect 2D skeletons from each camera, and design an optimization problem based on multi-view geometry to find the 3D location of each keypoint of each person. Of course, the cameras are used only during training to generate labeled examples. Once the network is trained, we can take the radio to new environments and use the RF signal alone to track the 3D skeletons and their movements.

RF-Pose3D has been evaluated empirically. We train and test our system using data collected in public environments around our campus.[1] The dataset has over one hundred people performing diverse indoor activities: walking, sitting, waiting for elevators, opening doors, talking to friends, etc. We train and test in different environments to ensure the network generalizes to new scenes. We summarize our results as follows:

- **Qualitative Results:** Figure 4-1 above provides a representative example of our results (more are provided in Section 4.6.3). The figure shows an important feature of our 3D skeletons. The radio in this experiment is situated behind the seated person, and hence captures signals from a specific perspective. Yet, RF-Pose3D generates 3D skeletons that can be shown from any perspective –e.g., you can look at them from the direction opposite to the radio.

- **Accuracy of Each Keypoint:** RF-Pose3D estimates simultaneously the 3D locations of 14 keypoints on the body. Its average error in localizing a keypoint is 6.5cm in the horizontal plane and 4.0cm along the vertical axis. To the best of our knowledge, this is the first work that localizes multiple keypoints on the human body at the same time.

- **Indoor Localization:** Once we have 3D skeletons, we can easily localize people. Our median localization error is 1.7cm, 2.8cm and 2.3cm along the X, Y and Z axes, which is a significant improvement over past work.

Contributions: RFPose3D makes the following contributions:

- Our system is the first to extract 3D skeletons and their keypoints from RF signals. Inferring the 3D skeleton is a complex task that requires mapping 14 keypoints on the human body to their 3D locations. It also involves generalization to unseen views that are different from the view of the radio.

- We present a novel CNN model that differs from all past work including models used in computer vision. The key property of this model is its ability to decompose 4D CNN to

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

3D convolutions over 2D planes and the time axis. This method allows us to maintain spatio-temporal relationship between human keypoints, yet operate on individual views of the signal over time, which both reduces complexity and allows for using common neural network platforms.

- We present an architecture that leverages deep learning to sense humans using RF signals. Our architecture consists of a component that generates training example, a component that separates RF data from different individuals, and a sensing component that infers properties related to a particular individual. We show how to build these components using deep neural networks and multi-view optimization of visual data. We believe that this architecture as well as our camera system can be used by researchers in the field to address other RF-based sensing tasks.

■ 4.1 Primer: Multi-Antenna FMCW Radio

RF-Pose3D uses a multi-antenna FMCW radio similar to the one used in [71]. The radio has a single transmit antenna, and two 1D antenna arrays for reception, one situated horizontally and the other vertically. The combination of FMCW and antenna arrays allows the radio to measure the signal from different 3D voxels in space. Specifically, the RF signals reflected from location (x, y, z) can be computed as [159]:

$$a(x, y, z, t) = \sum_k \sum_i s_{k,i}^t \cdot e^{j 2\pi \frac{d_k(x,y,z)}{\lambda_i}}, \tag{4.1}$$

where $s_{k,i}^t$ is the i-th sample of an FMCW sweep received on the k-th receive antenna at the time index t (i.e., the FMCW index), λ_i is the wavelength of the signal at the i-th sample in the FMCW sweep, and $d_k(x, y, z)$ is the round-trip distance from the transmit antenna to the voxel at (x, y, z), and back to the k-th receive antenna.

■ 4.2 RF-Pose3D Overview

RF-Pose3D is a system that estimates multi-people 3D poses based on RF signals. RF-Pose3D takes as input the RF reflections from the environment captured by a multi-antenna FMCW radio. Such reflections are a 4D function of space and time, which we refer to there-

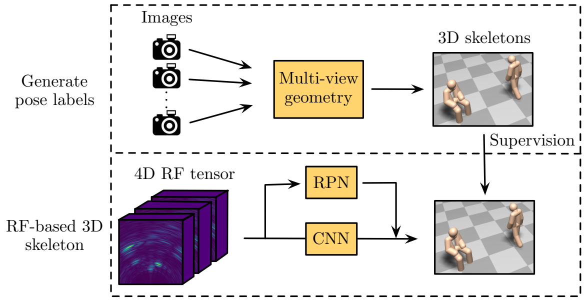


Figure 4-2: **RF-Pose3D's system overview.** Top graph shows the process of generating labeled 3D poses using our coordinated camera system. The labeled samples are used to train the model in the bottom graph. The model can be divided into two components: a region proposal network (RPN) that zooms in on RF data from one individual, and a CNN that extracts the 3D skeleton from the proposed region.

after as a 4D RF tensor.

RF-Pose3D's design is based on a deep neural network architecture (Figure 4-2). The system includes multiple components:

- A multi-camera sub-system that generates 3D poses from many 2D images taken from different viewpoints (top graph in Fig. 4-2). The output of this subsystem is used to provide labeled examples to train RF-Pose3D's neural networks.

- A neural network model that extracts multi-people 3D poses from RF signals (bottom graph in Fig. 4-2). The model is trained using labeled examples from the camera system. Once training is over, the model can infer 3D skeletons from RF signals alone. Furthermore, it can be taken to new environments that it did not see during training and would still work correctly. The model itself has two conceptual subcomponents:

  - A component that zooms in on the RF data from each individual separately. We refer to this network as the region proposal network (RPN) because it associates each person with the RF data in a particular region.

  - A component that operates on the RF data of each person and extracts his or her skeleton. We refer to this component as the CNN.

The following sections explain the above three components: the camera-system, the RPN, and the CNN. For clarity reason, we start by explaining the CNN assuming only one person in the scene. We then extend the model by adding the RPN, which takes care of separating the RF signals from different people in the scene. Finally, we explain the camera system and how it obtains labeled examples for training.

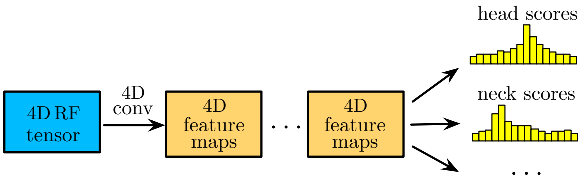


Figure 4-3: **Illustration of RF-Pose3D's 4D CNN model.** The model localizes human keypoints (e.g., head, neck, right knee) by classifying each keypoint to one voxel in space.

■ 4.3 3D Pose Estimation from RF

In this section, we describe our design of a CNN model that uses RF signal to estimate the 3D human pose. The problem of 3D pose estimation is defined as identifying the 3D locations of 14 anatomical keypoints on the body: head, neck, shoulders, elbows, wrists, hips, knees and ankles. We first focus on 3D pose estimation for a single person in this section, and extend it for multi-person scenarios in Section 4.4.

Manually designing a mapping from RF signals to 3D poses is an intractable task. Such a mapping has to take care of reflection properties, the presence of multi-path and other reflective objects, the deformable nature of the human body, and the constraints on the movements and locations of human body parts with respect to each other. Thus, rather than manually design filters or rules to decode 3D human poses from the RF signals, we consider neural networks, which have proved their advantage in learning complex mappings from training examples. Our goal is to design a CNN model that takes as input a 4D RF tensor (Section 4.1), and outputs a 3D human pose.

■ 4.3.1 CNN Model

We start by formulating keypoint localization as a CNN classification problem, then design a CNN architecture that solves the problem.

Keypoint localization as CNN classification: We first discretize the space of interests into 3D voxels. In our CNN classification problem, the set of classes are all 3D voxels, and our goal is to classify the location of each keypoint (head, neck, right elbow, etc.) into one of the voxels. Specifically, to localize a keypoint, our CNN outputs scores $\boldsymbol{s} = \{s_v\}_{v \in V}$ corresponding to all 3D voxels $v \in V$, and the target voxel v^* is the one that contains the keypoint. We use the Softmax loss $L_{\text{Softmax}}(\boldsymbol{s}, v^*)$ as the loss of keypoint localization. To localize all 14 keypoints, instead of having a separate CNN for each of the keypoint, we use a single CNN that outputs scores $\boldsymbol{s}^k$ for each of the 14 keypoints. This design forces the model to localize all the keypoints jointly, and will learn to infer the location of occluded keypoint based on the locations of other keypoints. The total loss of pose estimation is the sum of the Softmax loss of all 14 keypoints:

$$L_{\text{pose}} = \sum_k L_{\text{Softmax}}(\boldsymbol{s}^k, v^{k^*}), \tag{4.2}$$

where the index k refers to a particular keypoint. Once the model is trained, it can predict the location of each keypoint k as the voxel with the highest score:

$$\hat{v}_k = \arg\max_v \boldsymbol{s}_v^k. \tag{4.3}$$

CNN architecture: To localize keypoints in 3D space, our CNN model needs to aggregate information over space to analyze all RF reflections from a person's body and assign scores for each voxel. Also the model needs to aggregate information across time to infer keypoints that may be occluded at a specific time instance. Thus, as illustrated in Figure 4-3, our CNN model takes 4D RF tensors (space and time) as input and performs 4D convolution at each layer to aggregate information along space and time, that is:

$$\boldsymbol{a}^n = \boldsymbol{f}^n *_{(4\text{D})} \boldsymbol{a}^{n-1}, \tag{4.4}$$

where $\boldsymbol{a}^n$ and $\boldsymbol{a}^{n-1}$ are the feature maps at layer n and $n-1$, $\boldsymbol{f}^n$ is the 4D convolution filter at layer n and $*_{4\text{D}}$ is 4D convolution operator.

4.3.2 Challenge: Time and Space Complexity

The 4D CNN model described in Section 4.3.1 has practical issues. The time and space complexity of 4D CNN is so prohibitive that major machine learning platforms (PyTorch, TensorFlow) only support convolution operation up to 3D. To appreciate the computational complexity of such model, consider performing 4D convolutions on our 4D RF tensor. The size of the convolution kernel is fixed and relatively small. So the complexity stems from convolving with all 3 spatial dimensions and the time dimension. Say we want to span an area of 100 square meters with 3 meters of elevation. We want to divide this area to voxels of 1 cm^3 to have a good resolution of the location of a keypoint. Also say that we take a time window of 3 seconds and that we have 30 RF measurements per voxel per second. Performing a 4D convolution on such tensor involves $1,000 \times 1,000 \times 300 \times 90$, i.e., 27 giga operations. This process has to be repeated for each example in the training set, which contains over 1.2 million (Section 4.6.3) such examples. The training can take multiple weeks. Furthermore, the inference process cannot be performed in real-time.

In fact, the above analysis underestimates the required training and inference time since 4D convolution is one out of multiple high-complexity computations needed by a 4D CNN. Claim 0.1 below states the complexity of our 4D CNN, which depends on three equally complex computations: 4D convolution (Equation 4.4), Softmax loss computation (Equation 4.2) and maximum score selection (Equation 4.3).

Claim 0.1. *Assuming the time and space complexity of computing the response of a 4D filter at a single location and time is $O(1)$, the time and space complexity of each 4D convolution, Softmax loss computation and maximum score computation are all $O(XYRT)$, where X, Y, R, T are the size of the input 4D RF tensor along the space and time axes.*

■ 4.3.3 Model Decomposition

We present a model decomposition that allows us to reduce the complexity from $O(XYRT)$ to $O(XRT + YRT)$. For scenarios in which a resolution of a couple of centimeters is desirable for a space that spans 10×10 square meters, this decomposition translates to 3 orders of magnitude reduction in computation time. In Section 4.6.6 we show that such a reduction allows us to infer the 3D skeleton in real-time on a single GPU.

At a high-level, our model decomposition goes as follows: We first prove that our 4D RF tensor is planar decomposable (planar decomposition defined later in Definition 1 and 2). Then we prove that for a layer in a CNN, if its input is planar decomposable, its output is also planar decomposable. Thus, we can stack many convolution layers creating a deep CNN while maintaining decomposability. Finally, we prove that the computation of the loss function and the process of detecting which class has the maximum score are both decomposable when given a decomposable tensor as input. This last step means that we can train the network (and use it for inference) while operating on its decomposed version –i.e., the two 2D planar tensors and the time axis. This completes our model decomposition. Below, we define planar decomposition and state the theorem underlying the model decomposition process, and leave the proofs to the Appendix.

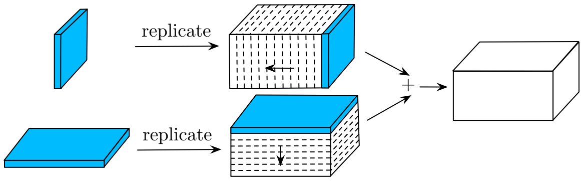


Figure 4-4: **Illustration of Planar Summation.**

We first define the concept of a planar summation. This concept allows us to create a 3D tensor from two 2D tensors simply by replicating and summing their entries, as shown in Fig. 4-4. Specifically:

Definition 1 (Planar Summation)**.** *If $\boldsymbol{A}$ is an $n \times l$ matrix and $\boldsymbol{B}$ is an $m \times l$ matrix, then the planar sum $\boldsymbol{A} \oplus \boldsymbol{B}$ is an $n \times m \times l$ 3D tensor $\boldsymbol{C}$, where $\boldsymbol{C}_{i,j,k} = \boldsymbol{A}_{i,k} + \boldsymbol{B}_{j,k}$.*

Analogously, we can define planer decomposition as taking a 3D tensor and decomposing it to two 2D tensors that can regenerate the original 3D tensor using planar summation. Specifically:

Definition 2 (Planar Decomposition)**.** *An $n \times m \times l$ 3D tensor $\boldsymbol{C}$ is planar decomposable if it can be written into the planar sum of an $n \times l$ matrix $\boldsymbol{A}$ and an $m \times l$ matrix $\boldsymbol{B}$, that is, $\boldsymbol{C} = \boldsymbol{A} \oplus \boldsymbol{B}$.*

Once we have defined planar summation and decomposition, the process of decomposing our 4D CNN becomes simple.

1. First we decompose the RF input.

   Theorem 3 (Decomposition of 4D RF tensor by decomposing its spatial dimensions)**.** *The 3D RF tensor from an FMCW array radio with a horizontal array and a vertical array (Section 4.1) is planar decomposable. It can be decomposed into the planar summation of the 2D RF tensors computed separately from the horizontal array and the vertical array.*

2. Then, we show that for every convolution layer, if its input is decomposable, its output is also decomposable.

   Theorem 4 (Decomposition of Convolution)**.** *For a decomposable 4D tensor $\boldsymbol{A} = \boldsymbol{H} \oplus \boldsymbol{V}$, the output of convolving $\boldsymbol{A}$ with a 4D filter $\boldsymbol{f}$ is also decomposable. That is, there exist 3D tensors $\boldsymbol{H'}$ and $\boldsymbol{V'}$, such that $\boldsymbol{H'} \oplus \boldsymbol{V'} = (\boldsymbol{H} \oplus \boldsymbol{V}) *_{(3D)} \boldsymbol{f}$.*

3. Next, we show that the loss function and the identification of the class with the maximum score are decomposable. Hence, allowing us to perform training and inference on the decomposed networks.

   Theorem 5 (Decomposition of Softmax Loss)**.** *For an $n \times l$ matrix $\boldsymbol{H}$ and an $m \times l$ matrix $\boldsymbol{V}$, Softmax loss $L(\boldsymbol{H} \oplus \boldsymbol{V}, (x^*, y^*, r^*))$ can be computed as:*

$$\log\big(\sum_r \big(\sum_x e^{\boldsymbol{H}_{x,r}}\big) \cdot \big(\sum_y e^{\boldsymbol{V}_{y,r}}\big)\big) - \boldsymbol{H}_{x^*,r^*} - \boldsymbol{V}_{y^*,r^*}$$

   Theorem 6 (Decomposition of Maximum Score)**.** *For an $n \times l$ matrix $\boldsymbol{H}$ and an $m \times l$ matrix $\boldsymbol{V}$, the maximum value of $\boldsymbol{H} \oplus \boldsymbol{V}$ can be computed as follows:*

$$\max\left(\boldsymbol{H} \oplus \boldsymbol{V}\right) = \max_r\left(\boldsymbol{h}_r + \boldsymbol{v}_r\right)$$

   where $\boldsymbol{h}_r = \max_x\left(\boldsymbol{H}_{x,r}\right)$ and $\boldsymbol{v}_r = \max_y\left(\boldsymbol{V}_{y,r}\right)$.

4. Finally, Claim 6.1 below states our final result of reducing the 4D CNN complexity from $O(XYRT)$ to $O(XRT + YRT)$, which is derived directly from the above theorems.

   Claim 6.1. *Assuming the time and space complexity of computing the response of a 4D filter at a single location and time are both $O(1)$, the time and space complexity of each 4D convolution (Theorem 4), Softmax loss computation (Theorem 5) and maximum value computation (Theorem 6) are all $O(XRT + YRT)$, where X, Y, R, T are the size of input 4D RF tensor on spatial and time axis.*

■ 4.4 Multi-Person 3D Pose Estimation

While the CNN described in the last section can handle single-person 3D pose estimation, the RF signal is capable of capturing multiple people at the same time. Therefore, it is desirable to extend the CNN model so that it can extract 3D skeletons of multiple people from the RF signal. To this end, we follow the divide-and-conquer paradigm by first detecting people regions and then zooming into each region to extract 3D skeleton for each individual. This leads to the design of a new neural network module called region proposal network (RPN), which generates potential people regions.

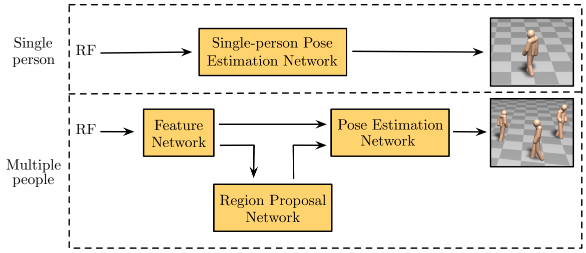


Figure 4-5: **Extension single-person model to multiple people.** The single-person pose estimation network is split into feature network and pose estimation network. Critically, region proposal network is inserted to detect individual person based on the output of FN. The skeleton of each individual is further estimated by pose estimation network.

The most straight-forward approach would have the RPN operate directly on the RF input to identify the 3D region in space that is associated with each person. We can then run our CNN pose-estimation model from last section on the 4D RF tensor after cropping it according to the RPN output. We actually take a different approach: We split the CNN model from last section and make the RPN operate on the output of an intermediate layer (i.e., a feature map), as shown in Fig 4-5. This approach is inspired by object detection in images; instead of trying to detect objects in the original image, it is preferable to detect objects at an intermediate layer after the information has been condensed. For our application, the reason why we crop the region associated with a person at an intermediate layer is twofold. First, the raw RF signal is cluttered and suffers from multipath effect. So we want to use a few convolutions layers to condense the information and remove clutter before asking the RPN to crop a specific region (Figure 4-11). Second, when multiple people are present, they may occlude each other from the RF device, resulting in missing reflections from the occluded person. Thus we want to perform a few 4D spatio-temporal convolutions to combine information across space and time to allow the RPN to detect a temporarily occluded person

The RPN is inserted in the middle as shown in Figure 4-5. The CNN model is split into two parts, which we name as feature network (FN) and pose estimation network (PEN). Feature network extracts abstract and high-level feature maps from raw RF signals. Based on these features maps, we first detect potential person regions with RPN. For each region detected by RPN, we zoom into the corresponding region on the feature maps, crop the features and feed them into our pose estimation network.

The single person network contains 18 convolutional layers totally. We split the first 12 layers into feature network (FN) and the remaining 6 layers into pose estimation network (PEN). Where to split is not unique, but generally the FN should have enough layers to aggregate spatial and temporal information for the subsequent RPN and PEN.

■ 4.4.1 Region Proposal Network

Region proposal network (RPN) is built to generate possible person regions for the subsequent pose estimation network. Ideally a region is a cuboid which tightly bounds a person. While it is inefficient to search over the 3D space to propose such cuboids, it is quite unlikely that one person stands over the head of another person. Therefore, we simplify the 3D cuboid detection as 2D bounding box detection on the horizontal plane (recall that we have decomposed our 4D convolutions to two 3D convolution over horizontal and vertical planes and the time axis).

The RPN takes as input feature maps output by the FN, and outputs a set of rectangular region proposals, each with a score describing the probability of the region containing a person. The RPN is implemented as a standard CNN. One way to train the RPN is to try to all possible regions, and for each region classify it as correct if it fits tightly around a real person in the scene. This approach is very slow since there are so many possible regions. Instead we sample potential regions using a sliding window. For each sampled window, we use a classifier to check whether it intersects reasonably well with a real person. If it does, RPN tries to adjust the boundaries of that window to make it fit better.

We assign a binary label to each window for training, to indicate whether it contains a person or not. To set the label, we use a simple intersection-over-union (IoU) metric, which is defined as:

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}} \tag{4.5}$$

Therefore, (1) a window that overlaps more than 0.7 IoU with any ground truth region (i.e., a region corresponding to a real person) is set as positive; (2) a window that overlaps less than 0.3 with all ground truth is set as negative; For other windows which satisfy neither of the above criteria, we simply ignore them during the training stage. For other details, we refer the reader to the literature on selecting regions for object detection in images [94].

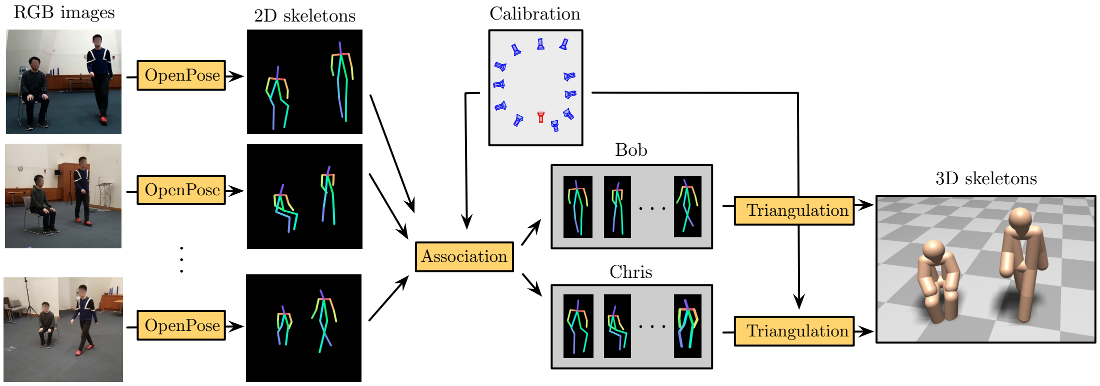


Figure 4-6: **Diagram of 3D skeleton generation using a set of RGB images.**

4.5 Generating 3D Pose Labels

To learn 3D skeletons from RF signals, RF-Pose3D needs many training examples –i.e., synchronized 4D RF tensors and the corresponding 3D skeletons. In this section, we describe a subsystem that generates such training examples. This subsystem is designed to satisfy the following requirements:

- **Portable and Passive:** It should be portable so that we can collect pose labels from different environments to make sure that our RF-based model can generalize to new scenes. It should also be passive without requiring people to wear any markers, as opposed to motion capture systems (e.g., VICON [160]) that require every person in the scene to put reflective markers around every keypoint.

- **Accurate and Robust:** It should generate accurate 3D skeletons and localize every keypoint on each person with respect to a global reference frame. It also should be robust to various types of occlusions including self-occlusion, inter-person occlusion and occlusion by furniture or walls. Such data is necessary to enable RF-Pose3D to estimate 3D skeletons from different perspectives despite occlusions.

- **Capable of dealing with multiple people:** It should track the 3D skeletons of multiple people simultaneously so that RF-Pose3D has training examples with multiple people and hence can scale to such scenarios.

We have designed and implemented a subsystem for generating labeled examples that satisfy all of the above requirements. Figure 4-6 illustrates the operation of this system, which involves the following steps:

Multi-camera system: Our system has 12 camera nodes, each of which consists of a Raspberry Pi, a battery, and a camera module board. Our nodes are small, light, and easy to deploy by attaching them on the wall. The camera nodes are synchronized via NTP and calibrated with respect to one global coordinate system using standard multi-camera calibration techniques [161]. Once deployed, the cameras image people from different view points.

2D skeleton generation: Next, our system uses the images captured by the cameras to generate 2D skeletons. To do so, we leverage a computer vision system called OpenPose [22], which given an image returns the 2D skeletons of the people in it, as shown in Figure 4-6. Ideally we would like the same skeletons to appear in the images of all 12 cameras. However, due to occlusions and the fact that 12 cameras are placed to cover different area, each camera may see different people or different keypoints of the same person.

2D skeleton association: Next, we identify 2D skeletons of the same person and associate them together as shown in Figure 4-6. To tell whether a pair of 2D skeletons are from the same person or not, we look at the geometric relationship between them. Specifically, given a 2D keypoint (e.g. head), the original 3D keypoint must lie on a line in the 3D space that is perpendicular to the camera view and intersects it at the 2D keypoint. The intuition is that when a pair of 2D skeletons are both from the same person, those two lines corresponding to the potential location of a particular keypoint will intersect in 3D space. On the other hand, if the pair of 2D skeletons are from two different people, those two lines in 3D space will have a large distance and no intersection. Based on this intuition, we use the average distance between the 3D lines corresponding to various keypoints as the distance metric of two 2D skeletons, and use hierarchical clustering [162] to cluster 2D skeletons from the same person.

Triangulating 3D skeletons: Once we have multiple 2D skeletons from the same person, we can triangulate their keypoints to generate the corresponding 3D skeleton. We estimate the 3D location of a particular keypoint $\boldsymbol{p}$ using its 2D projections $\boldsymbol{p^i}$ as the point in space whose projection minimizes the sum of distances from all such 2D projections, i.e.:

$$\boldsymbol{p} = \arg\min_{\boldsymbol{p}} \sum_{i \in I} \left\| \boldsymbol{C}_i \boldsymbol{p} - \boldsymbol{p}^i \right\|_2^2, \tag{4.6}$$

where the sum is over all cameras that detected that keypoint, and $\boldsymbol{C}_i$ is the calibration matrix that transforms the global coordinates to the image coordinates in the view of camera i [163].

4.6 Implementation and Evaluation

In this section, we describe our implementation, dataset and evaluation results.

4.6.1 Implementation

Neural Network Architecture. Today there are a few standard CNN designs that are widely used across tasks, we choose to use the ResNet [164] design that uses residual connections across different layers. For more detail about ResNet, please refer to [164]. Our feature network uses a ResNet with 12 layers. Our region proposal network and pose estimation network have another 2 and 6 layers on top of the feature network, respectively. All convolutional layers have a kernel size of 5 except the region proposal network where the kernel sizes are 3 and 1 for the first and second layer, respectively.

Training Details. All 3 subnetworks are trained jointly using ADAM optimizer [165] with a learning rate of 0.001. Both Residual Connection and Batch Normalization are adopted to benefit the training. To stabilize the training, we balance the loss weights between RPN and PEN as 1 and 0.3, respectively. We use RoiAlign [23] to crop and resize feature maps inside each region proposal.

Camera System. We have implemented a wireless camera system consisting of 12 camera nodes. Each camera node is built on a Raspberry Pi 3 single-board computer resulting in a small box design $(10 \times 7 \times 5cm)$ with a light weight $(290g)$.

RF Radio. RF-Pose3D uses an FMCW radio equipped with a vertical and horizontal antenna arrays, similar to the one used in [71]. The radio transmits an FMCW chirp sweeping the frequencies from 5.4 to 7.2 GHz. The transmission power is less than one millie Watt. The RF signal is processed using standard FMCW and antenna array equations to generate 30 vertical and horizontal heatmaps per second, which are then synchronized with the camera frames.

Synchronization. Our radio and cameras are synchronized using the network time protocol (NTP). When using a local NTP server, the clock synchronization error is less than 1ms on average. During experiments, we timestamp all the RF heatmaps and video frames and synchronize different streams based on their timestamps. We use an FPS of 30 for all the RF and video streams after synchronization.

4.6.2 Dataset

We have collected a diverse dataset of synchronized 3D skeletons and RF signals. Our dataset has people performing a variety of typical activities including walking, sitting, hand shaking, using mobile device, chatting, waving hands, etc.

- **Scale**: The dataset contains 16 hours of data. This results in 1,693,440 samples of synchronized 3D skeleton frames and 3D RF tensors.

- **Diversity**: Our data is collected from 22 different locations on a university campus including seminar rooms, open spaces, and offices. The average number of people in each frame is 2.3.

- **Accuracy of 3D skeleton labels**: To evaluate the accuracy of 3D skeletons generated by our camera system (Section 4.5), we compare the resulting skeletons against a VICON motion capturing system [160]. Table 4-1 shows the average distance between 3D skeletons from our camera system and from a VICON system. Our 3D skeletons have an average error of 1.1cm and 1.5cm along two axes on the horizontal plane and 0.7cm along the vertical axis. This result suggests that our 3D skeleton generation subsystem is very accurate and can serve as the ground-truth for training our RF-based model. Note that we could not use the VICON room to generate labeled examples for training since it would limit us to only one environment.

Axis	Avg	Hea	Nec	Sho	Elb	Wri	Hip	Kne	Ank
X	1.1	1.7	0.6	0.8	1.3	1.3	1.1	1.1	1.1
Y	0.7	0.4	0.3	0.4	0.9	1.5	0.6	0.9	0.9
Z	1.5	1.5	1.4	1.2	1.5	1.9	1.6	2.1	1.2

Table 4-1: **Average distance between labels from our camera system and labels from a VICON system.** The results show high accuracy and hence justify using our camera system as the ground truth for RF-based 3D skeleton estimation.

4.6.3 3D Pose Estimation Performance

The 3D pose estimation performance is evaluated by comparing the pose predicted from our model with the ground truth from the camera system. We ensure that the data used for testing and training do not include the same environments.

Training/Testing Split: Our dataset is split into 12 and 4 hours for training and testing, respectively. Our model is trained with data from 16 environments and tested in the remaining 6 environments that are not in the training set.

Metric: The spatial distance for each human keypoint between the model predictions and ground truth.

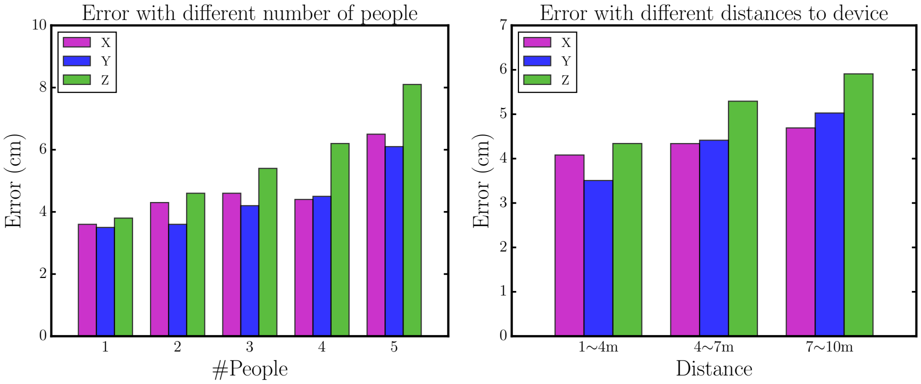
Axis	Avg	Hea	Nec	Sho	Elb	Wri	Hip	Kne	Ank
X	4.2	3.9	3.1	3.6	4.3	5.8	3.2	4.0	5.1
Y	4.0	4.4	4.2	4.3	4.0	5.1	3.5	3.3	3.5
Z	4.9	4.8	3.9	4.4	5.0	6.6	3.8	4.2	5.7

Table 4-2: **Average keypoint localization error (cm) of RF-based 3D skeleton prediction on the test set.**

Overall Performance: The keypoint localization performance of our model is shown in Table 4-2, where the X and Z axes define the horizontal plane and Y is the vertical axis. The average error in localizing a keypoint are 4.2, 4.0 and 4.9 cm in the X, Y and Z axes, respectively. The error along X and Z axes are larger than that of Y axis because of the larger variation of locations in the horizontal plane.

The table reports the localization accuracy for every keypoint type. It merges results for the left and right sides of the body. Evaluated keypoints include head, neck, shoulder, elbow, wrist, hip, knee, and ankle. The results show that our model achieves less error when localizing large or slow body parts, e.g., head or hip, than when localizing small and highly mobile parts, e.g., wrist or ankle. For example, the average error along X, Y and Z when localizing someone's head is 4.4cm, whereas the error in localizing their wrist is 5.8cm. This is expected and can be explained by two reasons. First, the amount of RF reflections highly depends on the size of a body part. Second, limbs such as wrist and ankles are more flexible and their movements usually have a larger degree of freedom than head or hip, thus are harder to be captured.

Overall the accuracy is significantly higher than past localization work, though the task is significantly harder since we are localizing small body parts. This may come as a



(a) Different number of people

(b) Different distances

Figure 4-7: **Keypoint localization performance.** Keypoint localization error (cm) with (a) different number of people; (b) people at different distances.

surprise to some readers. The reason however is threefold. First, a neural network model is much more powerful than a manually crafted model because it can capture dependencies that are unknown to the designer. Second, our model not only captures the information in the RF signal but also the general constraints on the shape and relationship between different body parts. This is because it is trained with many 3D skeletons and hence learns to abstract the relationship between their keypoints. Third, we operate over time and space. Thus, the model can learn the dynamics of how each keypoint moves and use the information to predict the location of a keypoint even when it is occluded.

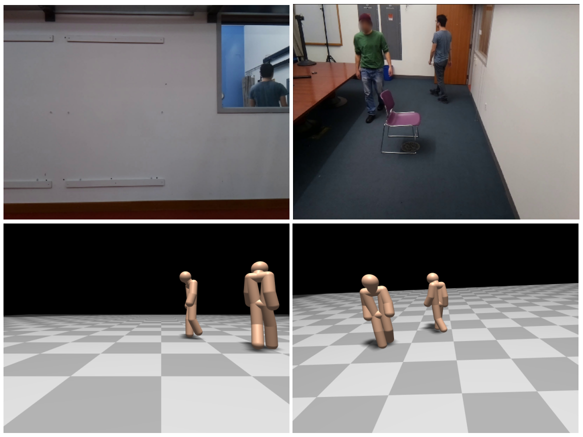


Figure 4-8: **Through-wall example.** The top left image represents the view of the radio, the top right image shows the view inside the room. Bottom row shows the detected skeletons in corresponding views.

Different Number of People: The performance on different number of people is reported in Figure 4-7(a). The average error along the spatial dimensions for a single person is $3.8cm$. As the number of subjects goes to 5, the average error increases to $7cm$, which is caused by heavy inter-people occlusion. The ability to sustain such accuracy with multiple people is due to our RPN module, which can zoom in on each person and reduce interference from other people and the environment. One major reason we do not train with more than 5 people is that the camera system starts to become unstable due to heavy occlusions. Potentially our model can be trained and tested with more people if a better camera system is constructed to provide supervision (for example by increasing the number of coordinated cameras).

Different Ranges: We evaluate the performance when people are located at different distances. Fig. 4-7(b) shows that as people move from $1m$ to $10m$, the error slightly increases from $3.8cm$ to $5.3cm$. The increase in error is expected since the spatial resolution of antenna arrays decreases with distance (an angular error of a few degrees leads to small errors at nearby distances but large errors at far distances.) We did not experiment with distances larger than 10 meters because at such distances the main limitation is the low power of the FMCW radio [8].

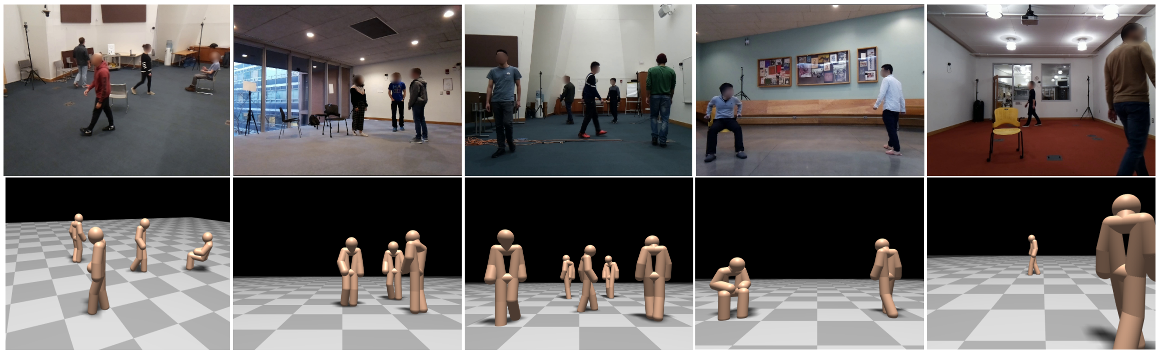


Figure 4-9: **Qualitative results on multi-person detection and pose estimation.**

Same v.s. Different Environment: All of the above results were for training and testing on different environment. In this section, we train and test our model in the same environment in order to compare with cross environment testing result. Note that though we use the same environment, we still use different examples for training and testing. The average error along X, Y, and Z is $3.7cm$ which is on par with cross environment error which is $4.4cm$. This clearly shows that our model is robust to environmental changes. Again, this benefit stems from the RPN module which enables the PEN to focus on individual people

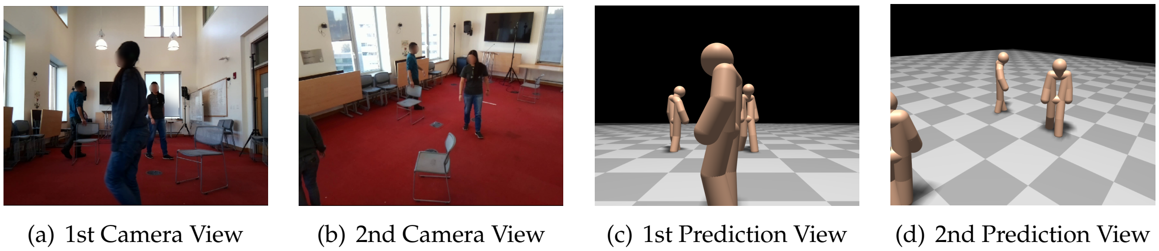

(a) 1st Camera View (b) 2nd Camera View (c) 1st Prediction View (d) 2nd Prediction View

Figure 4-10: **RF-Pose3D generates 3D skeletons from different perspectives.** (a)(b) shows two views out of the camera system, (c)(d) shows the detected skeletons in corresponding views.

and ignore environmental reflectors.

Through-Wall v.s. Line of Sight: We evaluate our system in through-wall scenarios where the radio is separated from the monitored people by a wall. The errors along the X, Y and Z axes are 5.2cm, 3.7cm and 4.7cm, respectively. These errors are comparable to the errors in line-of-sight scenarios which are reported in Table 4-2. One example is shown in Figure 4-8, where the top left image shows the viewpoint of the radio, the right image shows the view inside the room. The second row shows the 3D skeletons from the corresponding viewpoints.

Qualitative Results: Fig. 4-9 and Fig. 4-10 shows samples of 3D skeletons for multiple people generated using RF-Pose3D. It illustrates that RF-Pose3D works well in different environments and when people are doing a variety of activities, e.g., sitting, walking, interacting with each other, etc.

4.6.4 Performance of Human Detection

Recall that our model starts by detecting people and zooming on each of them to extract his or her skeleton. Thus we evaluate the human detection performance of our model–i.e., whether it correctly detects all the people in the environment despite fictitious people due to multipath or other objects.

Metrics: We use the following metrics that are commonly used in object detection tasks.

- *Precision:* Precision is defined as the fraction of detected regions that truly contain a person. It measures the robustness of our system against false positives, i.e., fictitious people.
- *Recall:* Recall is defined as the fraction of people that are detected over the total amount

of people. It measures our system's ability in detecting all the people without misses.

- *F1 score:* F1 considers both precision and recall, and is computed as the harmonic average of the two, i.e., $\frac{2 \cdot p \cdot r}{p+r}$.

Table 4-3 shows the precision and recall for test data with different number of people in the scene. Overall, our model achieves a precision of 95.8% and a recall of 99.6% on single-person data. As the number of people increases, the F1 score only drops slightly by 2.9%. This demonstrates the effectiveness of our region proposal network, which successfully detects multiple people in the environment without being fooled by multipath or objects in the environment. This is partly attributed to the feature network which learns to attenuate the side effect of multipath as well as aggregate beneficial temporal information.

#People	1	2	3	4	5
Precision (%)	95.8	96.2	94.9	95.2	96.3
Recall (%)	99.6	98.8	97.8	96.5	93.4
F-1 score	97.7	97.4	96.3	95.9	94.8

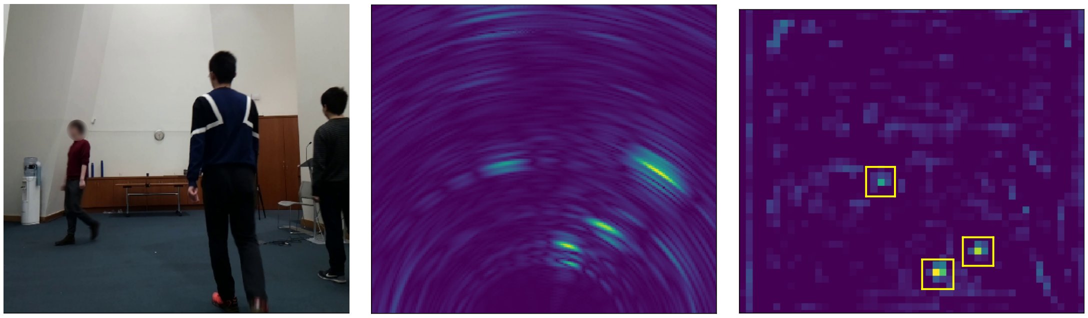
Table 4-3: **Precision and Recall when there are different number of people.**

To better understand how RPN works consider the example in Figure 4-11. The left part of Figure 4-11 shows an experiment where there are three people in the scene. The middle part of the figure shows the horizontal RF tensor at that instance of time, which contains multipath reflections from the wall. The right part of the figure shows one of the feature maps from the feature network together with the regions proposed by RPN. The feature map has a large value only at the locations of the three people, suggesting that the feature network has learned to differentiate reflections from real people from fictitious ones due to multipath and objects in the environment. In this example, the RPN successfully detects all the people and have 3 proposal boxes corresponding to each of them.

4.6.5 Localization Performance

Methods	Median			90-th Percentile		
	X	Y	Z	X	Y	Z
RF-Pose3D	1.7	2.8	2.3	5.1	8.3	6.4
WiTrack [8]	9.9	8.6	17.7	35.0	20.0	60.0

Table 4-4: **Comparison with previous device-free localization system.** Median and 90-th percentile localization error (cm) of RF-Pose3D and WiTrack.



(a) RGB image (b) Horizontal RF tensor (c) Feature map

Figure 4-11: **An example output of RPN.** Left: RGB images from the view of the device. Middle: horizontal RF tensor that contains fictitious people along with real ones. Right: Decomposed horizontal feature map after FN, marked with detected regions. Fictitious people are removed, real ones are detected.

We also compare our model with past work on indoor localization. Our trained model can derive people's location simply by computing the center of neck, two shoulders and two hips. We compare our method with previous RF-based device-free indoor localization system WiTrack [8] in Table 4-4. Our system achieves a median error of 1.7, 2.8 and 2.3 on X, Y and Z axes, respectively and 90-th percentile error of 5.1, 8.3 and 6.4, which is significantly better than past localization systems. This results demonstrates the power of the new model and the importance of the extra information it can get from the wireless signal even for more traditional tasks like localization.

■ 4.6.6 Running Time Analysis

As explained in Section 4.3.3, the proposed planar tensor decomposition technique enables us to train and test on 4D tensor data using 3D convolutions. Here, we provide a quantitative analysis of it. In Table 4-5, we benchmark the inference runtime of the three subnetworks of our model: FN, FPN and PEN. On a single NVIDIA Titan X GPU, one second of RF tensor data takes only 0.39 seconds to process. Estimated from the number of floating point operations, it would take a 4D CNN approximately 87 seconds to perform inference, which is way below real-time.

Model	FN	RPN	PEN	Total	4D CNN
Time (s)	0.04	0.01	0.34	0.39	87.0 (estimated)

Table 4-5: **Runtime analysis of our model during inference on a single NVIDIA Titan X GPU.** The table shows the time spent on each part of our model for every 1 second of RF signal. It suggests that our model can perform inference in real-time with our decomposition techniques while a vanilla 4D CNN could take 87.0 seconds from estimation.

4.7 Discussion

We present RF-Pose3D, a device-free system that for the first time estimates 3D human skeletons from RF signals. By designing a novel CNN model and leveraging camera system for supervision, RF-Pose3D is able to detect 3D skeletons for multiple people simultaneously. In terms of modeling, to avoid high dimensional convolution operations, we propose a tensor decomposition technique that is computationally efficient, making the system capable of running in realtime.

RF-Pose3D provides a leap in the quality and richness of human-related information learned from RF signals. However, the system exhibits some limitations: First, our dataset is focused on common activities in office buildings (e.g., walking, sitting, standing) and misses certain poses, e.g., dancing and doing sports. As a result, the trained model is good for poses common in office buildings and may degenerate with poses it did not see in the dataset. This problem can be addressed by expanding the dataset to include more actions. Second, the radio we use in our system can work up to 40 feet. Extra transmission power or multiple radios would be needed in order to cover a larger space. Third, the efficacy of RF-based pose estimation depends on the power reflected from each body part. Naturally, smaller body parts (e.g., hands and wrists) reflect less power than larger ones. Thus, learning actions that involve complex hand motion is more difficult. Despite these limitations, we see this work as an important step towards using wireless signals for human sensing. We believe this non-contact 3D pose tracking system can enable new applications in healthcare, smart homes and video gaming.

4.8 Proof of Theorems

Proof of Theorem 3. We prove that each 3D RF tensor is planar decomposable, and therefore the 4D RF tensor (3D RF tensor over time) is also planar decomposable. Consider an

FMCW array with M and N receivers for the horizontal and vertical arrays, respectively. Let (x, y, r) denotes 3D location in the (X, Y, R)-coordinate system as shown in Figure 4-12, where r is the distance from the point (x, y, r) to the origin. Let $d^h_m(x, y, r)$ denotes the round trip distance from transmit antenna to the point at the 3D voxel at (x, y, r) and back to the m-th horizontal receive antenna. $d^v_n(x, y, r)$ is similarly defined for the n-th vertical receive antenna.

Base on Equation 4.1, the 3D RF tensor is computed as:

$$A(x, y, r) = \sum_{m=1}^{M} \sum_{i} s^h_{m,i} \cdot e^{j2\pi \frac{d^h_m(x,y,r)}{\lambda_i}} + \sum_{n=1}^{N} \sum_{i} s^v_{n,i} \cdot e^{j2\pi \frac{d^v_n(x,y,r)}{\lambda_i}}$$

and the 2D RF tensor based on horizontal and vertical array are computed as:

$$H(x, r) = \sum_{m=1}^{M} \sum_{i} s^h_{m,i} \cdot e^{j2\pi \frac{d^h_m(x,0,r)}{\lambda_i}}$$

$$V(y, r) = \sum_{n=1}^{N} \sum_{i} s^v_{n,i} \cdot e^{j2\pi \frac{d^v_n(0,y,r)}{\lambda_i}}$$

It can be proved geometrically that $d^h_m(x, y, r) = d^h_m(x, 0, r)$ and $d^v_n(x, y, r) = d^v_n(0, y, r)$, therefore $A(x, y, r) = H(x, r) + V(y, r)$, that is $A = H \oplus V$. □

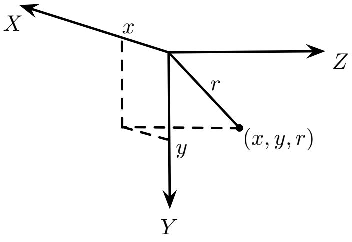


Figure 4-12: **(X,Y,R)-coordinate system.**

Proof of Theorem 4. Due to space limit, we only prove the decomposition of 3D spatial convolution below, and 4D convolution is a natural extension of it. For an $n \times l$ matrix $\boldsymbol{H}$ and an $m \times l$ matrix $\boldsymbol{V}$, we prove that: $(\boldsymbol{H} \oplus \boldsymbol{V}) *_{(\text{3D})} (\boldsymbol{f}^h \oplus \boldsymbol{f}^v) = \boldsymbol{H'} \oplus \boldsymbol{V'}$, where $\boldsymbol{H'} = (m\boldsymbol{H} + \mathbf{1}_{n\times m} \cdot \boldsymbol{V}) *_{(\text{2D})} \boldsymbol{f}^h$, $\boldsymbol{V'} = (n\boldsymbol{V} + \mathbf{1}_{m\times n} \cdot \boldsymbol{H}) *_{(\text{2D})} \boldsymbol{f}^v$, and $\mathbf{1}_{a\times b}$ is a-by-b all-one matrix. Let $A = (\boldsymbol{H} \oplus \boldsymbol{V}) *_{\text{(3D)}} (\boldsymbol{f}^h \oplus \boldsymbol{f}^v)$.

$$\boldsymbol{A}(x,y,r) = \sum_{i,j,k} \big(\boldsymbol{H}(x{+}i, r{+}k) + \boldsymbol{V}(y{+}j, r{+}k)\big) \cdot \big(\boldsymbol{f}^h(i,k) + \boldsymbol{f}^v(j,k)\big)$$

$$\boldsymbol{H'}(x,r) = m \sum_{i,k} \boldsymbol{H}(x{+}i, r{+}k)\boldsymbol{f}^h(i,k) + \sum_{i,k}\sum_{j} \boldsymbol{V}(j, r{+}k)\boldsymbol{f}^h(i,k)$$

$$\boldsymbol{V'}(y,r) = n \sum_{j,k} \boldsymbol{V}(y{+}j, r{+}k)\boldsymbol{f}^v(j,k) + \sum_{j,k}\sum_{i} \boldsymbol{H}(i, r{+}k)\boldsymbol{f}^v(j,k)$$

It can be examined that $\boldsymbol{A}(x,y,r) = \boldsymbol{H'}(x,r) + \boldsymbol{V'}(x,r)$, hence $\boldsymbol{A} = \boldsymbol{H'} \oplus \boldsymbol{V'}$ by definition.

□

Proof of Theorem 5.

$$L(\boldsymbol{H} \oplus \boldsymbol{V}, (x^*, y^*, r^*)) = \log\,(\sum_{x,y,r} e^{\boldsymbol{H}_{x,r}+\boldsymbol{V}_{y,r}}) - \boldsymbol{H}_{x^*,r^*} - \boldsymbol{V}_{y^*,r^*}$$

$$= \log\,(\sum_{r} (\sum_{x} e^{\boldsymbol{H}_{x,r}}) \cdot (\sum_{y} e^{\boldsymbol{V}_{y,r}})) - \boldsymbol{H}_{x^*,r^*} - \boldsymbol{V}_{y^*,r^*}$$

Proof of Theorem 6.

$$\begin{aligned}
\max\,(\boldsymbol{H} \oplus \boldsymbol{V}) &= \max_{x,y,r}\,(\boldsymbol{H}_{x,r} {+} \boldsymbol{V}_{y,r}) \\
&= \max_{r}\,(\max_{x} \boldsymbol{H}_{x,r} {+} \max_{y} \boldsymbol{V}_{y,r}) \\
&= \max_{r}\,(\boldsymbol{h}_r + \boldsymbol{v}_r)
\end{aligned}$$

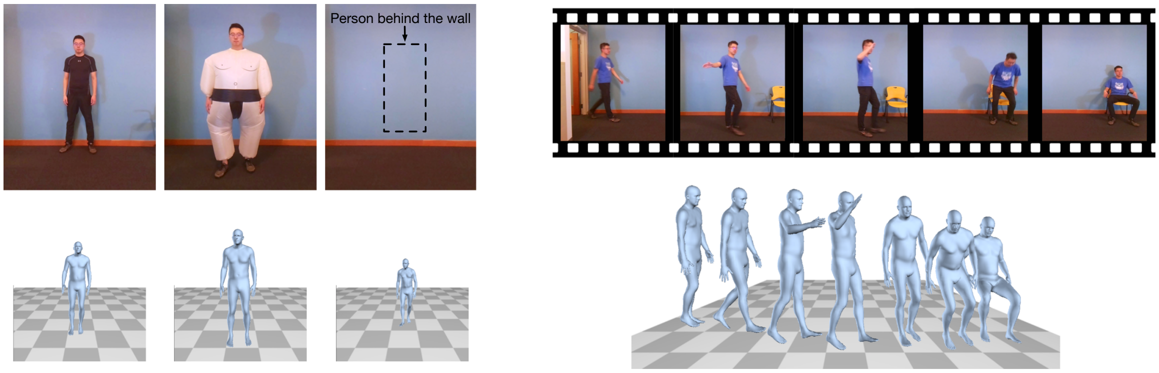
CHAPTER 5

Through-Wall Human Mesh Recovery using Radio Signals

Estimating a full 3D mesh of the human body, capturing both human pose and body shape, is a challenging task in computer vision. The community has achieved major advances in estimating 2D/3D human pose [86, 166], and more recent work has succeeded in recovering a full 3D mesh of the human body characterizing both pose and shape [167, 104]. However, as in any camera-based recognition task, human mesh recovery is still prone to errors when people wear baggy clothes, and in the presence of occlusions or under bad lighting conditions.

Recent research has proposed to use different sensing modalities that could augment vision systems and allow them to expand beyond the capabilities of cameras [168, 169, 170, 6, 5]. In particular, radio frequency (RF) based sensing systems have demonstrated through-wall human detection and pose estimation [1, 2]. These methods leverage the fact that RF signals in the WiFi range can traverse occlusions and reflect off the human body. The resulting systems are privacy-preserving as they do not record visual data, and can cover a large space with a single device, despite occlusions. However, RF signals have much lower spatial resolution than visual camera images, and therefore it remains an open question as to whether it is possible at all to capture dynamic 3D body meshes characterizing the human body and its motion with RF sensing.

In this section, we demonstrate how to use RF sensing to estimate dynamic 3D meshes



(a) Human mesh recovery through occlusions

(b) Dynamic human mesh captures the body motion

Figure 5-1: **Dynamic human meshes estimated using radio signals.** Images captured by a camera co-located with the radio sensor are presented here for visual reference. (a) shows the estimated human meshes of the same person in sportswear, a baggy costume and when he is behind the wall. (b) shows the dynamic meshes that capture the motion when the person walks, waves his hand, and sits.

for human bodies through walls and occlusions. We introduce RF-Avatar, a neural network framework that parses RF signals to infer dynamic 3D meshes. Our model can capture body meshes in the presence of significant, and even total, occlusion. It stays accurate in bad lighting conditions, and when people wear costumes or baggy clothes. Figure 5-1 shows RF-Avatar's performance on a few test examples. The left panel demonstrates that RF-Avatar can capture the 3D body mesh accurately even when the human body is obscured by a voluminous costume, or completely hidden behind a wall. Further, as shown in the right panel, RF-Avatar generates dynamic meshes that track the body movement. In Section 5.3.2, we show that RF-Avatar also works in dark settings and in scenarios with multiple individuals.

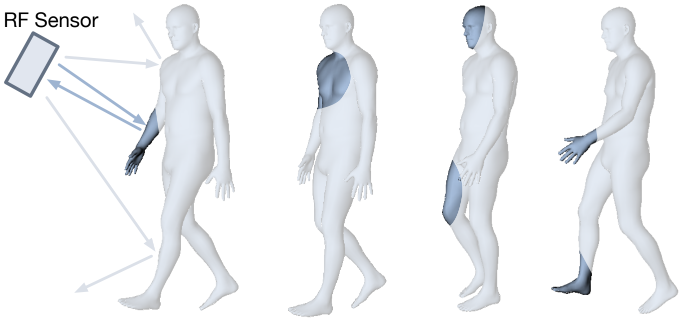


Figure 5-2: **Specularity of the human body with respect to RF.** The human body reflects RF signals as opposed to scattering them. A single RF snapshot can only capture a subset of limbs depending on the orientation of the surfaces.

Inferring 3D body meshes solely from radio signals is a difficult task. The human body is specular with respect to RF signals in the WiFi range –i.e., the human body reflects RF signals, as opposed to scattering them. As illustrated in Figure 5-2, depending on the orientation of the surface of each limb, the RF signal may be reflected towards our radio or away from it. Thus, in contrast to camera systems where any snapshot shows all unoccluded body parts, in radio systems, a single snapshot has information only about a subset of the limbs. This problem is further complicated by the fact that there is no direct relationship between the reflected RF signals from a person and their underlying 3D body mesh. We do not know which part of the body actually reflected the signal back. This is different from camera images, which capture a 2D projection of the 3D body meshes (modulo clothing). The fact that the reflected RF signal at a point in time has information only about a unknown subset of the body parts means that using RF sensing to capture 3D meshes is a highly unconstrained problem – at a point in time, the reflected RF signal could be explained by many different 3D meshes, most of which are incorrect.

RF-Avatar tackles the above challenge as follows. We first develop a module that uses the RF signal to detect and track multiple people over time in 3D space, and create trajectories for each unique individual. Our detection pipeline extends the Mask-RCNN framework [23] to handle RF signals. RF-Avatar then uses each person's detected trajectory, which incorporates multiple RF snapshots over time, to estimate their body mesh. This strategy of combining information across successive snapshots of RF signals allows RF-Avatar to deal with the fact that different RF snapshots contain information about different body parts due to the specularity of the human body. We incorporate a multi-headed attention module that lets the neural network selectively focus on different RF snapshots at different times, depending on what body parts reflected RF signals back to the radio. RF-Avatar also learns a prior on human motion dynamics to help resolve ambiguity about human motion over time. We introduce a temporal adversarial training method to encode human pose and motion dynamics.

To train our RF-based model, we use vision to provide cross-modality supervision. We use various types of supervision, ranging from off-the-shelf 2D pose estimators (for pose supervision) to vision-based 3D body scanning (for shape supervision). We design a data collection protocol that scales to multiple environments, while also minimizing overhead and inconvenience to subjects.

We train and test RF-Avatar using data collected in public environments around our campus. Our experimental results show that in visible scenes, RF-Avatar has mean joint position error of 5.84 cm and mean vertex-to-vertex distance of 1.89 cm. For through-wall scenes and subjects wearing loose costumes, RF-Avatar has mean joint position error of 6.26 cm and mean vertex-to-vertex distance of 1.97 cm whereas the vision-based system fails completely. We conduct ablation studies to show the importance of our self-attention mechanism and the adversarially learned prior for human pose and motion dynamics.

■ 5.1 RF Signals and CNN

Much of the work on sensing people using radio signals uses a technology called FMCW (Frequency Modulated Continuous Wave) [171, 172]. An FMCW radio works by transmitting a low power radio signal and receiving its reflections from the environment. Different FMCW radios are available [147, 148] and RF-Avatar uses one similar to that used in [71] and can be ordered from [173]. Our model is not specific to a particular radio, and applies generally to such RADAR-based radios. In RF-Avatar , the reflected RF signal is transformed into a function of the 3D spatial location and time [2]. This results in a 4D tensor that forms the input to our neural network. It can be viewed as a sequence of 3D tensors at different points of time. Each 3D tensor is henceforth referred to as the *RF frame* at a specific time.

It is important to note that RF signals have intrinsically different properties from visual data, i.e., camera pixels: first, the human body is specular in the frequency range that traverse walls (see Figure 5-2). Each RF frame therefore only captures a subset of the human body parts. Also, in the frequency range of interest (in which RF can pass through walls), RF signals have low spatial resolution – our radio has a depth resolution about 10 cm, and angular resolution of 15 degrees. This is a much lower resolution than what is obtained with a camera. The above properties have implications for human mesh recovery, and need to be taken into account in designing our model.

CNN with RF Signals: Processing the 4D RF tensor with 4D convolutions has prohibitive computational and space complexity. We use a decomposition technique [2] to decompose both the RF tensor and the 4D convolution into 3D ones. The main idea is to represent each 3D RF frame as a summation of multiple 2D projections. As a result, the operation in the original dimension is equivalent to a combination of operations in lower-dimensions.

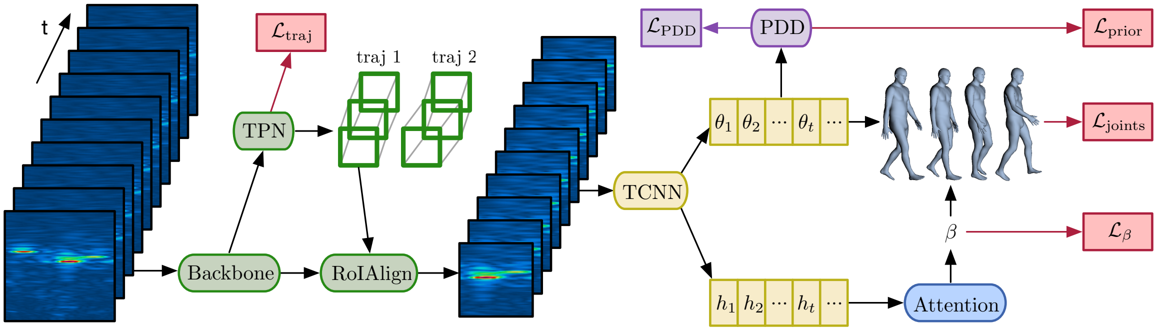

t
$\mathcal{L}_{\text{traj}}$
traj 1
traj 2
TPN
Backbone
RoIAlign
TCNN
$\mathcal{L}_{\text{PDD}}$
PDD
$\mathcal{L}_{\text{prior}}$
θ_1 θ_2 ⋯ θ_t ⋯
$\mathcal{L}_{\text{joints}}$
β
$\mathcal{L}_{\beta}$
h_1 h_2 ⋯ h_t ⋯
Attention

Figure 5-3: **Overview of the network model used in RF-Avatar.**

5.2 Method

We propose a neural network framework that parses RF signals and produces dynamic body meshes for multiple people. The design of our model is inspired by the Mask-RCNN framework [23]. Mask-RCNN is designed for instance-level recognition tasks in 2D images; we extend it to handle 4D RF inputs and generate 3D body meshes over time. Figure 5-3 illustrates the 2-stage network architecture used in RF-Avatar. In the first stage of the model, we use a Trajectory Proposal Network (TPN) to detect and track each person in 3D space (Sec. 5.2.2). TPN outputs a trajectory (a sequence of bounding boxes over time) for each person, and we use this trajectory to crop the spatial regions in the RF tensor that contain this particular person.

The second stage of the model takes the cropped features as input and uses a Trajectory-CNN (TCNN) to estimate the sequence of body meshes of this person (Sec. 5.2.3). TCNN introduces an attention module to adaptively combine features from different RF frames when predicting the body shape (Sec. 5.2.3). TCNN also outputs a sequence of joint angles capturing the body motion. It uses a Pose and Dynamics Discriminator (PDD) to help resolve the ambiguities about human motion (Sec. 5.2.4). We describe how we use various forms of supervision to train RF-Avatar in Sec. 5.2.5.

5.2.1 Human Mesh Representation

We use the Skinned Multi-Person Linear (SMPL) model [174] to encode the 3D mesh of a human body. SMPL factors the human mesh into a person-dependent shape vector and pose-dependent 3D joint angles. The shape vector $\boldsymbol{\beta} \in \mathbb{R}^{10}$ corresponds to the first 10 coefficients of a PCA shape model. The joint angles $\boldsymbol{\theta} \in \mathbb{R}^{72}$ define the global rotation of the body and the 3D relative rotations of 23 joints. SMPL provides a differentiable function $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$ that outputs $N = 6890$ vertices of a triangular mesh given $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. A 3D mesh of a human body in the world coordinates is represented by 85 parameters including $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ (describing shape and pose via SMPL) and a global translation vector $\boldsymbol{\delta}$. Note that the 3D location of body joints, $\boldsymbol{J}$, can be computed via a linear combination of mesh vertices.

RF-Avatar recovers dynamic body meshes, i.e., a sequence of SMPL parameters including a time-invariant $\boldsymbol{\beta}$ characterizing the body, and a time-variant $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_T)$ describing the joint angles, and a time-variant global translation vector $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_T)$ capturing the location.

■ 5.2.2 Trajectory Proposal Network

The first stage in our 3D mesh estimation pipeline is to detect regions containing individuals and then track them over time to form trajectories. Our Trajectory Proposal Network (TPN) takes as input the 4D RF tensor. It first extracts features using a backbone with spatial-temporal convolutions, and then uses a recurrent region proposal network to propose candidate regions for each RF frame. After a further candidate selection stage with a box head, we perform a lightweight optimization to link the detections over time. We describe each TPN component in detail:

Backbone: This takes the raw sequence of RF frames as input and uses a set of decomposed 4D convolutional layers (see Sec. 5.1) with residual connections to produce features.

Recurrent Region Proposal Network (Recurrent-RPN): In contrast to prior work using RPN in detection and tracking [94, 23, 175], our recurrent-RPN has two major differences. First, we wish to detect individuals in the 3D world space instead of the 2D image space. Thus, our model uses 3D bounding boxes as anchors and learns to propose 3D regions by transforming these anchors. Proposing regions in 3D space removes scale-variation of regions due to perspective projection to image space [176]. For tractability, we choose 3D anchors to be those close to the ground plane. Second, our RPN works in a recurrent manner to propose regions for each RF frame sequentially. It uses recurrent layers on top of convolutional layers to predict object scores and regression outputs for all anchor regions. Non-maximal suppression (NMS) is used to remove duplicated proposals.

Box Head: To improve detection precision, we use a box head to further classify proposals into correct/incorrect detections. We use standard box head with RoIAlign [23].

Tracker: The tracker module receives proposals from the Box Head output at each timestep. It then associates together proposals that belong to the person, and stitches them over time to form trajectory tubes. We use a lightweight optimization tracker based on bipartite matching [175].

5.2.3 Trajectory-CNN with Attention

Trajectory-CNN (TCNN) uses the cropped features from the TPN as input and estimates the body mesh parameters for each individual. To deal with the fact that different RF frames contain information about different body parts, we introduce a self-attention module to predict a temporally consistent shape $\boldsymbol{\beta}$. TCNN first extracts shape features at different timesteps as $\boldsymbol{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_T)$. Our self-attention module uses a function f to attend to different frames and combine all the shape features into a fixed-length feature vector: $\widetilde{\boldsymbol{h}} = \frac{1}{\mathcal{C}(\boldsymbol{H})} \sum_t (f(\boldsymbol{h}_t) \cdot \boldsymbol{h}_t)$, where $\mathcal{C}(\boldsymbol{H}) = \sum_t (f(\boldsymbol{h}_t)$ is a normalization factor. We utilize multi-headed self-attention [177], allowing the neural network to attend to different aspects of the shape features differently. Feature vectors from different heads are concatenated together to produce the $\boldsymbol{\beta}$ prediction.

Empirical results show that this temporal self-attention leads to improved shape estimation and model interpretability. We further believe that the benefits of temporal attention extend to video-based 3D mesh models, since it allows the model to recognize that different frames may have different importance for estimating a particular mesh parameter. For example, height is better estimated from frames where the subject is standing as opposed to sitting.

5.2.4 Learning Pose and Dynamics Priors

We would like to learn a prior that encodes feasible human pose and motion dynamics in order to ensure that the 3D meshes it produces over time are realistic. Without such a prior, and especially given the weak supervision for the 3D joint angles (see Sec. 5.2.5), our model could produce arbitrary rotations of joints and/or temporally inconsistent meshes. This issue is exacerbated in the case of pose estimation from RF signals, as we only get sparse observations at each timestep, due to human body specularity.

We introduce an adversarial prior that regularizes both human body pose and motion dynamics and ensures realistic predictions; we call this the Pose and Dynamics Discriminator (PDD). PDD is a data-driven discriminator that takes our predicted sequence of 3D joint angles, and aims to distinguish it from real human poses and dynamics data. We use MoSh-ed data from the CMU MoCap dataset [178] as real dynamics data. It covers a diverse set of human subjects performing different poses and actions. In contrast to previous work, which uses a separate discriminator for each joint at a single time instance [104, 106], PDD considers all keypoints over a temporal window, which improves the estimated pose results.

The PDD is trained using a binary cross entropy loss and a gradient penalty term on the real data. Its objective function takes the following form:

$$\begin{aligned} \mathcal{L}_{\text{PDD}} = &-\Big(\mathbb{E}_{\mathbf{\Phi}\sim p_{\text{data}}}[\log D(\mathbf{\Phi})] + \mathbb{E}_{\mathbf{\Theta}\sim p_E}\big[\log(1 - D(\mathbf{\Theta}))\big]\Big) \\ &+\gamma \cdot \mathbb{E}_{\mathbf{\Phi}\sim p_{\text{data}}}[\|\nabla D(\mathbf{\Phi})\|^2], \end{aligned} \tag{5.1}$$

where $\mathbf{\Theta}$ is the estimated joint angles from TCNN, and $D(\cdot)$ is our pose and dynamics discriminator.

Finally, we convert them to rotation matrices and feed to the discriminator. This technique allows for more stable training by bypassing the 2π wrapping nature of angle representations.

5.2.5 Training the Model

Past image-based solutions that recover 3D meshes use mostly weak supervision during training, in the form of the location of body joints. However, our empirical results (Sec. 5.3.3) show that weak supervision is insufficient for RF-based systems. Unfortunately, strong supervision that captures full information about 3D meshes is difficult to obtain, as it requires highly constrained setups involving a sophisticated multi-view camera setup, and minimally clothed subjects [179, 180]; such setups are not scalable.

To deal with this issue, we train our model using a combination of strong and weak supervision. The SMPL shape representation decomposes into a time-independent shape vector, $\boldsymbol{\beta}$, and time-dependent joint angles, $\boldsymbol{\theta}$. We obtain strong supervision for the time-independent shape vector by using an adapted version of the scanning/silhouette method from [101] once for each subject in our dataset, with each subject in a standard canonical pose. We need only perform this procedure once for each person, as the shape vector, $\boldsymbol{\beta}$, is constant for a given person. We adapt the procedure in [101] as follows. The original method solves an optimization problem to obtain both $\boldsymbol{\beta}$ and offsets for the N mesh vertices (to capture clothing and other small perturbations). We remove the optimization over the mesh vertices (as we wish to capture pure body shape, and do not wish to include clothing information) to obtain only $\boldsymbol{\beta}$. We henceforth refer to the mesh obtained from this method as a *VideoAvatar*.

Additionally, we use a system of 12 calibrated cameras and the AlphaPose algorithm [86, 166] to obtain ground truth information for 3D joint locations, obtained as subjects engage in activities (walking, standing up/sitting down, interacting with objects, etc). This serves as weak supervision for our system's joint angle predictions, $\boldsymbol{\theta}$.

Training TPN: We use standard anchor classification and regression losses [94, 23]. We compute ground truth 3D bounding boxes from the 3D poses reconstructed by 3D AlphaPose. The total loss $\mathcal{L}_{\text{traj}}$ is the sum of losses from the RPN and the Box Head.

Training TCNN: As illustrated in Figure 5-3, TCNN has three different loss terms. We compute shape loss $\mathcal{L}_{\beta}$ and 3D joint loss $\mathcal{L}_{\text{joints}}$ by comparing our predictions with the ground truth provided by corresponding vision algorithms. We use the smooth L1 loss [93] for both of them. We note that in order to compute the joint locations in 3D world space, our model needs to predict the global translations $\boldsymbol{\Delta}$ as well. We use the bounding box centers and predicted local translations with respect to the box centers to obtain the global translations. Our TCNN also performs a gender classification and uses the SMPL model of the predicted gender to compute the vertex and the joint locations.

When training TCNN together with the PDD, we follow standard adversarial training schemes [107, 181] and use the following loss term for TCNN:

$$\mathcal{L}_{\text{prior}} = -\mathbb{E}_{\boldsymbol{\Theta}\sim p_E}\log(D(\boldsymbol{\Theta})), \tag{5.2}$$

where $D(\cdot)$ is our pose and dynamics discriminator.

The total loss for the TCNN is a sum of the terms:

$$\mathcal{L}_{\text{TCNN}} = \mathcal{L}_{\boldsymbol{\beta}} + \mathcal{L}_{\text{joints}} + \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{gender}}. \tag{5.3}$$

5.3 Experiments

We describe our dataset, implementation details, quantitative and qualitative results on shape and pose estimation, and analyze what is learned by the attention module.

5.3.1 Dataset and Implementation

Dataset: To train and test our model, we build a dataset containing 84 subjects (male and female). For each subject, we use an adapted version of the approach in [101] to obtain ground truth β vectors with the subjects in a canonical pose (Sec. 5.2.5) – we refer to this method as VideoAvatar. We obtain data for the subjects walking around and engaging in activities in 16 different environments around our campus, and use a co-located calibrated camera system to obtain ground truth keypoint locations for the subjects. Our camera system is mobile, allowing us to collect data in varied environments and build a representative dataset.

Implementation details: We use decomposed 4D convolutions (Sec. 5.1) with residual blocks. Each uses ReLu activation and Group Normalization [182]. We use 12, 3, 12 and 12 layers of convolution in our backbone, RPN, box head and TCNN, respectively. We also use 1 and 2 layers of spatially-distributed GRU for TPN and RCNN. Our self-attention module uses two fully connected layers with $tanh(\cdot)$ activation in the middle. Our PDD model uses 12 layers of 1D temporal convolution, followed by a fully connected layer. We implement our model in PyTorch. Our model is trained with the Adam [165] optimizer for 40000 iterations.

5.3.2 Qualitative Evaluation for Shape and Pose

RF-Avatar produces realistic meshes: Figure 5-4 shows the 3D meshes produced by our model for different poses and subjects, as compared to the RGB images captured by a co-located camera. As can be seen, qualitatively, the estimated meshes are realistic, and agree well with the body shapes of different subjects. Our model also handles different body shapes (for male and female subjects), poses, and multi-person scenarios effectively. In addition, considering the bottom row of images in Figure 5-4, our model can produce accurate meshes for partially occluded subjects, subjects behind a wall, and subjects in poor lighting conditions; a vision-based system cannot produce full meshes in these situations.

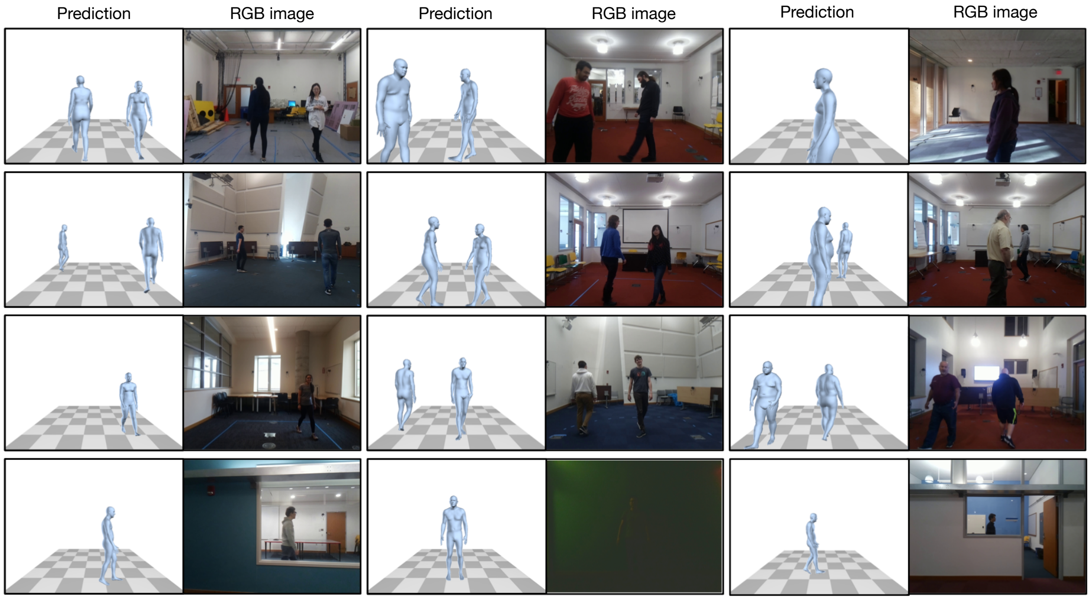


Figure 5-4: **Human mesh prediction from RF-Avatar.** We show images for visual reference. Our model captures different body shapes, poses, and multi-person scenarios effectively. The bottom row shows that RF-Avatar works despite occlusion and bad lighting conditions.

RF-Avatar effectively captures variation in body shape: To evaluate the quality of body shape predicted by RF-Avatar, we compare our prediction with the body shape captured by VideoAvatar [101], shown in Figure 5-6. VideoAvatar leverages a sequence of images to estimate a body mesh. The recovered mesh is overlaid on the RGB image of each person and is shown on the right side of each pair. To better compare the difference in body shape, we take the predicted shape of a subject (obtained by averaging predictions over a window of 10 seconds) from RF-Avatar and render the resulting mesh (in the same pose as VideoAvatar) and overlay it on the same background. This is shown on the left side of each pair. We see a close qualitative agreement between the ground truth and the output from RF-Avatar for male and female subjects with different body shapes.

RF-Avatar encodes human motion dynamics: Figure 5-5 demonstrates how our model can produce dynamic 3D meshes for different people over time, and how these meshes look realistic. We can see how the two subjects perform walking and lifting actions, and the produced meshes over time closely map to the performed actions.

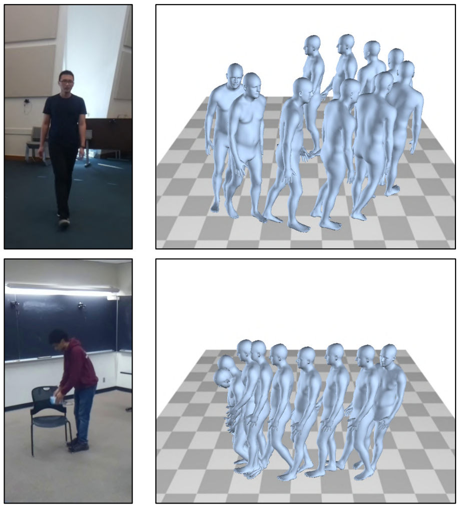


Figure 5-5: **Dynamic human meshes predicted from RF-Avatar.** RF-Avatar can capture dynamic mashes for different actions, including walking (top image) and lifting an object (bottom image).

5.3.3 Quantitative Evaluation for Shape and Pose

We now present quantitative results for our method, evaluating its performs on standard pose and body shape metrics. We also conduct ablation studies comparing with variants of our model that lack a particular component, namely variants that do not have supervision on the β parameters, do not use an attention mechanism, and use a frame-based discriminator (as in [104, 106]).

Metrics: We report the commonly used 3D joint metric Mean Per Joint Position Error (MPJPE). We also compute the per-vertex error as the average vertex to surface distance between the predicted mesh and the ground truth.

Table 5-1 shows the results for MPJPE and Per-vertex error respectively. As can be seen, for both MPJPE and per-vertex error, assessing recovered pose and shape quality respectively, the model that incorporates supervision for $\boldsymbol{\beta}$, self-attention, and the temporal discriminator, performs the best across all metrics. Of particular note is how the MPJPE drops from 6.05 cm to 6.88 cm when we do not use the temporal discriminator, demonstrating the value of the PDD in learning motion dynamics to help resolve ambiguities. We also see the importance of adding strong supervision for $\boldsymbol{\beta}$: the per-vertex error increases from 1.88 cm to 4.70 cm when it is removed. We also note here that the previous image-based mesh recovery methods have an MPJPE error around 8.8 cm [104] and a Per-

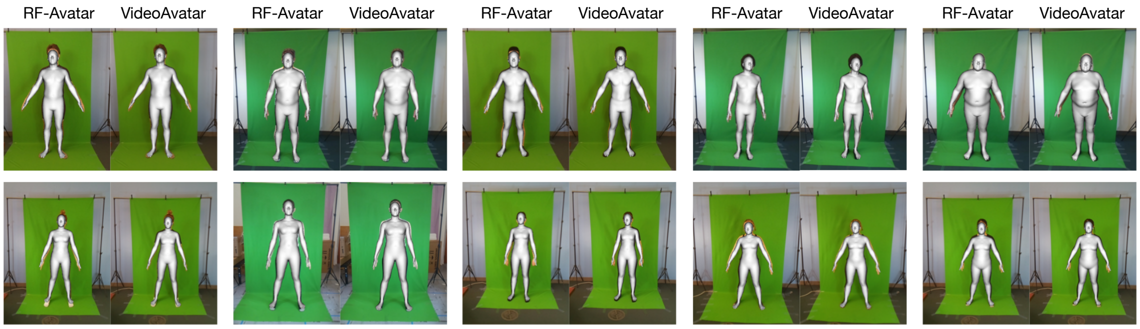


Figure 5-6: **Comparison of body shape recovered from RF-Avatar and VideoAvatar.** We render the mesh with the predicted shape estimated by RF-Avatar and the ground truth shape estimated by VideoAvatar and overlaid both on top of the corresponding RGB image.

vertex error around 11.8 cm [105]. Aside from the difference in datasets, we believe this difference in performance can be attributed to the fact that RF signals capture information about 3D space and our RF-based model is trained with stronger supervision than image-based methods.

We further see that the results using the TPN output (top row) are similar to the results using the ground truth bounding boxes (bottom row), illustrating the effectiveness of our entire detection, tracking, and shape estimation pipeline. This applies for both pose and shape metrics.

	MPJPE (cm)	Per-vertex error (cm)
RF-Avatar	**6.05**	**1.88**
No β loss	6.72	4.70
No attention	6.43	2.55
Frame-based disc.	6.88	2.24
With ground truth boxes	**5.75**	**1.65**

Table 5-1: **Joint and vertex errors, assessing pose and body shape quality respectively.**

Table 5-2 compares the results of our model for the shape and pose metrics for the total occlusion (through-wall) and line-of-sight scenarios. We see that our model performs well in the through-wall setting, even though it was never trained directly on through-wall data.

	3D MPJPE (cm)	Per-vertex errors (cm)
Line-of-sight	5.84	1.79
Through-wall	6.26	1.97

Table 5-2: **Results in the line-of-sight and through-wall settings.**

5.3.4 Analysis of Self-Attention

Table 5-1 shows that adding the self-attention module helps our quantitative results on shape and pose metrics. Self-attention helps our model better combine information over time when estimating the shape vector. We visualize the learned multi-headed attention maps in Figure 5-7. Focusing on the second attention component first, we see that it has high activation for timesteps 11 and 12. The high activation at these times indicates that they may contain important shape information. When comparing with the RGB images around timesteps 11 and 12, we see that the subject is facing the radio and waving at these times, so these timesteps likely contain reflections from his arm and provide important information about his upper limbs.

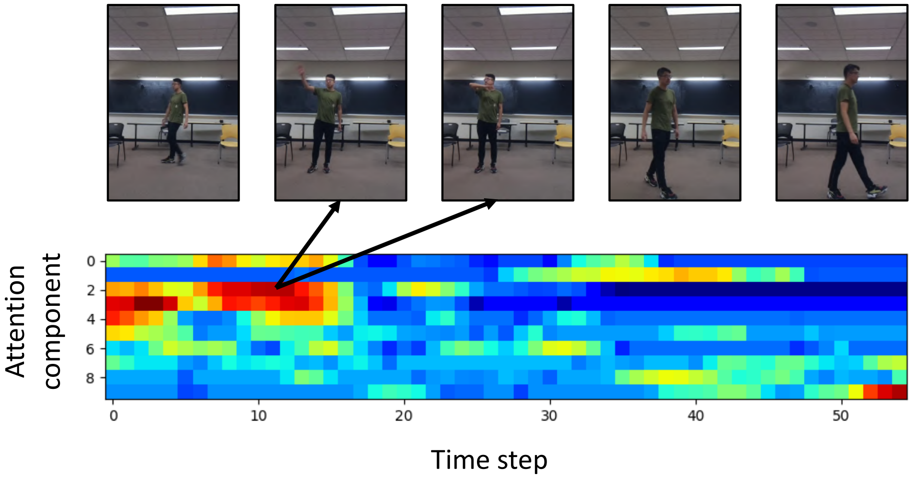
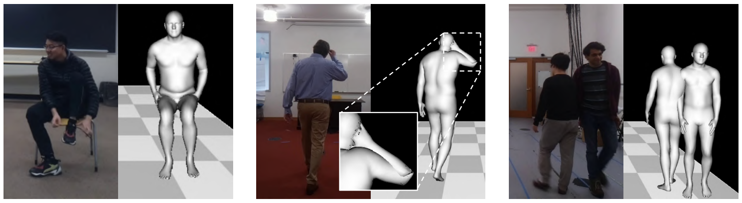


Figure 5-7: **Learned attention maps over time for the different attention heads.** We see that different attention components activate differently when the person is turning, waving hands and showing his side to the sensor.

5.3.5 Failure Modes

We analyze the failure cases of RF-Avatar. Typical failure examples are caused by (a) unusual body poses, (b) interpenetration of body meshes [102, 183], and (c) highly crowded



(a) Unusual body pose (b) Interpenetration (c) Crowded scene

Figure 5-8: **Typical failure cases of RF-Avatar.**

scenes where people are very close to each other. In Figure 5-8, we present examples of the typical failure cases. Figure 5-8(b) shows that RF-Avatar fails to handle unusual body poses (e.g., tying shoes). In Figure 5-8(b), interpenetration of estimated body meshes happens when the person raise his hand to hold glasses. In crowded scenes (e.g., Figure 5-8(c)) where people are very close to each other, RF-Avatar produces overlapped body meshes. Failure modes (a) and (b) are related to our choice of body mesh model, while failure mode (c) is due to the relatively low spatial resolution of RF signals in comparison to visible light.

■ 5.4 Conclusion

This section presented RF-Avatar a system that recovers dynamic 3D mesh models of the human body using RF signals. RF-Avatar is trained using cross-modality supervision from state-of-the-art vision algorithms, yet remains effective in situations that challenge vision systems, such as in poor lighting, and when subjects are occluded. We believe this work paves the way for many new applications in health monitoring, gaming, smart homes, etc. RF-Avatar significantly extends the capabilities of existing RF-based sensing systems, and the principles involved in its design could be utilized to improve the performance of existing computer vision methodologies.

CHAPTER 6
Assessment of Medication Self-Administration using Artificial Intelligence

Poor medication adherence is a major healthcare problem, contributing to 10% of hospitalizations, 125,000 deaths per year and up to $290 billion in annual cost in the United States alone [39, 40]. A crucial step toward achieving medication adherence is ensuring proper MSA – that is, ensuring that patients take medications at the prescribed time and use the proper self-administration technique. When patients fail to perform MSA properly, the net result is that the medication is not delivered to its intended action site, causing failures in managing the condition [125]. Unfortunately, MSA errors are common, particularly in chronic diseases where up to 50% of patients do not take medications as prescribed [184, 185]. The problem is exacerbated when medication delivery requires devices such as insulin pens or inhalers. MSA errors associated with medication delivery devices (for example, not shaking the inhaler before use or not priming the insulin dose) result in administration failures, subsequent high levels of non-adherence, reduced disease control and unscheduled use of healthcare resources [41, 42, 43, 44]. Physicians report that up to 70% of their patients do not take their insulin as prescribed [41, 42]. Similarly, over 50% of patients who use inhalers do so erroneously regardless of the inhalation device used [43]. These statistics are alarming given that hundreds of millions of patients worldwide depend on these devices for their medication administration [44, 186].

Addressing the above problems requires adequately assessing patients' MSA and detecting MSA errors. Existing solutions to assess MSA typically require direct observation by health professionals – that is, a clinician or pharmacist watches the patient as she uses her inhaler or insulin pen and guides her through the proper administration technique. For example, the clinician would explain to the patient that she needs to shake the inhaler, fully exhale, inhale a dose and hold her breath for 10 seconds before exhaling. The clinician would also watch the patient performing these steps and alert her if she fails to follow the proper administration technique. Although this approach can be used in the clinic, most MSA errors occur at home and away from the observation of a clinician. Furthermore, patients' performance in front of a clinician might be unrepresentative of their actual MSA, as patients tend to perform better when assessed by a clinician, a phenomenon known as 'white-coat compliance' [187]. Even when patients receive initial training on their devices by a clinician, MSA errors occur over time due to forgetfulness or recklessness in adhering to the prescribed administration time, frequency or technique [188]. As a result, many MSA errors end up undetected until they manifest as serious health problems or admissions to the emergency room [188].

This study was motivated by the question of whether we could use AI to assist with the observation of patients at home and provide a continuous assessment of their MSA. We present an AI-based solution that achieves this goal in an accurate, efficient and cost-effective manner. Our solution avoids cameras, which many patients find to be intrusive when deployed in their homes [189]. Instead, our solution uses a Wi-Fi-like sensor deployed in patient homes. (The sensor transmits signals around the Wi-Fi frequency range using the frequency-modulated continuous-wave (FMCW) technique. A detailed description of the radio sensor can be found in the Methods section.) The sensor transmits a very low-power wireless signal (1,000 times lower power than standard Wi-Fi), and our system analyzes the reflections of the signal from the environment using AI techniques. Because up to 60% of the human body is water, it reflects the surrounding radio signals and modulates them with the person's movements [8]. Past work has shown that such radio reflections can be used to capture breathing and heart rate, detect falls and monitor sleep [68, 55, 5]. In this study, we focused on movements associated with MSA events. Our AI system, embedded in the sensor, analyses the radio reflections from the environ-

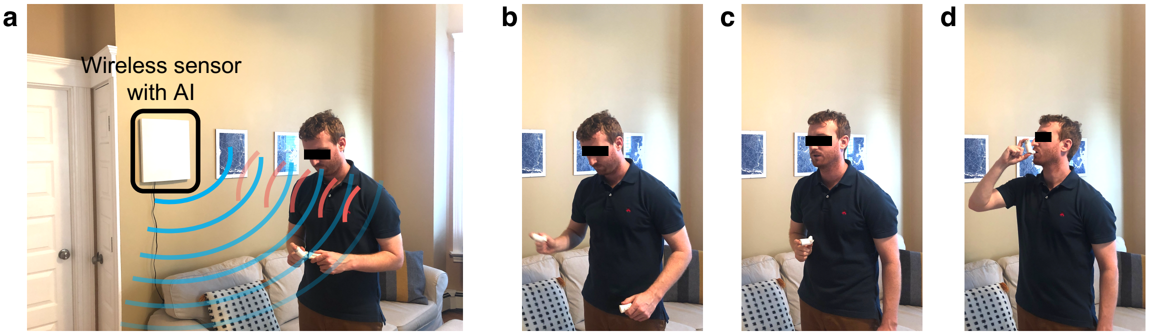


Figure 6-1: **A use case illustration of the wireless Ai-based system to monitor individual MSA with an inhaler device. a,** The wireless sensor is mounted on the wall, analyzing the surrounding radio signals using AI. The AI solution would detect when the person started to use an inhaler. **b–d,** Our AI solution also tracks the motion during the MSA event and detects that the person shook the device, exhaled before use and, finally, inhaled a dose. (We obtained informed consent from the participant for the use of his photographs.)

ment to track the specific movements associated with MSA and to detect when a patient administers her medication using an inhaler or insulin pen. It further examines the wireless reflections to detect whether the patient has followed the required steps of using the medication device and generates an alert if the patient fails to follow the proper technique (for example, forgot to prime her insulin pen or shake her inhaler). This AI-based solution works in a contactless and passive manner, introducing no burden on the patient, caregiver or health personnel.

Figure 6-1 illustrates a use case of our AI system at home, where it assesses the individual's MSA with an inhaler. The wireless sensor is mounted on the wall like a Wi-Fi box (Figure 6-1.a). There is no need for cameras, wearable sensors or any additional smart devices. The wall-mounted sensor analyzes the surrounding radio signal using AI methods. In this case, it would detect an instance of MSA using an inhaler and document the corresponding time. The AI solution also tracks the motion of the person and detects that the person shook the device, exhaled before use and, finally, inhaled a dose (Figure 6-1.b–d), which are required steps for MSA with an inhaler.

To build the AI-based solution, we designed a study where health professionals and the AI-based solution simultaneously observe MSA events with insulin pens and inhaler devices. To emulate real-world standard practice for first-time users of these medications, participants were trained by a pharmacist to perform MSA according to current guidelines and recommendations [45, 46] and were then asked to demonstrate their MSA technique.

During the experiment, human observers provided an MSA assessment that included the time window of each MSA event and the errors made during each event (if any). We used the MSA assessment provided by human observers to train and evaluate our AI algorithm. In total, we collected a large dataset that consists of 47,788 examples, where each example is a 2-min recording of radio signals. This dataset has 1,203 positive examples of MSA events with insulin pens and inhaler devices; about half of these MSA events are performed with no errors, whereas the other half includes some errors (for example, not shaking the inhaler before use or not holding one's breath after inhaling the dose). The dataset also includes 46,585 negative MSA examples corresponding to common home activities that do not involve MSA, such as cooking, eating, typing and interacting with objects such as glasses, clothes, microwaves and hairdryers. The MSA examples in the dataset were performed by 107 healthy individuals whose ages varied from 18 to 72 years. The dataset was divided into training and testing sets that we used to train and evaluate the AI system, respectively.

Extensive experimental results (that are detailed in the Results section) demonstrate that our AI-based solution can reliably detect the occurrence of MSA events. Specifically, the AUC was 0.992 for detecting the use of an inhaler and 0.967 for detecting the use of an insulin pen. These results indicate that an AI system could be used at home to monitor whether patients use their inhalers and insulin pens following the prescribed time and frequency.

The experimental results also show that the AI solution can accurately evaluate whether the individual correctly followed the required steps for administering her medication using an inhaler or insulin pen. Adherence to the proper steps while performing MSA is crucial or disease management and therapeutic effectiveness [126]. For example, failure to follow the correct steps when using an insulin pen can lead to hyperglycemia or severe hypoglycemic episodes for patients with diabetes [190, 191]. Similarly, failure to follow the recommended steps during inhaler administration contributes to symptom exacerbations and subsequent reduced quality of life for patients with asthma and patients with chronic obstructive pulmonary disease (COPD) [192, 193, 194]. Our results show that the AI system reliably detects both 1) missing key steps during the administration process (for example, not shaking the inhaler before use or not priming the insulin pen) and 2) patients not following duration-based requirements (for example, not holding the insulin pen after injection for 10 seconds).

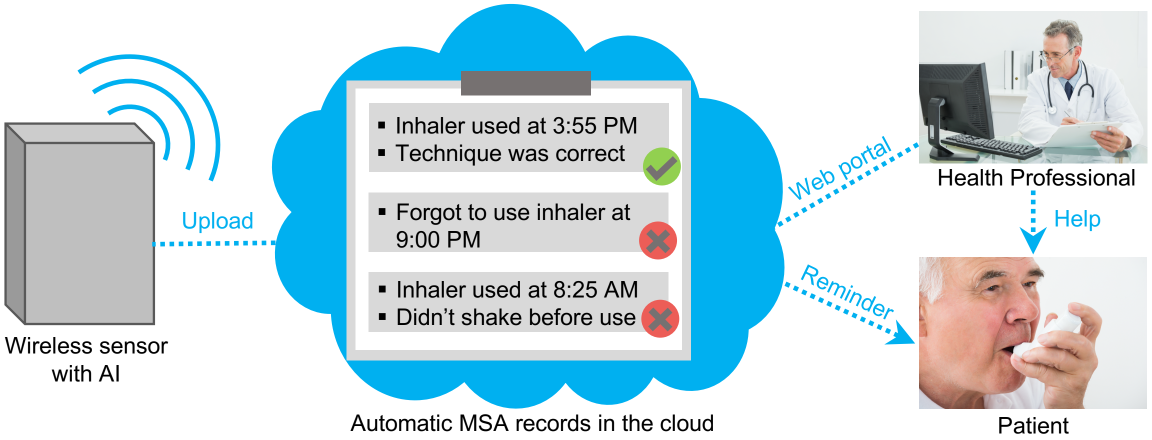


Figure 6-2: **Potential integration of our system into care management.** Our wireless sensor with AI will continuously and automatically analyze the radio signals and document MSA assessment results in the cloud. The patient will receive reminders if she fails to take the medication at the prescribed time. Authorized health professionals can also access these records via a web portal to learn which patients have difficulties with their MSA and the types of errors they experience. The health professionals can reach out to the patient to corroborate these results and make a clinical judgment if necessary.

Figure 6-2 illustrates how we envision such an AI-based solution that could be used in patient homes to help detect and address MSA errors. Our wireless sensor would be deployed in the patient's home. The AI system would continuously and automatically analyze the radio signals and document MSA assessment results, which are uploaded over the internet and appended to the patient's digital health record. Reminders will be sent to the patient if she fails to take the medication at the prescribed time. Authorized health professionals will also be able to access these records via a web portal to learn which patients have difficulties with their MSA and the types of errors they experience. The health professionals can then reach out to the patient to corroborate these results and make a clinical judgment (for example, whether more training on medication device administration is needed for the patient).

The recent outbreak of Coronavirus Disease 2019 (COVID-19) emphasizes the need for an automated and contactless solution for assessing MSA at home. The stay-at-home orders make it even more difficult to assess MSA through direct observation by health professionals. At the same time, individuals suffering from asthma, COPD and diabetes are at higher risk for severe illness from COVID-19 [195, 196]; hence, it is even more critical to ensure that they take their medications with the proper administration technique. Our automated and contactless AI-based MSA assessment solution could help these vulnerable populations to control their chronic conditions. It also enables health professionals to remotely monitor the MSA of their patients, without risks of contagion.

Our work shows how advances in AI can address an important unmet need in healthcare [41, 42, 43, 44, 186], by continuously monitoring the MSA of patients in their homes, detecting when patients fail to use their medication devices as prescribed and providing patients with feed-back on their medication administration technique and whether it follows the required steps. More generally, the work opens the door to the integration of AI-based solutions in care management through in-home passive, unobtrusive and contactless patient monitoring. Such integration could improve outcomes for patients and reduce the cost of healthcare.

■ 6.1 Method

Experiment Design. When designing the experiments in this study, we aimed to emulate the real-world scenarios of how patients use medication delivery devices. Patients typically receive training from pharmacists or other health professionals on how to use their medication delivery device when they are prescribed such medication for the first time41. To emulate the real-world scenarios, we chose individuals without prior experience with insulin pens and inhalers and had them trained by a pharmacist to use those devices. The pharmacist followed a standard procedure where he first taught the individual the MSA process and then asked the individual to demonstrate their MSA technique and ensured that the individual correctly simulated the administration of their insulin pen and inhaler device. After the initial training session, the individual performed MSA in front of a wireless sensor and a camera that recorded videos for annotation purposes. In addition to performing MSA, individuals were instructed to perform other activities, such as typing, cooking, eating and interacting with surrounding objects. We annotated the exact time window for every step involved in each MSA event and the types of errors that were made. To mitigate the imbalance between MSA events with and without errors and facilitate the development of AI models, we asked the individuals to purposely simulate errors during the experiment sessions. Note that, during both the initial training session and the later experiment session, all the MSA events were performed using placebo devices, and no medication dose was actually administered.

Individuals and dataset. A total of 107 healthy individuals (18–72 years of age) were recruited for this study. The individuals performed 1,203 positive examples of MSA events with insulin pens and inhaler devices at 40 different locations (offices, lounges, seminar rooms, kitchens, halls, etc). Positive MSA events were compared against a total of 46,585 instances of negative MSA examples corresponding to common activities that do not involve MSA. Of the 1,203 MSA events, 620 used insulin pens, and 583 used inhalers for administration. For the MSA events with insulin pens, 150 of them missed a common step, and 155 of them failed to comply with duration requirements. For the MSA events with inhaler devices, 149 of them missed a common step, and 168 of them failed to comply with duration requirements. None of the MSA events simultaneously missed a step and failed to comply with duration requirements.

During the experiments, the individuals were allowed to move freely in the space and perform the MSA at a location of their choice, within 10 m from the wireless sensor. The individuals were recorded both in sitting and standing positions via the wireless sensor. They were allowed to pick any orientation with respect to the wireless sensor (that is, face the sensor or show their sides to it) except for having their back facing the sensor.

RF sensing technology. Recent advances in RF sensing have developed systems that can capture human motion and infer biometric information, such as respiration, heart rate, gait speed, mobility, sleep stages and human pose [8, 68, 55, 1, 197, 2, 6]. Similarly to this past work, we used a radio sensor that employs FMCW and antenna arrays. The system works by transmitting low-power RF signals (1,000 times weaker than Wi-Fi) and receiving reflections from nearby people. Because up to 60% of the human body is water, it reflects the radio signals and modulates them with the person's movements, capturing important information about the person's MSA technique. The radio is equipped with vertical and horizontal antenna arrays, each of which has 12 antennas. It transmits an FMCW chirp sweeping the frequencies from 5.4 to 7.2 GHz. The combination of FMCW and antenna arrays allows the radio to separate RF reflections from different areas based on their distance (that is, range) and spatial direction (that is, angle of arrival) with respect to the radio sensor [8, 2]. This property allows the system to separate reflections from different people and process them independently.

We processed the RF signal into three-dimensional (3D) tensors indicating the amount of RF reflection from each point in the 3D space. We generated 30 such tensors (that is, frames) every second.

Building the AI-based model. Our AI-based model processes RF signals through three stages to detect and assess MSA events. The first stage uses a neural network model to localize and track people in the environment, zooming in on each individual while eliminating noise and interference from other people and objects in the environment. The second stage uses another neural network model to perform frame-wise prediction of MSA steps, where each frame is a snapshot of the 3D RF tensor at one point in time. Finally, the third stage decodes the frame-wise predictions into start time and end time of each MSA step and analyzes the sequence of steps to determine whether an MSA event has occurred. Below, we describe all three stages in detail.

Stage 1: The first stage of the processing takes a stream of radio signals as input and outputs bounding boxes43 representing the spatial locations of each individual. By focusing on RF reflections from the spatial locations indicated by the bounding boxes, our model zooms in on each individual while eliminating noise and interference from other people and objects in the environment. We used the same neural network as previous work [2] to localize and track people in the environment. This neural network model uses a 12-layer ResNet to extract features from RF signals together with a region proposal network that outputs bounding boxes [2].

Stage 2: The second stage takes the RF frames focused on a specific individual from the previous stage as input and outputs for each RF frame a probability score of the frame belonging to each of the MSA steps. The neural network used in this stage has a UNet structure [198] with 3D convolutional layers. Specifically, it has eight residual blocks, each of which consists of three convolutional layers, along with group normalization layers and exponential linear unit layers [182, 199]. It also interleaves four long short-term memory layers [200] within the last four residual blocks to capture temporal information. This subnetwork is trained using human annotations of MSA steps. The model is implemented in PyTorch. During training, the weights of the model are randomly initialized, and we use cross-entropy loss computed for each RF frame. Adam optimizer is used with a learning rate of 3×10^{-4}. We use a batch size of 4 on four NVIDIA TITAN Xp graphical processing units with distributed data parallelization. The model is trained for 100,000 iterations with a $10\times$ learning rate decay after 20,000 and 50,000 iterations.

Stage 3: The third stage of the processing decodes the frame-wise MSA step probabilities to estimate the start time and end time of each MSA step and determine whether an MSA event has occurred. We adopt beam search decoding, which is widely used in speech and handwritten text recognition for decoding the output of neural network models [201]. At a high level, there is an analogy between recognizing a spoken word by detecting the sequence of its phonemes and detecting an MSA event by detecting the sequence of its steps (and their corresponding RF frames). The beam search decoding algorithm considers all the frames jointly and uses language models as prior knowledge to output a coherent sequence of characters/words as opposed to a greedy decoding scheme that decodes each frame independently. Similarly, the beam search decoding algorithm in our model uses priors of the transition probability between MSA steps and the step duration, which are based on the statistics of the training data. The beam search decoding also computes a score (that is, log likelihood) for the decoded results. The score is normalized by the duration of the detected MSA event, and our model rejects all MSA events if their final score is less than a threshold. (The threshold is set to 0.4, which balances sensitivity with specificity).

Statistical methods. To evaluate the performance of MSA event detection, we used the following metrics: sensitivity, specificity, ROC curves, AUC and the estimation error of start time and end time, with sample sizes as given. To evaluate the performance of MSA error detection, we used the following metrics: estimation error of step duration, sensitivity and specificity of MSA error detection, ROC curves and AUC.

Sensitivity and specificity are calculated as (TP: true positive; FN: false negative; TN: true negative; FP: false positive):

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP + FN}}, \tag{6.1}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN + FP}}. \tag{6.2}$$

We plotted ROC curves that demonstrate the tradeoff between sensitivity and specificity, as the detection thresholds are varied. When reporting the sensitivity and specificity, we used a detection threshold of 0.6 for MSA event detecting, a detection threshold of 0.25 s for detecting missing steps and a detection threshold of 8 s for detecting the error of failing to comply with duration-based requirements. We followed standard procedures to calculate the 95% CI for sensitivity and specificity [202]. We also reported AUC, which is the area under the corresponding ROC curves showing an aggregate measure of detection performance.

We computed the error between the predicted start (or end) time and the ground truth start (or end) time of events as (t_s: ground truth start time: t_e: ground truth end time: $\hat{t_s}$: predicted start time; $\hat{t_e}$: predicted end time):

$$\mathrm{error_s} = |\hat{t_s} - t_s|, \tag{6.3}$$

$$\mathrm{error_s} = |\hat{t_s} - t_s|. \tag{6.4}$$

The error of step duration estimation is computed as (d: ground truth duration; $\hat{d}$: predicted duration):

$$\mathrm{error_d} = |\hat{d} - d|. \tag{6.5}$$

We reported the error of start/end time estimation and step duration estimation with box plots. For each box plot, the central line indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to 1.5 times the interquartile range. Points beyond the whiskers are plotted individually using the '+' symbol.

6.2 Evaluation

Our system was built and validated on a large dataset that consists of 47,788 examples of MSA events and home activities that do not involve MSA from 107 participants (53 females and 54 males), where each example is a recording of radio signals up to 2min. A total of 40 different locations (such as offices, lounges, seminar rooms, kitchens and halls) were chosen to allow for variation between environmental conditions. Participants were allowed to move freely in the space and perform the MSA at a location of their choice, within 10 m from the wireless sensor. They were either standing or sitting when performing MSA and were allowed to pick any orientation with respect to the wireless sensor except for having their back to the sensor. For the MSA events with insulin pens, 150 events missed a common step, and 155 of them failed to comply with a specific duration requirement. For the MSA events with inhaler devices, 149 of them missed a required step, and 168 of them failed to comply with duration requirements.

We trained a neural network model that takes a stream of radio signals as input. It first detects and tracks the location of each individual in the environment. It then zooms in on the radio signals pertaining to each individual and predicts the occurrence of an MSA event. Instead of detecting the whole MSA event directly, our model detects the steps involved during administration and only claims an occurrence of an MSA event when multiple steps happen one after another. Detecting MSA events in this way improves our model's robustness to variations among different people and enables the detection of MSA errors. Details of the model are described in the Methods section.

We performed K-fold cross-validation (K=5) to evaluate our model. Specifically, the dataset was randomly split into five equal-sized subsets. A single fold was retained as test data, and the remaining four folds were used for training. This process was repeated five times, with each of the five folds used exactly once as the test data. The folds were divided such that participants who appear in the training data do not appear in the testing data and vice versa. To balance the number of participants across folds, we imposed no constraints on allocating sites to folds. We note that the cross-validation was not used for hyperparameter tuning.

■ 6.2.1 Detection of MSA events

Our model detects MSA events in a sliding-window fashion. Specifically, it detects whether an MSA has happened for each 2-min window. To evaluate the performance of our model, we compared its predictions with ground truth provided by human annotations. Figure 6-3.a-b show the receiver operating characteristic (ROC) curves for detecting MSA events with insulin pens and inhaler devices, respectively. When computing sensitivity and specificity, positive examples indicate MSA events, whereas negative examples indicate non-MSA events. Our system detected the occurrence of an insulin pen administration event with a sensitivity of 87.58% (95% confidence interval (CI), 84.7–90.0%) and a specificity of 96.06% (95% CI, 95.9–96.2%) and an AUC of 0.967. Similarly, inhaler administration events were detected with a sensitivity of 91.08% (95% CI, 88.4–93.2%) and a specificity of 99.22% (95% CI, 99.1–99.3%) and an AUC of 0.992. We note that the specificity or the false-positive ratio is computed for windows of non-MSA events such as eating, drinking or putting on clothes, not just any window of radio frequency (RF) signals. The number of false positives when considering any window of RF signals is significantly smaller in real-world deployment. Specifically, we leveraged a dataset in which the radio was used to monitor patients with Parkinson's disease and control individuals (that is, healthy individuals) for over 1 month [127]. Because none of the individuals in this dataset used inhalers or insulin pens, all detected MSA events could be considered false positives. We considered five homes from the study and used one full month of RF signals from each home. On average, the number of false positives over a whole month was 2.2 for insulin pens and 6.6 for inhalers.

We also looked at errors that our model made when estimating the start time and end time of an MSA event. Figure 6-3.c shows the box plots for the estimation errors, and Figure 6-3.d-e shows the cumulative distribution functions of the absolute estimation errors. Our system made an unbiased (that is, median-unbiased) estimation of the start time and the end time for both devices. For the start time estimation, the 50th percentile error was 0.6 s and 0.4 s, whereas the 90th percentile error was 2.0 s and 1.3 s for insulin pens and inhaler devices, respectively. Similarly, for the end time estimation, the 50th percentile error was 0.4 s and 0.3 s, whereas the 90th percentile error was 1.4 s and 0.9 s for insulin pens and inhaler devices, respectively. To put these errors in context, the average duration of MSA events based on human annotations was 65.27 ± 13.22 s for insulin pens and 34.30 ± 7.12 s for inhalers.

6.2.2 Evaluation of MSA techniques.

To evaluate the MSA technique, we partitioned an MSA event into constituent key steps based on recommendations pertaining to insulin pen and inhaler device administration [45, 46]. Figure 6-4 illustrates the details of these steps. Besides detecting the occurrence of an MSA event, our model also predicts the time window for each of the individual steps involved during the administration. To evaluate its performance, we compared the predicted duration of each individual step with human annotations. 6-5.a shows the box plots of the duration estimation errors for eight different steps during MSA events with insulin pens. Similarly, 6-5.b plots the duration estimation errors for six different steps during MSA events with inhaler devices. Our model made an unbiased duration estimation for all the steps, and the interquartile range was smaller than 1.5 seconds for all the steps.

Based on the detection of individual steps and estimation of their duration, we further

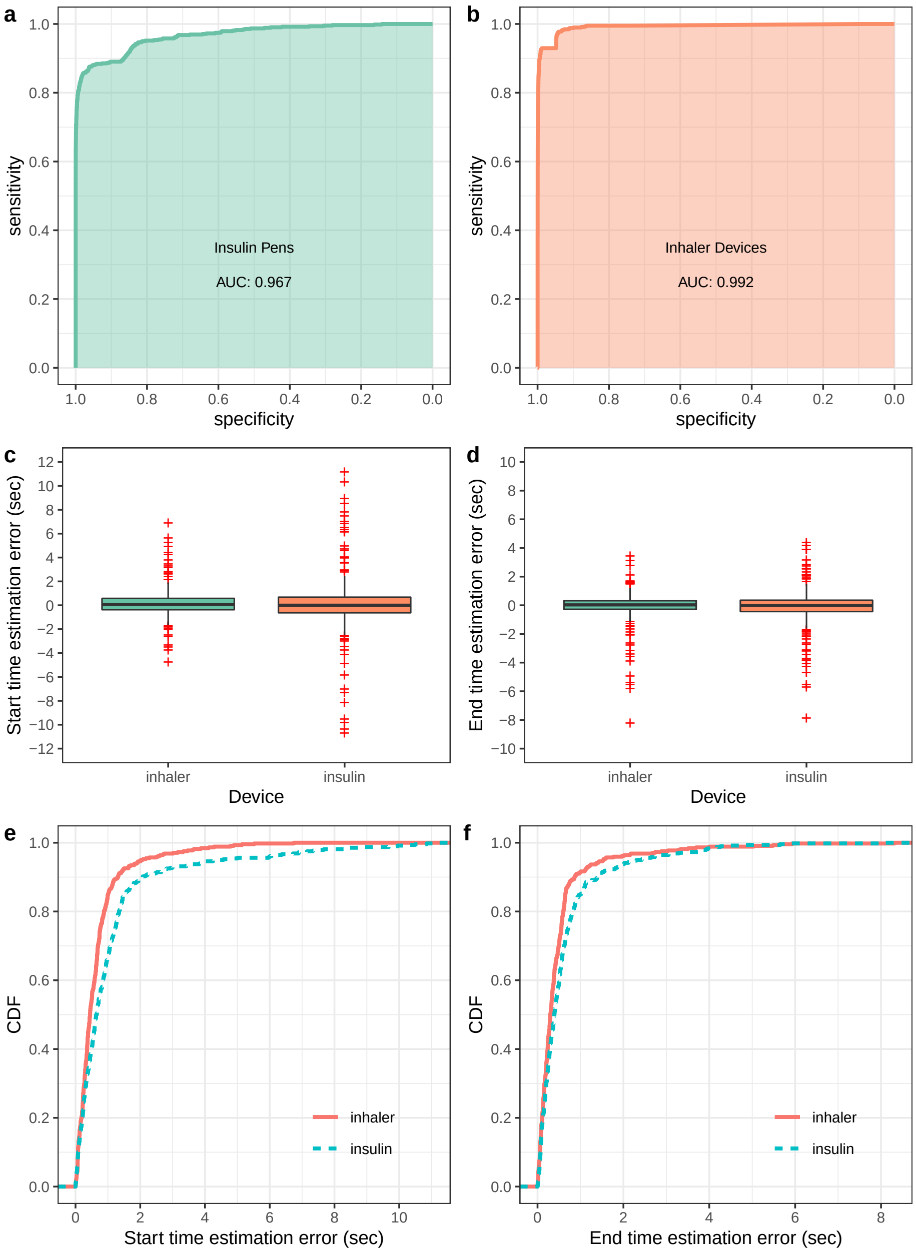


Figure 6-3: **Evaluation results for the detection of MSA events with insulin pens and inhaler devices.** **a,b** ROC curves for detecting insulin (n=47,205) and inhaler (n=47,168) administration. ROC curves demonstrate the tradeoff between sensitivity and specificity as the detection thresholds are varied. The AUC is an aggregate measure of detection performance (a model whose predictions are 100% correct will have an AUC of 1.0). **c,** Distribution of the errors for start time and end time estimation (n = 620 for insulin pens and n = 583 for inhalers). On each box plot, the central line indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to 1.5 times the interquartile range. Points beyond the whiskers are plotted individually using the '+' symbol. **d,e,** Cumulative distribution functions (CDFs) of the absolute error for start time and end time estimation (n=620 for insulin pens and n=583 for inhalers).

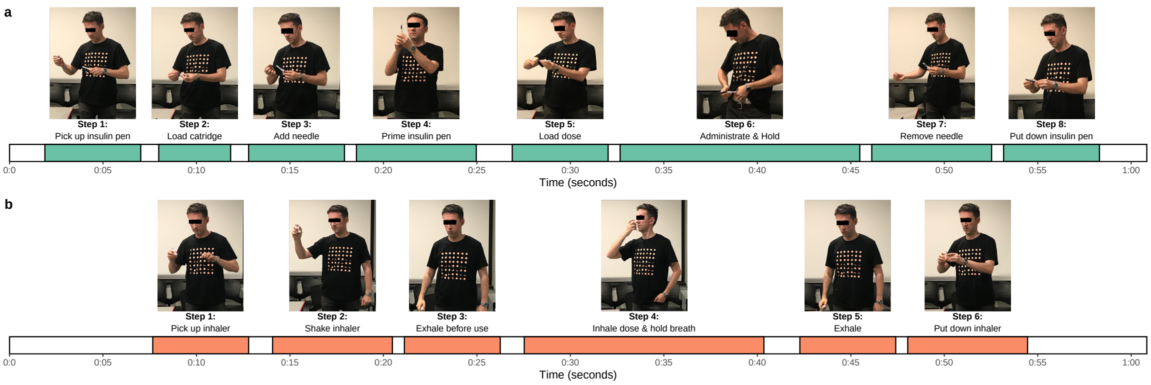


Figure 6-4: **Partitioning of key constituent steps of insulin pen and inhaler device self-administration enabling evaluation of administration technique.** Based on recommendations pertaining to insulin pen and inhaler device administration, we partitioned an MSA event into eight steps and six steps for insulin and inhaler administration, respectively. (We obtained informed consent from the participant for the use of his photographs).

looked at two common errors of MSA: 1) missing a key step during the administration process and 2) failure to comply with the duration requirements of device administration. The above MSA errors occur frequently and are associated with poor disease control outcomes [125, 126, 203]. For example, studies have reported that 37% of patients fail to shake their inhalers [203], and patients not holding their breath after inhalation was a prevalent (53%) error during inhalation device administration [203].

MSA errors with a missing step considered in this study were not priming the insulin pen (Step 4) and not shaking the inhaler device before use (Step 2). These steps are crucial to make sure these devices deliver the medication at the right dose. Specifically, priming the insulin pen ensures an unobstructed and free flow of insulin [126], and shaking the inhaler ensures proper mixture of particles and consistent dose delivery [192]. Figure 6-5.c-d show the ROC curves for detecting such errors during insulin and inhaler administration, respectively. Our system detected not priming the insulin pen with a sensitivity of 84.00% (95% CI, 76.9–89.2%) and a specificity of 92.55% (95% CI, 89.7–94.7%) and an AUC of 0.905. Similarly, our system detected not shaking the inhaler device before use with a sensitivity of 96.64% (95% CI, 91.9–98.7%) and a specificity 94.47% (95%, CI 91.8–96.3%) and an AUC of 0.967.

To evaluate our system's performance in detecting errors of failing to comply with duration requirements, we considered two duration-related common steps—namely, holding the insulin pen still for 10s after injection (Step 6 of insulin administration) and holding

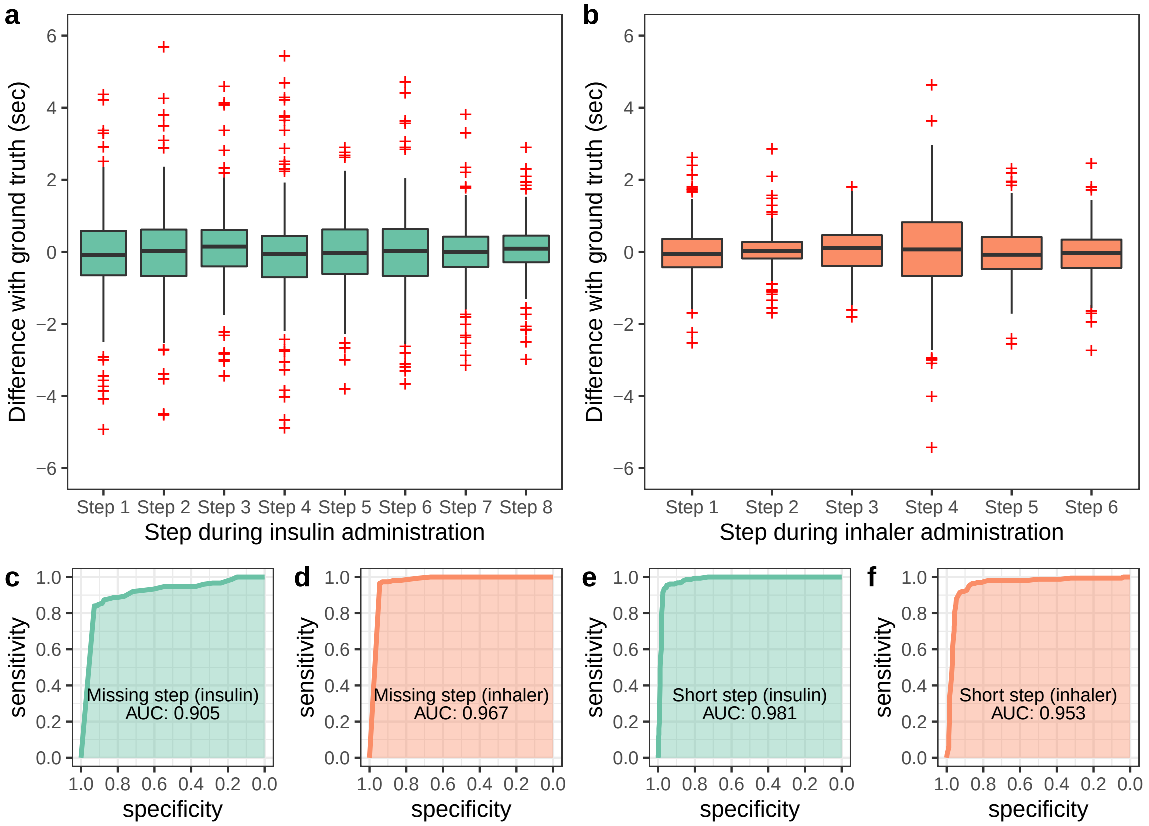


Figure 6-5: **Evaluation results for estimating the duration for constituent key steps and detecting MSA errors during MSA events with insulin pens and inhaler devices.** **a,b,** Distribution of the step duration estimation errors for insulin (n=620) and inhaler (n=583) administration, respectively. On each box plot, the central line indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to 1.5 times the interquartile range. Points beyond the whiskers are plotted individually using the '+' symbol. **c,d,** ROC curves for detecting MSA errors of missing a key step during insulin (n=620) and inhaler (n=583) administration. ROC curves demonstrate the tradeoff between sensitivity and specificity as the detection thresholds are varied. The shaded AUC is an aggregate measure of detection performance. **e,f,** ROC curves for detecting MSA errors of failing to comply with duration requirements during insulin (n=620) and inhaler (n=583) administration.

one's breath for 10 s after inhaling a dose (Step 4 of inhaler administration). We use the cutoff of 10 s base on clinical recommendations [126, 192, 204]. Specifically, holding the insulin pen for 10 s after injection ensures no insulin leakage or dribbling [126], and holding the breath after dose inhalation ensures adequate lung deposition, which occurs through sedimentation of particles [126, 192]. Figure 6-5.e-f show the ROC curves for detecting MSA errors when individuals failed to comply with the duration requirements during insulin and inhaler administration, respectively. Our system detected not holding the insulin pen still for 10 s after injection with a sensitivity of 94.19% (95% CI, 88.9–97.1%) and a specificity of 95.48% (95% CI, 93.1–97.1%) and an AUC of 0.981. Similarly, our system detected not holding breath after inhaling a dose with a sensitivity of 89.88% (95% CI, 84.0–93.8%) and a specificity of 92.04% (95% CI, 88.9–94.4%) and an AUC of 0.953. This performance rose further for detecting significant deviations from the recommendations (for example, failing to hold breath/pen even for a few seconds). For the insulin pen, the AUC increased to 0.986 and 0.993 for detecting duration shorter than 5s and 3s, respectively. Similarly, the AUC increased to 0.983 and 0.988 in the case of inhalers.

In Fig. 6 we show example outputs from our system. Figure 6a shows an example MSA event with the insulin pen. The top panel plots the predictions of our system on the time axis—that is, a detected MSA event with multiple boxes corresponding to all eight steps when using an insulin pen. The bottom panel shows the human annotation during the corresponding MSA event. Figure 6b shows an example when the individual missed a key step during the administration process. Note that the box corresponding to the step of priming the insulin dose (Step 4) is missing, and this error was successfully detected by our AI model. Figure 6c shows another insulin pen example where the individual failed to comply with administration duration requirements—namely, the individual failed to hold the pen for 10s after injecting the dose. Again, our model was able to detect this error, as the detected step of holding the insulin pen after injection (Step 6) was much shorter than 10 s. Similarly, Fig. 6d–f show example outputs with inhalers.

In Figure 6-6.a we show example outputs from our system. Figure 6a shows an example MSA event with the insulin pen. The top panel plots the predictions of our system on the time axis—that is, a detected MSA event with multiple boxes corresponding to all eight steps when using an insulin pen. The bottom panel shows the human annotation during the corresponding MSA event. Figure 6-6.b shows an example when the individual missed

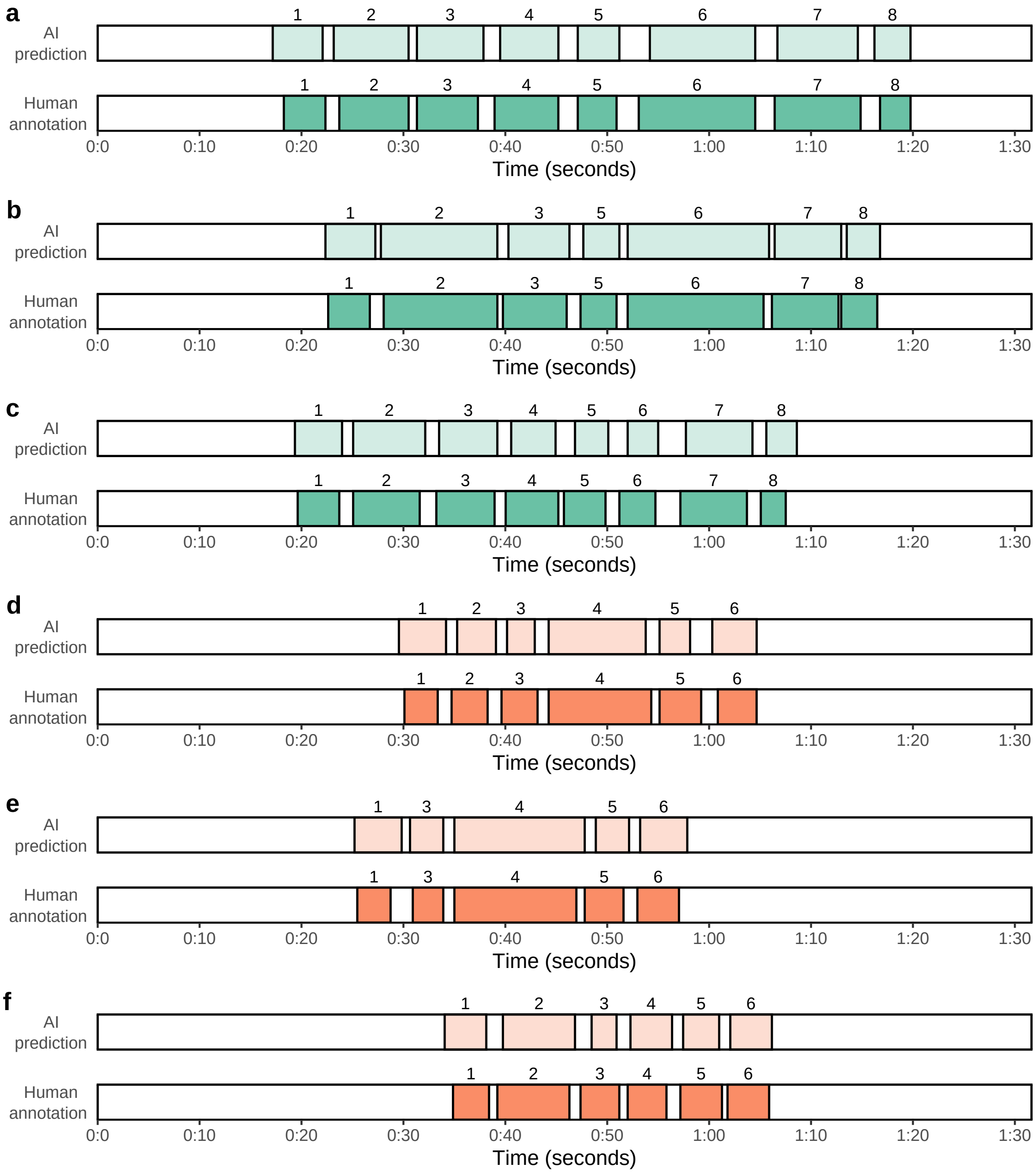


Figure 6-6: **Example outputs from our AI-based system. a,** An insulin administration event without any error. **b,** An insulin administration event that missed the step of priming the insulin pen (Step 4). **c,** An insulin administration event where the individual did not hold the pen for 10 s after injection (Step 6). **d,** An inhaler administration event without any error. **e,** An inhaler administration event that missed the step of shaking the inhaler device before use (Step 2). **f,** An inhaler administration event where the individual did not hold their breath after inhaling a dose (Step 4).

a key step during the administration process. Note that the box corresponding to the step of priming the insulin dose (Step 4) is missing, and this error was successfully detected by our AI model. Figure 6-6.c shows another insulin pen example where the individual failed to comply with administration duration requirements—namely, the individual failed to hold the pen for 10s after injecting the dose. Again, our model was able to detect this error, as the detected step of holding the insulin pen after injection (Step 6) was much shorter than 10 s. Similarly, Figure 6-6.d–f show example outputs with inhalers.

■ 6.3 Discussion

Here we described an AI-based solution for contactless at-home assessment of patient MSA using inhalers and insulin pens. Our solution is characterized by three properties: low overhead, informative and accurate. It is low overhead because it works in a passive and contactless manner without requiring patients or health professionals to observe, report or measure any parameters. It is informative because, in addition to detecting patient medication administration, it also assesses the patient's self-administration technique and informs her of errors and omissions of required steps. It is also accurate as demonstrated through our empirical results.

We think that the above three properties are important for the success of an MSA assessment solution. Past solutions for assessing MSA at home fall short of delivering all three properties. In particular, solutions that attach sensors to medication devices to monitor MSA [122, 123, 124] can impose a new burden on the patient, as they require the patient to regularly charge or replace their battery and bring the devices in the vicinity of a smartphone so they can upload their data. Although such solutions can detect dose release, they lack information on whether the patient followed the proper MSA technique to ensure adequate dose delivery – that is, the sensor captures the actuation and movements of the medication device itself but cannot capture the patient's actions and their sequence, which are crucial for correct MSA. To our knowledge, this system is the first to introduce an automated solution for assessing an individual's MSA technique and whether it follows the proper steps. Being able to assess MSA techniques is essential because failures to follow the proper techniques are common and have been associated with high non-adherence levels and subsequent poor disease outcomes [125, 126, 127, 128].

Another feature of our approach is its ability to generalize to different types of insulin pens and inhalers. The neural network models used in this study support both reusable and disposable insulin pens and the widely used metered-dose inhalers, as the constituent key steps that our system learned are similar regardless of the type of insulin or the medication delivered by the inhaler. Specifically, in the case of insulin pens, our model is trained to support both reusable and disposable pens by considering the cartridge-loading step as optional. In the case of inhalers, our model is trained on the MSA steps recommended for the widely used metered-dose inhalers. Because the main difference between different types of metered-dose inhalers is the actual drug administered (for example, salbutamol, ipratropium ot fluticasone) rather than the administration technique itself, our system works with all such inhalers. Furthermore, our model can be extended to work with dry powder inhalers, which do not require shaking before use. This can be done by using a flag to indicate that the patient uses a dry powder inhaler and, therefore, not declaring an MSA error when the shaking step is missed.

We think that the clinical implications of our system could be significant. We envision that this system will be able to provide continuous feedback for clinicians on their patients' MSA. Based on the feedback from our system, health professionals can then make a clinical judgment (for example, whether more training and education on medication device administration techniques is needed for the patient). Additionally, this system could contribute to patient empowerment and engagement in their health by giving them feed- back about their MSA technique and allowing them to avoid common MSA errors.

Although our AI-based solution provides an important improvement over the status quo, we also note that it has several limitations. First, our system was developed and tested with healthy individuals in laboratory conditions. We designed the experiment in this study to emulate the real-world scenarios of how patients use medication delivery devices after the initial training from pharmacists or other health professionals. Thus, we chose individuals without prior experience with the insulin pen and inhaler and had them trained by a pharmacist to use these devices. This also ensured that our participants' level of education and training in using medication devices was standardized, therefore mitigating reported barriers associated with individuals' lack of training and education when using an insulin pen and inhaler device [125, 205]. We think that this study provides an important first step toward enabling automatic MSA assessment at home. We envision that future work would validate the system with actual patients in their homes and study the effect of having such a system for automatic at-home MSA assessment on medication adherence. Future work could also evaluate potential confounding factors that might affect MSA errors, such as the patient's chronic conditions, dexterity issues, health literacy and education level.

Second, we focused on insulin pens and metered-dose inhalers and their common errors, but there are many other MSA devices and potential technique errors. Although this is a limitation of the specific neural network that we trained, the AI approach that we propose is general and can be adapted to other MSA devices and MSA errors.

Third, the ability of our system to detect MSA events in various locations in the home is limited by the coverage area of the radio. The radio device that we used in this study can assess MSA events in locations up to 10 m from the device. This is usually enough to cover several rooms in a home. If desirable, however, the whole home can be covered by deploying multiple radio devices. Still, patients might take their medications outside the home (for example, at work), leading to some MSA events being missed. Even when MSA detection is incomplete, the system continues to be useful. Specifically, it would provide health professionals with a list of missing MSA events, which allows them to discuss this information with patients to clarify whether the missing MSA events are due to incomplete information or the patients indeed did not take their medication. Furthermore, it would detect MSA technique errors, which are typically repeated by patients, and today often go undetected until direct observation from clinicians or poor disease outcomes [187, 206].

Fourth, the system does not detect MSA events if the person has his back to the radio, because most of the RF signals are blocked by the person's own body. Similarly to the previous limitation, this issue can be addressed by deploying a second radio in the environment with a different orientation.

Additionally, the exposition in this work focused on scenarios where the house has a single person who uses an inhaler and/or insulin pen. For homes with multiple patients who use inhalers or insulin pens, a user identification system based on RF reflections [1, 197, 14, 31] can be employed to resolve the ambiguity. Such systems use RF signals to accurately identify a person from a small set of people—for example, other residents in a home or co-workers in an office scenario. Because we only require identification from others at home, their methods apply to this scenario.

In summary, we developed an AI system that can successfully detect MSA events and assess a patient's MSA technique. Our system demonstrates how AI can be applied to ensure medication safety, specifically with device-based administration, in a manner that has minimal potential overhead for patients and health professionals.

Chapter 7
Learning Sleep Stages from Radio Signals

Sleep plays a vital role in an individual's health and well-being. Sleep progresses in cycles that involve multiple sleep stages: Awake, Light sleep, Deep sleep and REM (Rapid Eye Movement). Different stages are associated with different physiological functions. For example, deep sleep is essential for tissue growth, muscle repair, and memory consolidation, while REM helps procedural memory and emotional health. At least, 40 million Americans each year suffer from chronic sleep disorders [207]. Most sleep disorders can be managed once they are correctly diagnosed [207]. Monitoring sleep stages is beneficial for diagnosing sleep disorders, and tracking the response to treatment [208].

Prevailing approaches for monitoring sleep stages are inconvenient and intrusive. The medical gold standard relies on Polysomnography (PSG), which is typically conducted in a hospital or sleep lab, and requires the subject to wear a plethora of sensors, such as EEG-scalp electrodes, an ECG monitor, multiple chest bands, and nasal probes. As a result, patients can experience sleeping difficulties, which renders the measurements unrepresentative [209]. Furthermore, the cost and discomfort of PSG limit the potential for long term sleep studies.

Recent advances in wireless systems have demonstrated that radio technologies can capture physiological signals without body contact [79, 68, 6]. These technologies transmit a low power radio signal (i.e., 1000 times lower power than a cell phone transmission) and analyze its reflections. They extract a person's breathing and heart beats from the radio frequency (RF) signal reflected off her body. Since the cardio-respiratory signals are correlated with sleep stages, in principle, one could hope to learn a subject's sleep stages by analyzing the RF signal reflected off her body. Such a system would significantly reduce the cost and discomfort of today's sleep staging, and allow for long term sleep stage monitoring.

There are multiple challenges in realizing the potential of RF measurements for sleep staging. In particular, we must learn RF signal features that capture the sleep stages and their temporal progression, and the features should be transferable to new subjects and different environments. The problem is that RF signals carry much information that is irrelevant to sleep staging, and are highly dependent on the individuals and the measurement conditions. Specifically, they reflect off all objects in the environment including walls and furniture, and are affected by the subject's position and distance from the radio device. These challenges were not addressed in past work which used hand-crafted signal features to train a classifier [210, 211]. The accuracy was relatively low (∼64%) and the model did not generalize beyond the environment where the measurements were collected.

We present a new model that delivers a significantly higher accuracy and generalizes well to new environments and subjects. The model adapts a convolutional neural network (CNN) to extract stage-specific features from RF spectrograms, and couples it with a recurrent neural network (RNN) to capture the temporal dynamics of sleep stages.

However, a CNN-RNN combination alone would remain liable to distracting features pertaining to specific individuals or measurement conditions (i.e., the source domains), and hence would not generalize well. To address this issue, we introduce a new adversarial training regime that discards extraneous information specific to individuals or measurement conditions, while retaining all information relevant to the predictive task –i.e., the adversary ensures conditional independence between the learned representation and the source domains.

Our training regime involves 3 players: the feature encoder (CNN-RNN), the sleep stage predictor, and the source discriminator. The encoder plays a cooperative game with the predictor to predict sleep stages, and a minimax game against the source discriminator. Our source discriminator deviates from the standard domain-adversarial discriminator in that it takes as input also the predicted distribution of sleep stages in addition to the encoded features. This dependence facilitates accounting for inherent correlations between stages and individuals, which cannot be removed without degrading the performance of the predictive task.

We analyze this game and demonstrate that at equilibrium, the encoded features discard all extraneous information that is specific to the individuals or measurement conditions, while preserving all information relevant to predicting the sleep stages. We also evaluate our model on a dataset of RF measurements and corresponding sleep stages[1]. Experimental results show that our model significantly improves the prediction accuracy of sleep stages as shown in Table 7-1. In particular, our model has a prediction accuracy of 79.8% and a Cohen's Kappa of 0.70, whereas the best prior result for predicting sleep stages from RF signals [211] has an accuracy of 64% and a Cohen's Kappa of 0.49. This improvement is due to intrinsic differences between past models and the model in our work, which avoids hand-crafted features, and learns features that capture the temporal dependencies and transfer well to new subjects and different environments.

Table 7-1: **Automated Sleep Staging Systems**

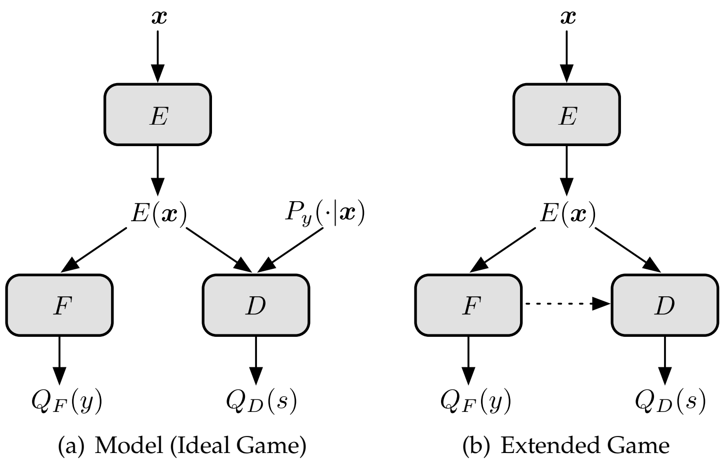
Signal Source		**Accuracy (*acc*/κ)**[1]	**Comfort**
EEG		High (83%/0.76)[2]	Low
Cardiorespiratory		Medium (71%/0.56)	Medium
Actigraphy		Low (65%/-)[3]	High
RF	State-of-the-art	Low (64%/0.49)	High
	Ours	**High (**79.8%/0.70)	**High**

[1] Four-class subject-independent classification accuracy on every 30-second segment.
[2] Some studies achieve accuracy over 90% [212] but they discard artifacts and use segments from the same night to train and test.
[3] Three-class classification based on 5-minute segment.

■ 7.1 Conditional Adversarial Model

Let $x \in \Omega_x$ be an input sample, and $y \in \{1, 2, ..., n_y\}$ an output label. Let $s \in \{1, 2, ..., n_s\}$ denote an auxiliary label that refers to the source of a specific input sample. We define $\boldsymbol{x} = [x_1, x_2..., x_t] \in \Omega_{\boldsymbol{x}}$ as the sequence of input samples from the beginning of time until the current time t.

[1] Dataset is available at: http://sleep.csail.mit.edu/



(a) Model (Ideal Game) (b) Extended Game

Figure 7-1: **Model and Extended Game.** Dotted arrow indicates that the information does not propagate back on this link.

In the context of our application, the above notation translates into the following: The input sample x is a 30-second RF spectrogram, and the output label y is a sleep stage that takes one of four values: Awake, Light Sleep, Deep Sleep, or REM. The vector $\boldsymbol{x}$ refers to the sequence of RF spectrograms from the beginning of the night until the current time. Since RF signals carry information about the subject and the measurement environment, we assign each input x an auxiliary label s which identifies the subject-environment pair, hereafter referred to as the source.

Our goal is to learn a latent representation (i.e., an encoder) that can be used to predict label y; yet, we want this representation to generalize well to predict sleep stages for new subjects without having labeled data from them. Simply making the representation invariant to the source domains could hamper the accuracy of the predictive task. Instead we would like to *remove conditional dependencies between the representation and the source domains*.

We introduce a multi-domain adversarial model that achieves the above goal. Our model is shown in Figure 7-1(a). It has three components: An encoder E, a label predictor F, and a source discriminator D. Our model is set up as a game, where the representation encoder plays a cooperative game with the label predictor to allow it to predict the correct labels using the encoded representation. The encoder also plays a minimax game against the source discriminator to prevent it from decoding the source label from the encoded representation.

A key characteristic of our model is the conditioning of the source discriminator on the label distribution, $P_y(\cdot|\boldsymbol{x})$ (see Figure 7-1(a)). This conditioning of the adversary allows the learned representation to correlate with the domains, but only via the label distribution –i.e., removes conditional dependencies between the representation and the sources.

The rest of this section is organized as follows. We first formally define three players E, F, and D and the representation invariance they are trained to achieve. In Section 7.1.1, we analyze the game and prove that at equilibrium the encoder discards all extraneous information about the source that is not beneficial for label prediction (i.e., predicting y). Training the ideal model in Figure 7-1(a) is challenging because it requires access to the label distribution $P_y(\cdot|\boldsymbol{x})$. To drive an efficient training algorithm, we define in Section 7.1.2 an extended game where the source discriminator uses the output of the label predictor as an approximation of the posterior probabilities, as shown in Figure 7-1(b). We prove that the equilibriums of the original game are also equilibriums in the extended one.

Encoder E: An encoder $E(\cdot) : \Omega_{\boldsymbol{x}} \to \Omega_{\boldsymbol{z}}$ is a function that takes a sequence of input samples $\boldsymbol{x}$, and returns a vector summary of $\boldsymbol{x}$ as $\boldsymbol{z} = E(\boldsymbol{x})$.

Label Predictor F: A label predictor $F(\cdot) : \Omega_{\boldsymbol{z}} \to [0,1]^{n_y}$ takes a latent representation $E(\boldsymbol{x})$ as input and predicts the probability of each label y associated with input $\boldsymbol{x}$ as $Q_F(y|E(\boldsymbol{x}))$. The goal of an ideal predictor F is to approximate $P_y(\cdot|\boldsymbol{x})$ with $Q_F(\cdot|E(\boldsymbol{x}))$.

The loss of the label predictor, F, given the encoder E, is defined as the cross-entropy between the label distribution $P_y(\cdot|\boldsymbol{x})$ and $Q_F(\cdot|E(\boldsymbol{x}))$:

$$\mathcal{L}_f(F;E) = \mathbb{E}_{\boldsymbol{x},y}[-\log Q_F(y|E(\boldsymbol{x}))] \tag{7.1}$$

During training, the encoder E and predictor F play a co-operative game to minimize the label prediction loss.

Source Discriminator D: We define a source discriminator as $D(\cdot,\cdot) : \Omega_{\boldsymbol{z}} \times [0,1]^{n_y} \to [0,1]^{n_s}$. It takes the latent representation $E(\boldsymbol{x})$ and the label distribution $P_y(\cdot|\boldsymbol{x})$ as inputs, and predicts which source domain (i.e., subject and environment) they are sampled from as $Q_D(\cdot|E(\boldsymbol{x}), P_y(\cdot|\boldsymbol{x}))$.

Next, we define the desired representation invariance.

Definition 7 (Representation invariance)**.** *We say that representation E is invariant if $E(\boldsymbol{x})$ contains no information about s beyond what is already contained in $P_y(\cdot|\boldsymbol{x})$; that is, $Q_D(\cdot|E(\boldsymbol{x}), P_y(\cdot|\boldsymbol{x})) = Q_D(\cdot|P_y(\cdot|\boldsymbol{x}))$ for the optimal D.*

To measure the invariance of an encoder E, we define the loss of the source discriminator D as the cross-entropy between $P_s(\cdot|\boldsymbol{x})$ and $Q_D(\cdot|E(\boldsymbol{x}), P_y(\cdot|\boldsymbol{x}))$:

$$\mathcal{L}_d(D; E) = \mathbb{E}_{\boldsymbol{x},s}[-\log Q_D(s|E(\boldsymbol{x}), P_y(\cdot|\boldsymbol{x}))] \tag{7.2}$$

During training, encoder E and discriminator D play a minimax game: while D is trained to minimize the source prediction loss, encoder E is trained to maximize it in order to achieve the above invariance.

■ 7.1.1 Ideal Game

During training, encoder E plays a co-operative game with predictor F, and a minimax game with discriminator D. We define a value function of E, F and D with $\lambda > 0$:

$$\mathcal{V}(E, F, D) = \mathcal{L}_f(F; E) - \lambda \cdot \mathcal{L}_d(D; E) \tag{7.3}$$

The training procedure can be viewed as a three-player minimax game of E, F and D:

$$\min_E \min_F \max_D \mathcal{V}(E, F, D) = \min_{E,F} \max_D \mathcal{V}(E, F, D) \tag{7.4}$$

Proposition 8 (Optimal predictor)**.** *Given encoder E,*

$$\mathcal{L}_f(E) \triangleq \min_F \mathcal{L}_f(F; E) \geq H(y|E(\boldsymbol{x})), \tag{7.5}$$

where $H(\cdot)$ is entropy.

The optimal predictor F^* that achieves equality is:

$$Q_{F^*}(y|E(\boldsymbol{x})) = p(y|E(\boldsymbol{x})) \tag{7.6}$$

Proof.

$$\begin{aligned}
&\mathcal{L}_f(F;E) \\
=&\mathbb{E}_{\boldsymbol{x},y}[-\log Q_F(y|E(\boldsymbol{x}))] \\
=&\mathbb{E}_{E(\boldsymbol{x}),y}[-\log Q_F(y|E(\boldsymbol{x}))] \\
=&\mathbb{E}_{\boldsymbol{z}\sim P(E(\boldsymbol{x}))}\,\mathbb{E}_{y\sim P(y|\boldsymbol{z})}[-\log Q_F(y|\boldsymbol{z})] \\
=&\mathbb{E}_{\boldsymbol{z}\sim P(E(\boldsymbol{x}))}[H(y|\boldsymbol{z}) + D_{\mathrm{KL}}(P(y|\boldsymbol{z}) \parallel Q_F(y|\boldsymbol{z}))] \\
\geq&\mathbb{E}_{\boldsymbol{z}\sim P(E(\boldsymbol{x}))}[H(y|\boldsymbol{z})] \\
=&H(y|E(\boldsymbol{x}))
\end{aligned}$$

The equality holds when $D_{\mathrm{KL}}(P(y|E(\boldsymbol{x})) \parallel Q_F(y|E(\boldsymbol{x}))) = 0$ for almost every $\boldsymbol{x} \in Supp(P_{\boldsymbol{x}})$. That is $Q_{F^*}(y|E(\boldsymbol{x})) = p(y|E(\boldsymbol{x}))$ for almost every y and $\boldsymbol{x} \in Supp(P_{\boldsymbol{x}})$. □

Similarly we can prove the following Proposition.

Proposition 9 (Optimal discriminator)**.** *Given encoder E,*

$$\mathcal{L}_d(E) \triangleq \min_D \mathcal{L}_d(D;E) \geq H(s|E(\boldsymbol{x}), P_y(\cdot|\boldsymbol{x})) \tag{7.7}$$

The optimal discriminator D^* that achieves this value is:

$$Q_{D^*}(s|E(\boldsymbol{x}), P_y(\cdot|\boldsymbol{x})) = P(s|E(\boldsymbol{x}), P_y(\cdot|\boldsymbol{x})) \tag{7.8}$$

Corollary 9.1. *$H(s)$ is an upper bound of the loss of the optimal discriminator D^* for any encoder E.*

Next, we state the virtual training criterion of the encoder.

Proposition 10. *If predictor F and discriminator D have enough capacity and are trained to achieve their optimal losses, the minimax game Equation 7.4 can be rewritten as the following training procedure of encoder E:*

$$\min_E[H(y|E(\boldsymbol{x})) - \lambda \cdot H(s|E(\boldsymbol{x}), P_y(\cdot|\boldsymbol{x}))] \tag{7.9}$$

Proof. Based on the losses of the optimal predictor F^* and the optimal discriminator D^* in Proposition 8 and Proposition 9, the minimax game Equation 7.4 can be rewritten as (7.9).

Thus, encoder E is trained to minimize a virtual training criterion $C(E) = H(y|E(\boldsymbol{x})) - \lambda \cdot H(s|E(\boldsymbol{x}), P_y(\cdot|\boldsymbol{x}))$. □

Next, we describe the optimal encoder.

Theorem 11 (Optimal encoder)**.** *If encoder E, predictor F and discriminator D have enough capacity and are trained to reach optimum, any global optimal encoder E^* has the following properties:*

$$H(y|E^*(\boldsymbol{x})) = H(y|\boldsymbol{x}) \tag{7.10a}$$

$$H(s|E^*(\boldsymbol{x}), P_y(\cdot|\boldsymbol{x})) = H(s|P_y(\cdot|\boldsymbol{x})) \tag{7.10b}$$

Proof. Since $E(\boldsymbol{x})$ is a function of $\boldsymbol{x}$:

$$\mathcal{L}_f(E) = H(y|E(\boldsymbol{x})) \geq H(y|\boldsymbol{x}) \tag{7.11a}$$

$$\mathcal{L}_d(E) = H(s|E(\boldsymbol{x}), P_y(\cdot|\boldsymbol{x})) \leq H(s|P_y(\cdot|\boldsymbol{x})) \tag{7.11b}$$

Hence, $C(E) = H(y|E(\boldsymbol{x})) - \lambda \cdot H(s|E(\boldsymbol{x}), P_y(\cdot|\boldsymbol{x})) \geq H(y|\boldsymbol{x}) - \lambda \cdot H(s|P_y(\cdot|\boldsymbol{x}))$. The equality holds if and only if both Equation 7.10a and Equation 7.10b are satisfied. Therefore, we only need to prove that the optimal value of $C(E)$ is equal to $H(y|\boldsymbol{x}) - \lambda \cdot H(s|P_y(\cdot|\boldsymbol{x}))$ in order to prove that any global encoder E^* satisfies both Equation 7.10a and Equation 7.10b.

We show that $C(E)$ can achieve $H(y|\boldsymbol{x}) - \lambda \cdot H(s|P_y(\cdot|\boldsymbol{x}))$ by considering the following encoder E_0: $E_0(\boldsymbol{x}) = P_y(\cdot|\boldsymbol{x})$. It can be examined that $H(y|E_0(\boldsymbol{x})) = H(y|\boldsymbol{x})$ and $H(s|E_0(\boldsymbol{x}), P_y(\cdot|\boldsymbol{x})) = H(s|P_y(\cdot|\boldsymbol{x}))$. □

Adversarial training of D can be viewed as a regularizer, which leads to a common representation space for multiple source domains. From Theorem 11, the optimal encoder E^* using adversarial training satisfies $H(y|E^*(\boldsymbol{x})) = H(y|\boldsymbol{x})$, which is the maximal discriminative capability that any encoder E can achieve. Thus, we have the following corollary.

Corollary 11.1. *Adversarial training of the discriminator does not reduce the discriminative capability of the representation.*

Remark 11.1. *During the proof of Theorem 11, we construct an encoder $E_0(\boldsymbol{x}) = P_y(\cdot|\boldsymbol{x})$ that can achieve the optimal value of $C(E)$. However, we argue that training will not converge to this* *trivial encoder in practice. This is because $P_y(\cdot|\boldsymbol{x})$ is a mapping from the full signal history to the distribution over stages at the current step, therefore itself highly complex. Since we use the RNN state as the encoding $E(\boldsymbol{x})$, and it feeds into the LSTM gates, distribution over stages at previous step does not represent a sufficient summary of the history until the current one. Therefore, $E(\boldsymbol{x})$ must be able to anticipate the temporal evolution of the signal and contain a more effective summary than $P_y(\cdot|\boldsymbol{x})$ would be.*

Corollary 11.2. *If encoder E and predictor F have enough capacity and are trained to reach optimum, the output of F is equal to $P_y(\cdot|\boldsymbol{x})$.*

Proof. When predictor F is optimal (Proposition 8), $Q_F(y|E(\boldsymbol{x})) = p(y|E(\boldsymbol{x}))$. When E is optimal (Theorem 11), $H(y|E(\boldsymbol{x})) = H(y|\boldsymbol{x})$, that is $p(y|E(\boldsymbol{x})) = p(y|\boldsymbol{x})$. Therefore, $Q_F(y|E(\boldsymbol{x})) = p(y|\boldsymbol{x})$. □

■ 7.1.2 Extended Game

In practice, estimating the posterior label distribution $P_y(\cdot|\boldsymbol{x})$ from labeled data is a non-trivial task. Fortunately however our predictor F and encoder E are playing a cooperative game to approximate this posterior label distribution $P_y(\cdot|\boldsymbol{x})$ with $Q_F(\cdot|E(\boldsymbol{x}))$. Therefore, we use $Q_F(\cdot|E(\boldsymbol{x}))$, the output of predictor F, as a proxy of $P_y(\cdot|\boldsymbol{x})$ and feed it as input to discriminator D (Figure 7-1(b)).

An extended three-player game arises: while encoder E still plays a cooperative game with predictor F and a minimax game with discriminator D, discriminator D depends strategically on predictor F but not vice versa. The dotted line in Fig. 7-1(b) illustrates this dependency.

The relationship between the ideal minimax game (Section 7.1.1) and the extended one is stated below.

Proposition 12. *If encoder E, predictor F and discriminator D have enough capacity, the solution that encompasses the optimal encoder, E^*, predictor, F^* and discriminator, D^*, in the ideal minimax game is also an equilibrium solution of the extended game.*

Proof. By Corollary 11.2, when encoder E and predictor F are optimal, $Q_F(\cdot|E(\boldsymbol{x}))$ is equal to $P_y(\cdot|\boldsymbol{x})$. Thus, the extended game becomes equivalent to the ideal game, and E^*, F^* and D^* is an equilibrium solution of both games. □

Algorithm 1 Encoder, predictor and discriminator training

```
Input: Labeled data $\{(\boldsymbol{x}_i, y_i, s_i)\}_{i=1}^{M}$, and learning rate $\eta$.
Compute stop criterion for inner loop: $\delta_d \leftarrow H(s)$
for number of training iterations do
    Sample a mini-batch of training data $\{(\boldsymbol{x}_i, y_i, s_i)\}_{i=1}^{m}$:
        $\mathcal{L}_f^i \leftarrow -\log Q_F(y_i|E(\boldsymbol{x}_i))$
        $\boldsymbol{w}_i \leftarrow Q_F(\cdot|E(\boldsymbol{x}_i))$ ▶ stop gradient along this link
        $\mathcal{L}_d^i \leftarrow -\log Q_D(s_i|E(\boldsymbol{x}_i), \boldsymbol{w}_i)$
        $\mathcal{V}^i = \mathcal{L}_f^i - \lambda \cdot \mathcal{L}_d^i$
    Update encoder $E$:
        $\theta_e \leftarrow \theta_e - \eta_e \nabla_{\theta_e} \frac{1}{m} \sum_{i=1}^{m} \mathcal{V}^i$
    Update predictor $F$:
        $\theta_f \leftarrow \theta_f - \eta_f \nabla_{\theta_f} \frac{1}{m} \sum_{i=1}^{m} \mathcal{V}^i$
    repeat
        Update discriminator $D$:
            $\theta_d \leftarrow \theta_d + \eta_d \nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \mathcal{V}^i$
    until $\frac{1}{m} \sum_{i=1}^{m} \mathcal{L}_d^i \leq \delta_d$
```

■ 7.1.3 Training Algorithm

We implement the extended three-player game with iterative updates of the players (Algorithm 1). Note that, since the output of the label predictor is a proxy of the underlying posterior, and since the source discriminator depends strategically on the predictor but not vice versa, the gradient does not back-propagate from the discriminator to the predictor (i.e., the dotted link in Figure 7-1(b)).

The number of training steps in the inner loop usually needs to be carefully chosen [107]. A large number of steps is computationally inefficient but a small one will cause the model to collapse. This is because the outer players, E and F, can be over-trained against a non-optimal inner player D, and they will try to maximize $\mathcal{L}_d$ at the cost of increasing $\mathcal{L}_f$. To prevent the model collapse phenomenon, we use an adaptive number of training steps in the inner loop and adjust it dynamically based on $\mathcal{L}_d$ (Algorithm 1). The idea is to use the upper bound in Corollary 9.1 as the stopping criterion for the inner loop.

■ 7.1.4 Discussion of the Model Benefits

While we described our model in the context of sleep staging, we believe the model can be applied more broadly. Our model is characterized by the 3-way game and the adversarial conditioning on the label distribution. This combination yields the following benefits: 1) It guarantees an equilibrium solution that fully preserves the ability to perform the predictive task while removing any distracting information specific to the source domains. Guarantees of this kind are particularly important in healthcare where the measurements are noisy and have a variety of dependencies that need to be controlled. 2) It allows to properly leverage the adversarial feedback even when the target labels are uncertain. For example, in the sleep staging problem, each 30-second window is given one label. Yet, many such windows include transitions between sleep stages, e.g., a transition from light to deep sleep. These transitions are gradual and hence the transition windows can be intrinsically different from both light and deep sleep. It would be desirable to have the learned representation capture the concept of transition and make it invariant to the source (see the results in Section 7.2.5). 3) It allows the conditioning to remain available for additional guiding of representations based on unlabeled data. The model can incorporate unlabeled data for either semi-supervised learning or transductive learning within a unified framework.

■ 7.2 Experiments

In this section, we empirically evaluate our model.

■ 7.2.1 RF-Sleep Dataset

RF-Sleep is a dataset of RF measurements during sleep with corresponding sleep stage labels. All studies that involve human subjects were approved by our IRB.

Study setup: The sleep studies are done in the bedroom of each subject. We install a radio device in the bedroom. It transmits RF signals and measure their reflections while the subject is sleeping alone in the bed.

Ground truth: During the study, each subject sleeps with an FDA-approved EEG-based sleep monitor [132], which collects 3-channel frontal EEG. The monitor labels every 30-second of sleep with the subject's sleep stage. This system has human-level comparable accuracy [132], and has already been used in several sleep studies[213, 214].

Size of dataset: The dataset collects 100 nights of sleep from 25 young healthy subjects (40% females). It contains over 90k 30-second epochs of RF measurements and their corre-

Table 7-2: **Sleep Stage Classification Accuracy and Kappa.**

Approach	*Accuracy*	κ
Tataraidze et al. [211]	0.635	0.49
Zaffaroni et al. [210]	0.641	0.45
Ours	**0.798**	**0.70**

sponding sleep stages provided by the EEG-based sleep monitor. Each epochs has one of four labels Awake, REM, Light Sleep (N1 or N2) and Deep Sleep (N3).

7.2.2 Parameterization

We parameterize encoder E, predictor F and discriminator D as neural networks. Encoder E is parameterized by a hybrid CNN-RNN model. We adapt a residual networks architecture [164] with 24 convolutional layers to extract features from each 30-second RF spectrogram, and an RNN with LSTM cell [200] that takes sequences of CNN features as input. Both predictor F and discriminator D are parameterized by networks with two fully-connected layers.

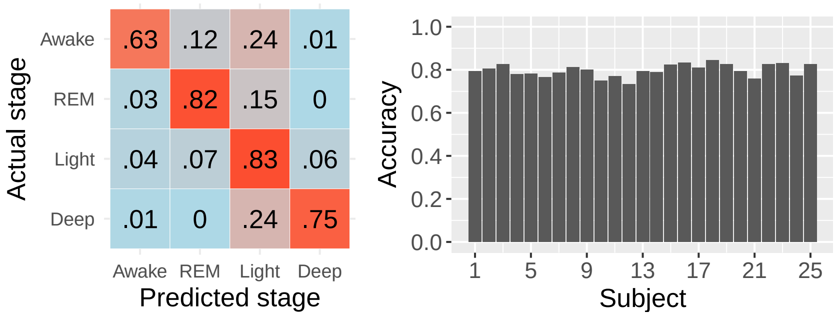
7.2.3 Classification Results

We evaluate the model on every subject while training on the data collected from the other subjects (i.e., the model is never trained on data from the test subject). The training data is randomly split into a training set and validation set ($75\%/25\%$).

We use two metrics commonly used in automated sleep staging, namely Accuracy and Cohen's Kappa. While accuracy measures the percent agreement with ground truth, Cohen's Kappa coefficient κ [125] takes into account the possibility of the agreement occurring by chance and is usually a more robust metric. $\kappa > 0.4$, $\kappa > 0.6$, $\kappa > 0.8$ are considered to be moderate, substantial and almost perfect agreement [215].

Table 7-2 shows the accuracy and Cohen's Kappa of our model compared to the state-of-the-art in classifying sleep stages using RF reflections. Since neither the dataset nor the code used in past papers is publicly available, we compare with their published results. We note however that the Cohen's Kappa provides some normalization since it accounts for the underlying uncertainty in the data. The table shows that our model has an accuracy of 79.8% and a $\kappa = 0.70$, which significantly outperforms past solutions.

Figure 7-2(a) shows the confusion matrix of our model. Figure 7-2(b) also shows the accuracy on each subject. It has a standard deviation of 2.9%, suggesting that our model is



(a) Confusion Matrix

(b) Accuracy on each subject

Figure 7-2: **Sleep Staging Accuracy.** 7-2(a) shows that our model can distinguish different sleep stages with high accuracy. And 7-2(b) illustrates that our model works well for different subjects and environments.

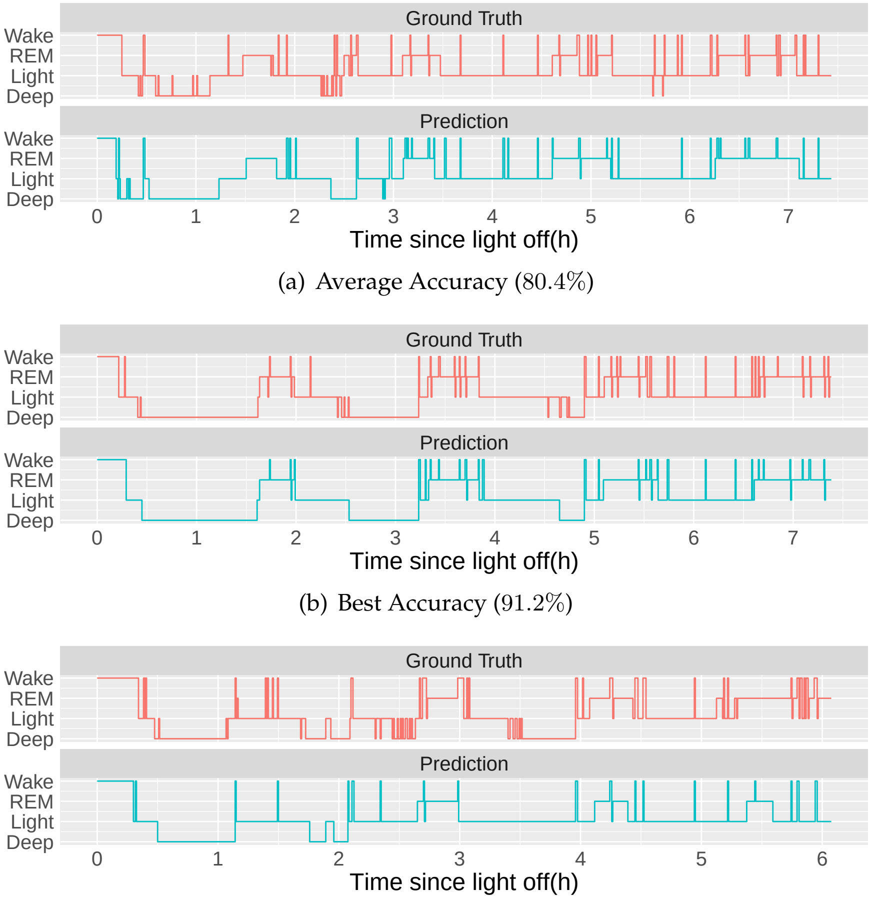
capable of adapting to different subjects and environments.

Finally, we show in Figure 7-3 the full-night predictions along with the ground truth for the average, best, and worst classification accuracy.

7.2.4 Understanding the Role of CNN & RNN

We analyze the role of CNN and RNN in predicting sleep stages. To do so, we use t-SNE embedding [216] to visualize the response of our network after CNN and RNN, respectively. Figure 7-4 shows the visualization results from one of the subjects. Data points are randomly sub-sampled for better viewing. The result shows that the CNN succeeds at separating the Wake, REM from Light and Deep Sleep. However it fails at separate Light Sleep and Deep Sleep from each other. In contrast, Light Sleep and Deep Sleep form different clusters in the RNN response. These results demonstrate the role of CNN and RNN in our model: CNN learns stage-specific features that can distinguish Wake, REM and from Deep and Light Sleep. RNN captures the dynamics of those features to further determine whether the sleep is light or deep. Note that Light and Deep Sleep are more similar to each other and are typically referred to as NREM, i.e., non-REM.

We have trained a similar model without the RNN layer on top of CNN. In this case, the overall accuracy decreases by 12.8%, specifically the precision light and deep sleep decreases by 23.5%. This suggests that there are stage-specific information embedded in the temporal dynamics of the RF measurements, and therefore can only be captured and exploited with RNN. Moreover, these temporal dynamics are particularly crucial for distinguishing light and deep sleep. Indeed, there are known temporal patterns that govern



(a) Average Accuracy (80.4%)

(b) Best Accuracy (91.2%)

(c) Worst Accuracy (71.2%)

Figure 7-3: **Example outputs of RF-Sleep.** Three examples of full night predictions corresponding to the average, best and worst classification accuracy.

the progression of light and deep sleep through the night [217]. For example, the probability of being in deep sleep decreases as sleep progresses. Also, people usually need to go through light sleep before they can get into deep sleep. These temporal dynamics of sleep stages can be captured by RNN and might be exploited to distinguish light and deep sleep.

■ 7.2.5 Role of Our Adversarial Discriminator

We evaluate the role of our adversarial discriminator in learning transferable features for predicting sleep stages. We first look at the losses on the validation set as training progresses to check whether the extraneous information specific to the individuals and environments can be removed. As a baseline, we compare with a version of our model without the source discriminator. For this baseline, we train a (non-adversarial) discriminator to

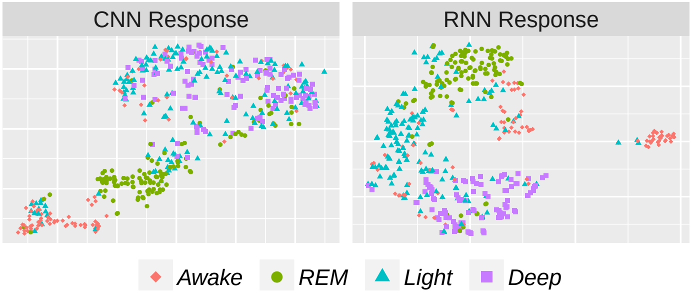


Figure 7-4: **Visualizations of the CNN and RNN responses.** CNN can separate Wake REM and from the other stages, yet Deep and Light Sleep can only be distinguished by RNN.

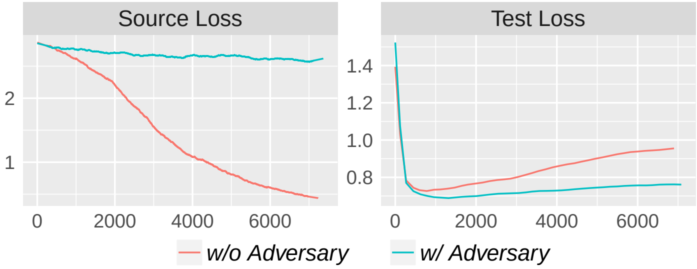


Figure 7-5: **Baseline model and ours are evaluated on same dataset.** A higher source loss indicates the removal of source specific information, and a lower test loss shows that the proposed setup can better avoid overfitting.

determine the source of features. Figure 7-5 shows that the loss of the source discriminator in the baseline model decreases very quickly while ours stays high (upper bounded by $H(s) = 2.81$ in this case), suggesting that our learned representation is invariant across sources. The figure also shows that adding an adversarial discriminator increases the performance on the test set and can be helpful in reducing over-fitting.

To check that our adversarial model has learned transferable features, we visualize the learned features $E(\boldsymbol{x})$ on the test data for both models. Color-coding the sources, Figure 7-6 shows that our learned features have almost the same distribution on different sources, while the baseline model learns features that are separable.

Next, we illustrate the benefits of conditioning on the posterior distribution, and that it can recover underlying concepts not specified in the labels. We consider the learned features for transition periods between light and deep sleep, which might be a class that is different from both light and deep sleep. We define transition periods as epochs that have both light and deep sleep as neighbors. We visualize it with a different color. Color-coding stages and shape-coding sources, Figure 7-7 shows the learned features from transition periods are segregated, as those from light sleep and deep sleep. This indicates that our

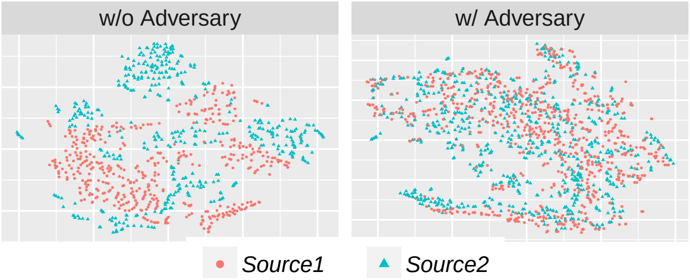


Figure 7-6: **Visualization of learned latent representations from two sources.** Datapoints are separated when no adversary, yet they are well aligned by proposed setup.

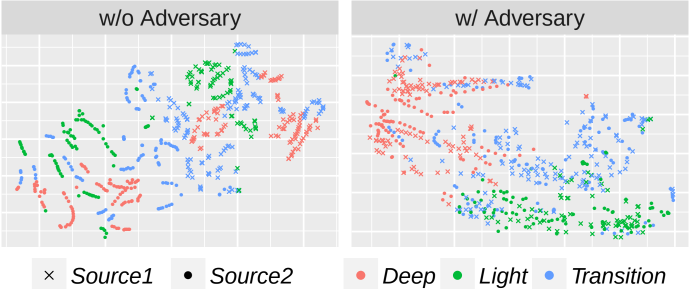


Figure 7-7: **Visualization of fine-grained alignment on test data.** Our model, which conditions the adversary on the posterior distribution, not only aligns deep and light stages, but also aligns the transition periods, which are not directly specified by the labels.

learned features have recovered the concept of a transition period, which is helpful in understanding and predicting sleep stages.

■ 7.3 Conclusion

We introduced a new predictive model that learns sleep stages from RF signals and achieves a significant improvement over the state-of-the-art. We believe this work marks an important step in sleep monitoring. We also believe that the proposed adversarial setup, which extracts task-specific domain-invariant features, is applicable to other predictive tasks, particularly in health sensing where variations across subjects and measurement conditions could be a major challenge.

CHAPTER 8
Emotion Recognition using Wireless Signals

Emotion recognition is an emerging field that has attracted much interest from both the industry and the research community [218, 219, 220, 221, 222]. It is motivated by a simple vision: Can we build machines that sense our emotions? If we can, such machines would enable smart homes that react to our moods and adjust the lighting or music accordingly. Movie makers would have better tools to evaluate user experience. Advertisers would learn customer reaction immediately. Computers would automatically detect symptoms of depression, anxiety, and bipolar disorder, allowing early response to such conditions. More broadly, machines would no longer be limited to explicit commands, and could interact with people in a manner more similar to how we interact with each other.

Existing approaches for inferring a person's emotions either rely on audiovisual cues, such as images and audio clips [143, 220, 223], or require the person to wear physiological sensors like an ECG monitor [53, 224, 145, 225]. Both approaches have their limitations. Audiovisual techniques leverage the outward expression of emotions, but cannot measure inner feelings [52, 224, 226]. For example, a person may be happy even if she is not smiling, or smiling even if she is not happy. Also, people differ widely in how expressive they are in showing their inner emotions, which further complicates this problem [227]. The second approach recognizes emotions by monitoring the physiological signals that change with our emotional state, e.g., our heartbeats. It uses on-body sensors – e.g., ECG monitors –

to measure these signals and correlate their changes with joy, anger, etc. This approach is more correlated with the person's inner feelings since it taps into the interaction between the autonomic nervous system and the heart rhythm [50, 228]. However, the use of body sensors is cumbersome and can interfere with user activity and emotions, making this approach unsuitable for regular usage.

In this section, we introduce a new method for emotion recognition that achieves the best of both worlds –i.e., it directly measures the interaction of emotions and physiological signals, but does not require the user to carry sensors on his body.

Our design uses RF signals to sense emotions. Specifically, RF signals reflect off the human body and get modulated with bodily movements. Recent research has shown that such RF reflections can be used to measure a person's breathing and average heart rate without body contact [68, 69, 77, 78, 79]. However, the periodicity of the heart signal (i.e., its running *average*) is of little relevance to emotion recognition. Specifically, to recognize emotions, we need to measure the minute variations in *each individual beat length* [50, 51, 52].

Yet, extracting individual heartbeats from RF signals incurs multiple challenges, which can be seen in Figure 8-1. First, RF signals reflected off a person's body are modulated by both breathing and heartbeats. The impact of breathing is typically orders of magnitude larger than that of heartbeats, and tends to mask the individual beats (see the top graph in Figure 8-1); to separate breathing from heart rate, past systems operate over multiple seconds (e.g., 30 seconds in [68]) in the frequency domain, forgoing the ability to measure the beat-to-beat variability. Second, heartbeats in the RF signal lack the sharp peaks which characterize the ECG signal, making it harder to accurately identify beat boundaries. Third, the difference in inter-beat-intervals (IBI) is only a few tens of milliseconds. Thus, individual beats have to be segmented to within a few milliseconds. Obtaining such accuracy is particularly difficult in the absence of sharp features that identify the beginning or end of a heartbeat. Our goal is to address these challenges to enable RF-based emotion recognition.

We present EQ-Radio, a wireless system that performs emotion recognition using RF reflections off a person's body. EQ-Radio's key enabler is a new algorithm for extracting individual heartbeats and their differences from RF signals. Our algorithm first mitigates the impact of breathing. The intuition underlying our mitigation mechanism is as follows:

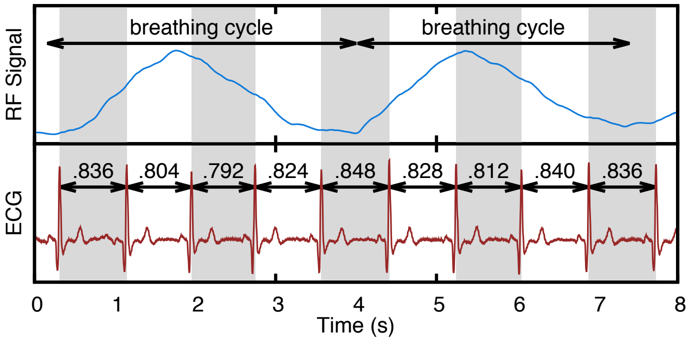


Figure 8-1: **Comparison of RF signal with ECG signal.** The top graph plots the RF signal reflected off a person's body. The envelop of the RF signal follows the inhale-exhale motion. The small dents in the signal are due to heartbeats. The bottom graph plots the ECG of the subject measured concurrently with the RF signal. Individual beats are marked by grey and white shades. The numbers report the beat-length in seconds. Note the small variations in consecutive beat lengths.

while chest displacement due to the inhale-exhale process is orders of magnitude larger than minute vibrations due to heartbeats, the acceleration of breathing is smaller than that of heartbeats. This is because breathing is usually slow and steady while a heartbeat involves rapid contraction of the muscles (which happen at localized instances in time). Hence, EQ-Radio operates on the acceleration of RF signals to dampen the breathing signal and emphasize the heartbeats.

Next, EQ-Radio needs to segment the RF reflection into individual heartbeats. In contrast to the ECG signal which has a known expected shape (see the bottom graph in Fig. 8-1), the shape of a heartbeat in RF reflections is unknown and varies depending on the person's body and exact posture with respect to the device. Thus, we cannot simply look for a known shape as we segment the signal; we need to learn the beat shape as we perform the segmentation. We formulate the problem as a joint optimization, where we iterate between two sub-problems: the first sub-problem learns a template of the heartbeat given a particular segmentation, while the second finds the segmentation that maximizes resemblance to the learned template. We keep iterating between the two sub-problems until we converge to the best beat template and the optimal segmentation that maximizes resemblance to the template. Finally, we note that our segmentation takes into account that beats can shrink and expand and hence vary in beat length. Thus, the algorithm finds the beat segmentation that maximizes the similarity in the morphology of a heartbeat signal across consecutive beats while allowing for flexible warping (shrinking or expansion) of the beat signal.

We have built EQ-Radio into a full-fledged emotion recognition system. EQ-Radio's system architecture has three components: The first component is an FMCW radio that transmits RF signals and receives their reflections. The radio leverages the approach in [68] to zoom in on human reflections and ignore reflections from other objects in the scene. Next, the resulting RF signal is passed to the beat extraction algorithm described above. The algorithm returns a series of signal segments that correspond to the individual heartbeats. Finally, the heartbeats – along with the captured breathing patterns from RF reflections – are passed to an emotion classification sub-system as if they were extracted from an ECG monitor. The emotion classification sub-system computes heartbeat-based and respiration-based features recommended in the literature [145, 52, 224] and uses an SVM classifier to differentiate among various emotional states.

We evaluate EQ-Radio by conducting user experiments with 30 subjects. We design our experiments in accordance with the literature in the field [145, 52, 224]. Specifically, the subject is asked to evoke a particular emotion by recalling a corresponding memory (e.g., sad or happy memories). She/he may use music or photos to help evoking the appropriate memory. In each experiment, the subject reports the emotion she/he felt, and the period during which she/he felt that emotion. During the experiment, the subject is monitored using both EQ-Radio and a commercial ECG monitor. Further, a video is taken of the subject then passed to the Microsoft image-based emotion recognition system [229].

Our experiments show that EQ-Radio's emotion recognition is on par with state-of-the-art ECG-based systems, which require on-body sensors [53]. Specifically, if the system is trained on each subject separately, the accuracy of emotion classification is 87% in EQ-Radio and 88.2% in the ECG-based system. If one classifier is used for all subjects, the accuracy is 72.3% in EQ-Radio and 73.2% in the ECG-based system.[1] For the same experiments, the accuracy of the image-based system is 39.5%; this is because the image-based system performed poorly when the emotion was not visible on the subject's face.

Our results also show that EQ-Radio's performance is due to its ability to accurately extract heartbeats from RF signals. Specifically, even errors of 40-50 milliseconds in estimating heartbeat intervals would reduce the emotion recognition accuracy to 44% (as we

[1] The ECG-based system and EQ-Radio use exactly the same classification features but differ in how they obtain the heartbeat series. In all experiments, training and testing are done on different data.

show in Figure 8-12 in Section 8.5.3). In contrast, our algorithm achieves an average error in inter-beat-intervals (IBI) of 3.2 milliseconds, which is less than 0.4% of the average beat length.

EQ-Radiomakes the following three contributions:

- To our knowledge, this is the first system that demonstrates the feasibility of emotion recognition using RF reflections off one's body. As such, this work both expands the scope of wireless systems and advances the field of emotion recognition.
- We introduce a new algorithm for extracting individual heartbeats from RF reflections off the human body. The algorithm presents a new mathematical formulation of the problem, and is shown to perform well in practice.
- We also present a user study of the accuracy of emotion recognition using RF reflections, and an empirical comparison with both ECG-based and image-based emotion recognition systems.

8.1 EQ-Radio Overview

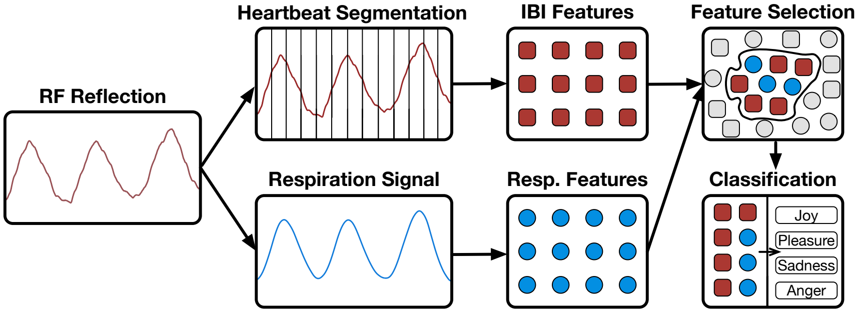


Figure 8-2: **EQ-Radio Architecture.** EQ-Radio has three components: a radio for capturing RF reflections (Section 8.2), a heartbeat extraction algorithm (Section 8.3), and a classification subsystem that maps the learned physiological signals to emotional states (Section 8.4).

EQ-Radio is an emotion recognition system that relies purely on wireless signals. It operates by transmitting an RF signal and capturing its reflections off a person's body. It then analyzes these reflections to infer the person's emotional state. It classifies the person's emotional state according to the known arousal-valence model into one of four basic emotions [230, 145]: anger, sadness, joy, and pleasure (i.e., contentment).

EQ-Radio's system architecture consists of three components that operate in a pipelined manner, as shown in Figure 8-2:

- *An FMCW radio,* which transmits RF signals and captures their reflections off a person's body.
- *A beat extraction algorithm,* which takes the captured reflections as input and returns a series of signal segments that correspond to the person's individual heartbeats.
- *An emotion-classification subsystem,* which computes emotion-relevant features from the captured physiological signals – i.e., the person's breathing pattern and heartbeats – and uses these features to recognize the person's emotional state.

In the following sections, we describe each of these components in detail.

8.2 Capturing the RF Signal

EQ-Radio operates on RF reflections off the human body. To capture such reflections, EQ-Radio uses a RADAR technique called Frequency Modulated Carrier Waves (FMCW) [8]. There is a significant literature on FMCW radios and their use for obtaining an RF signal that is modulated by breathing and heartbeats [68, 231, 232]. We refer the reader to [68] for a detailed description of such methods, and summarize below the basic information relevant to our emotion recognition system.

The radio transmits a low power signal and measures its reflection time. It separates RF reflections from different objects/bodies into buckets based on their reflection time. It then eliminates reflections from static objects which do not change across time and zooms in on human reflections. It focuses on time periods when the person is quasi-static. It then looks at the phase of the RF wave which is related to the traveled distance as follows [233]:

$$\phi(t) = 2\pi \frac{d(t)}{\lambda},$$

where $\phi(t)$ is the phase of the signal, λ is the wavelength, $d(t)$ is the traveled distance, and t is the time variable. The variations in the phase correspond to the compound displacement caused by chest expansion and contraction due to breathing, and body vibration due to heartbeats.[2]

[2] When blood is ejected from the heart, it exercises a force on the rest of the body causing small jitters in the head and skin, which are picked up by the RF signal [68].

The phase of the RF signal is illustrated in the top graph in Figure 8-1. The envelop shows the chest displacements as the inhale-exhale process. The small dents are due to minute skin vibrations associated with blood pulsing. EQ-Radio operates on this phase signal.

■ 8.3 Beat Extraction Algorithm

Recall that a person's emotions are correlated with small variations in her/his heartbeat intervals; hence, to recognize emotions, EQ-Radio needs to extract these intervals from the RF phase signal described above.

The main challenge in extracting heartbeat intervals is that the morphology of heartbeats in the reflected RF signals is unknown. Said differently, EQ-Radio does not know how these beats look like in the reflected RF signals. Specifically, these beats result in distance variations in the reflected signals, but the measured displacement depends on numerous factors including the person's body and her exact posture with respect to EQ-Radio's antennas. This is in contrast to ECG signals where the morphology of heartbeats has a known expected shape, and simple peak detection algorithms can extract the beat-to-beat intervals. However, because we do not know the morphology of these heartbeats in RF a priori, we cannot determine when a heartbeat starts and when it ends, and hence we cannot obtain the intervals of each beat. In essence, this becomes a chicken-and-egg problem: if we know the morphology of the heartbeat, that would help us in segmenting the signal; on the other hand, if we have a segmentation of the reflected signal, we can use it to recover the morphology of the human heartbeat.

This problem is exacerbated by two additional factors. First, the reflected signal is noisy; second, the chest displacement due to breathing is orders of magnitude higher than the heartbeat displacements. In other words, we are operating in a low SINR (signal-to-interference-and-noise) regime, where "interference" results from the chest displacement due to breathing.

To address these challenges, EQ-Radio first processes the RF signal to mitigate the interference from breathing. It then formulates and solves an optimization problem to recover the beat-to-beat intervals. The optimization formulation neither assumes nor relies on perfect separation of the respiration effect. In what follows, we describe both of these steps.

8.3.1 Mitigating the Impact of Breathing

The goal of the preprocessing step is to dampen the breathing signal and improve the signal-to-interference-and-noise ratio (SINR) of the heartbeat signal. Recall that the phase of the RF signal is proportional to the composite displacement due to the inhale-exhale process and the pulsing effect. Since displacements due to the inhale-exhale process are orders of magnitude larger than minute vibrations due to heartbeats, the RF phase signal is dominated by breathing. However, the acceleration of breathing is smaller than that of heartbeats. This is because breathing is usually slow and steady while a heartbeat involves rapid contraction of the muscles. Thus, we can dampen breathing and emphasize the heartbeats by operating on a signal proportional to acceleration as opposed to displacement.

By definition, acceleration is the second derivative of displacement. Thus, we can simply operate on the second derivative of the RF phase signal. Since we do not have an analytic expression of the RF signal, we have to use a numerical method to compute the second derivative. There are multiple such numerical methods which differ in their properties. We use the following second order differentiator because it is robust to noise [234]:

$$f_0'' = \frac{4f_0 + (f_1 + f_{-1}) - 2(f_2 + f_{-2}) - (f_3 + f_{-3})}{16h^2}, \tag{8.1}$$

where f_0'' refers to the second derivative at a particular sample, f_i refers to the value of the time series i samples away, and h is the time interval between consecutive samples.

In Figure 8-3, we show an example RF phase signal with the corresponding acceleration signal. The figure shows that in the RF phase, breathing is more pronounced than heartbeats. In contrast, in the acceleration signal, there is a periodic pattern corresponding to each heartbeat cycle, and the breathing effect is negligible.

8.3.2 Heartbeat Segmentation

Next, EQ-Radio needs to segment the acceleration signal into individual heartbeats. Recall that the key challenge is that we do not know the morphology of the heartbeat to bootstrap this segmentation process. To address this challenge, we formulate an optimization prob-

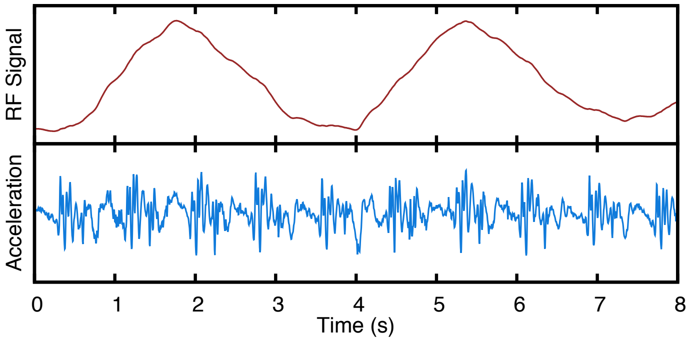


Figure 8-3: **RF Signal and Estimated Acceleration.** The figure shows the RF signal (top) and the acceleration of that signal (bottom). In the RF acceleration signal, the breathing motion is dampened and the heartbeat motion is emphasized. Note that while we can observe the periodicity of the heartbeat signal in the acceleration, delineating beat boundaries remains difficult because the signal is noisy and lacks sharp features.

lem that jointly recovers the morphology of the heartbeats and the segmentation.

The intuition underlying this optimization is that successive human heartbeats should have the same morphology; hence, while they may stretch or compress due to different beat lengths, they should have the same overall shape. This means that we need to find a segmentation that minimizes the differences in shape between the resulting beats, while accounting for the fact that we do not know a priori the shape of a beat and that the beats may compress or stretch. Further, rather than seeking locally optimal choices using a greedy algorithm, our formulation is an optimization problem over all possible segmentations, as described below.

Let $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ denote the sequence of length n. A segmentation $\mathcal{S} = \{s_1, s_2, ...\}$ of $\boldsymbol{x}$ is a partition of it into non-overlapping contiguous subsequences (segments), where each segment s_i consists of $|s_i|$ points.

In order to identify each heartbeat cycle, our idea is to find a segmentation with segments most similar to each other –i.e., to minimize the variation across segments. Since statistical variance is only defined for scalars or vectors with the same dimension, we extend the definition for vectors with different lengths as follows.

Definition 8.3.1. *Variance of segments* $\mathcal{S} = \{s_1, s_2, ...\}$ *is*

$$Var(\mathcal{S}) = \min_{\boldsymbol{\mu}} \sum_{s_i \in \mathcal{S}} \|s_i - \omega(\boldsymbol{\mu}, |s_i|)\|^2, \tag{8.2}$$

where $\omega(\boldsymbol{\mu}, |s_i|)$ is linear warping[3] *of $\boldsymbol{\mu}$ into length $|s_i|$.*

Note that the above definition is exactly the same as statistical variance when all the segments have the same length. $\boldsymbol{\mu}$ in the definition above represents the central tendency of all the segments –i.e., a template for the beat shape (or morphology).

The goal of our algorithm is to find the optimal segmentation $\mathcal{S}^*$ that minimizes the variance of segments, which can be formally stated as follows:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathrm{Var}(\mathcal{S}). \tag{8.3}$$

We can rewrite it as the following optimization problem:

$$\begin{aligned} \underset{\mathcal{S},\boldsymbol{\mu}}{\text{minimize}} \quad & \sum_{s_i \in \mathcal{S}} \|s_i - \omega(\boldsymbol{\mu}, |s_i|)\|^2, \\ \text{subject to} \quad & b_{\text{min}} \leq |s_i| \leq b_{\text{max}}, \; s_i \in \mathcal{S}, \end{aligned} \tag{8.4}$$

where b_{min} and b_{max} are constraints on the length of each heartbeat cycle.[4] It is trying to find the optimal segmentation $\mathcal{S}$ and template (i.e., morphology) $\boldsymbol{\mu}$ that minimize the sum of the square differences between segments and template. This optimization problem is difficult as it involves both combinatorial optimization over $\mathcal{S}$ and numerical optimization over $\boldsymbol{\mu}$. Exhaustively searching all possible segmentations has exponential complexity.

8.3.3 Algorithm

Instead of estimating the segmentation $\mathcal{S}$ and the template $\boldsymbol{\mu}$ simultaneously, our algorithm alternates between updating the segmentation and template, while fixing the other. During each iteration, our algorithm updates the segmentation given the current template, then updates the template given the new segmentation. For each of these two subproblems, our algorithms can obtain the global optimal with linear time complexity.

Update segmentation $\mathcal{S}$. In the l-th iteration, segmentation $\mathcal{S}^{l+1}$ is updated given template $\boldsymbol{\mu}^l$ as follows:

$$\mathcal{S}^{l+1} = \arg\min_{\mathcal{S}} \sum_{s_i \in \mathcal{S}} \|s_i - \omega(\boldsymbol{\mu}^l, |s_i|)\|^2. \tag{8.5}$$

[3] Linear warping is realized through a cubic spline interpolation [235].

[4] b_{min} and b_{max} capture the fact that human heartbeats cannot be indefinitely short or long. The default setting of b_{min} and b_{max} is 0.5s and 1.2s respectively.

Though the number of possible segmentations grows exponentially with the length of $\boldsymbol{x}$, the above optimization problem can be solved efficiently using dynamic programming. The recursive relationship for the dynamic program is as follows: if D_t denotes the minimal cost of segmenting sequence $\boldsymbol{x}_{1:t}$, then:

$$D_t = \min_{\tau \in \tau_{t,\mathbf{B}}} \{D_\tau + \|\boldsymbol{x}_{\tau+1:t} - \omega(\boldsymbol{\mu}, t-\tau)\|^2\}, \tag{8.6}$$

where $\tau_{t,\mathbf{B}}$ specifies possible choices of τ based on segment length constraints. The time complexity of the dynamic program based on Eqn. 8.6 is $O(n)$ and the global optimum is guaranteed.

Update template $\boldsymbol{\mu}$. In the l-th iteration, template $\boldsymbol{\mu}^{l+1}$ is updated given segmentation $\mathcal{S}^{l+1}$ as follows:

$$\begin{aligned} \boldsymbol{\mu}^{l+1} &= \arg\min_{\boldsymbol{\mu}} \sum_{s_i \in \mathcal{S}^{l+1}} \|s_i - \omega(\boldsymbol{\mu}, |s_i|)\|^2 \\ &= \arg\min_{\boldsymbol{\mu}} \sum_{s_i \in \mathcal{S}^{l+1}} |s_i| \cdot \|\boldsymbol{\mu} - \omega(s_i, m)\|^2 \end{aligned} \tag{8.7}$$

where m is the required length of template. The above optimization problem is a weighted least squares with the following closed-form solution:

$$\boldsymbol{\mu}^{l+1} = \frac{\sum_{s_i \in \mathcal{S}^{l+1}} |s_i| \omega(s_i, m)}{\sum_{s_i \in \mathcal{S}^{l+1}} |s_i|} = \frac{1}{n} \sum_{s_i \in \mathcal{S}^{l+1}} |s_i| \omega(s_i, m) \tag{8.8}$$

Figure 8-4 shows the final beat segmentation for the data in Figure 8-3. The figure also shows the ECG data of the subject. The segmented beat length matches the ECG of the subject to within a few milliseconds. There is a small delay since the ECG measures the electric signal of the heart, whereas the RF signal captures the heart's mechanical motion as it reacts to the electric signal [236].

Initialization. Initialization is typically important for optimization algorithms; however, we found that our algorithm does not require sophisticated initialization. Our algorithm can converge quickly with both random initialization and zero initialization. We choose to initialize the template $\boldsymbol{\mu}^0$ as the zero vector.

Running time analysis. The pseudocode of our algorithm is presented in 2. The complexity of this algorithm is $O(kn)$, where k is the number of iterations the algorithm takes be-

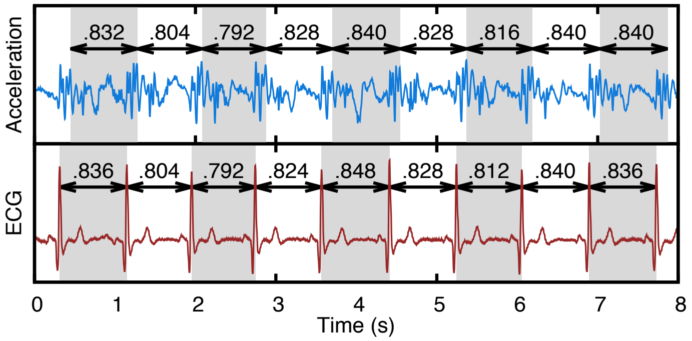


Figure 8-4: **Segmentation Result Compared to ECG.** The figure shows that the length of our segmented beats in RF (top) is very similar to the length of the segmented beats in ECG (bottom). There is a small delay since the ECG measures the electric signal of the heart, whereas the RF signal captures the heart's mechanical motion as it reacts to the electric signal.

Algorithm 2 Heartbeat Segmentation Algorithm

Input: Sequence $\boldsymbol{x}$ of n points, heart rate range $\mathbf{B}$.
Output: Segments $\mathcal{S}$, template $\boldsymbol{\mu}$ of length m.

```
Initialize $\boldsymbol{\mu}^0$ as zero vector
$l \leftarrow 0$                                                ▷ number of iterations
repeat
    $\mathcal{S}^{l+1} \leftarrow$ UPDATESEGMENTATION$(\boldsymbol{x}, \boldsymbol{\mu}^l)$
    $\boldsymbol{\mu}^{l+1} \leftarrow$ UPDATETEMPLATE$(\boldsymbol{x}, \mathcal{S}^{l+1})$
    $l \leftarrow l + 1$
until convergence
return $\mathcal{S}^l$ and $\boldsymbol{\mu}^l$

procedure UPDATESEGMENTATION$(\boldsymbol{x}, \boldsymbol{\mu})$
    $\mathcal{S}_0 \leftarrow \emptyset$
    $D_0 \leftarrow 0$
    for $t \leftarrow 1$ to $n$ do
        $\tau^* \leftarrow \arg\min_{\tau \in \mathcal{T}_{t,\mathbf{B}}} \{D_\tau + \|\boldsymbol{x}_{\tau+1:t} - \omega(\boldsymbol{\mu}, t-\tau)\|^2\}$
        $D_t \leftarrow D_{\tau^*} + \|\boldsymbol{x}_{\tau^*+1:t} - \omega(\boldsymbol{\mu}, t-\tau)\|^2$
        $\mathcal{S}_t \leftarrow \mathcal{S}_{\tau^*} \cup \{\boldsymbol{x}_{\tau^*+1:t}\}$
    return $\mathcal{S}_n$

procedure UPDATETEMPLATE$(\boldsymbol{x}, \mathcal{S})$
    $\boldsymbol{\mu} \leftarrow \frac{1}{n} \sum_{s_i \in \mathcal{S}} |s_i| \omega(s_i, m)$
    return $\boldsymbol{\mu}$
```

fore it converges. The algorithm is guaranteed to converge because the number of possible segmentations is finite and the cost function monotonically decreases with each iteration before it converges. In practice, this algorithm converges very quickly: for the evaluation experiments reported in Section 8.5, the number of iteration k is on average 8 and at most 16.

Finally, we note that the overall algorithm is not guaranteed to achieve a global optimum, but each of the subproblems achieves its global optimum. In particular, as detailed above, the first subproblem has a closed form optimal solution, and the second subproblem can be solved optimally with a dynamic program. As a result, the algorithm converges to a local optimum that works very well in practice as we show in Section 8.5.2.

■ 8.4 Emotion Classification

After EQ-Radio recovers individual heartbeats from RF reflections, it uses the heartbeat sequence along with the breathing signal to recognize the person's emotions. Below, we describe the emotion model which EQ-Radio adopts, and we elaborate on its approach for feature extraction and classification.

(a) 2D Emotion Model: EQ-Radio adopts a 2D emotion model whose axes are *valence* and *arousal*; this model serves as the most common approach for categorizing human emotions in past literature [230, 145]. The model classifies between four basic emotional states: Sadness (negative valence and negative arousal), Anger (negative valence and positive arousal), Pleasure (positive valence and negative arousal), and Joy (positive valence and positive arousal).

(b) Feature Extraction: EQ-Radio extracts features from both the heartbeat sequence and the respiration signal. There is a large literature on extracting emotion-dependent features from human heartbeats [145, 224, 237], where past techniques use on-body sensors. These features can be divided into time-domain analysis, frequency-domain analysis, time-frequency analysis, Poincaré plot [238], Sample Entropy [239], and Detrend Fluctuation Analysis [240]. EQ-Radio extracts 27 features from IBI sequences as listed in Table 8-1. These particular features were chosen in accordance with the results in [145]. We refer the reader to [145, 237] for a detailed explanation of these features.

EQ-Radio also employs respiration features. To extract the irregularity of breathing, EQ-Radio first identifies each breathing cycle by peak detection after low pass filtering. Since past work that studies breathing features recommends time-domain features [224], EQ-Radio extracts the time-domain features in the first row of Table 8-1.

(c) Handling Dependence: Physiological features differ from one subject to another for the same emotional state. Further, those features could be different for the same subject on different days. This is caused by multiple factors, including caffeine intake, sleep, and baseline mood of the day.

In order to extract better features that are user-independent and day-independent, EQ-Radio incorporates a baseline emotional state: neutral. The idea is to leverage changes of physiological features instead of absolute values. Thus, EQ-Radio calibrates the computed features by subtracting for each feature its corresponding values calculated at the neutral state for a given person on a given day.

(d) Feature Selection and Classification: As mentioned earlier, the literature has many features that relate IBI to emotions. Using all of those features with a limited amount of training data can lead to over-fitting. Selecting a set of features that is most relevant to emotions not only reduces the amount of data needed for training but also improves the classification accuracy on the test data.

Previous work on feature selection [224, 145] uses wrapper methods which treat the feature selection problem as a search problem. However, since the number of choices is exponentially large, wrapper methods have to use heuristics to search among all possible subsets of relevant features. Instead, EQ-Radio uses another class of feature selection mechanisms, namely embedded methods [241]; this approach allows us to learn which features best contribute to the accuracy of the model while training the model. To do this, EQ-Radio uses l_1-SVM [242] which selects a subset of relevant features while training an SVM classifier. Table 8-1 shows the selected IBI and respiration features in bold and italic respectively. The performance of the resulting classifier is evaluated in Section 8.5.3.

Table 8-1: **Features used in EQ-Radio.**

Domain	Name
Time	Mean, Median, SDNN,**pNN50**, RMSSD, SDNNi, meanRate, *sdRate*, HRVTi, *TINN*.
Frequency	Welch PSD: **LF/HF**, peakLF, peakHF. Burg PSD: **LF/HF**, peakLF, peakHF. Lomb-Scargle PSD: **LF/HF**, peakLF, peakHF.
Poincaré	SD_1, $\mathbf{SD}_2$, $\mathbf{SD}_2/\mathbf{SD}_1$.
Nonlinear	$\mathbf{SampEn}_1$, $\mathbf{SampEn}_2$, $\mathbf{DFA}_{all}$, DFA_1, DFA_2.

selected IBI features in **bold**;
selected respiration features in *italic*.

8.5 Evaluation

In this section, we describe our implementation of EQ-Radio and its empirical performance with respect to extracting individual heartbeats and recognizing human emotional states. All experiments were approved by our IRB.

8.5.1 Implementation

We reproduced a state-of-the-art FMCW radio designed by past work on wireless vital sign monitoring [68]. The device generates a signal that sweeps from 5.46 GHz to 7.25 GHz every 4 milliseconds, transmitting sub-mW power. The parameters were chosen as in [68] such that the transmission system is compliant with FCC regulations for consumer electronics. The FMCW radio connects to a computer over Ethernet. The received signal is sampled (digitized) and transmitted over the Ethernet to the computer. EQ-Radio's algorithms are implemented on an Ubuntu 14.04 computer with an i7 processor and 32 GB of RAM.

8.5.2 Evaluation of Heartbeat Extraction

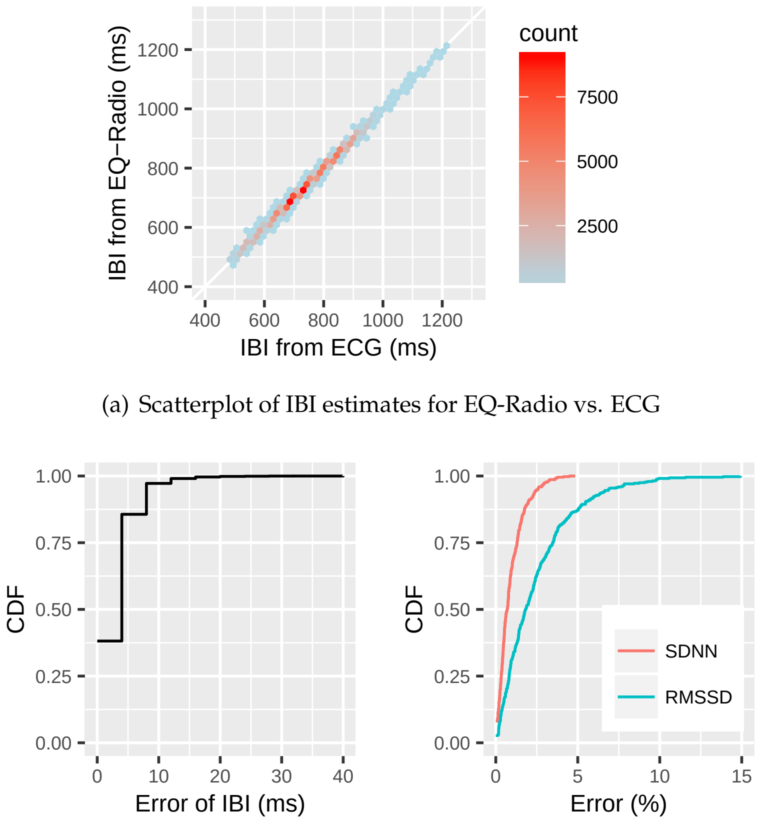
First, we would like to assess the accuracy of EQ-Radio's segmentation algorithm in extracting heartbeats from RF signals reflected off a subject's body.

Experimental Setup

Participants: We recruited 30 participants (10 females). Our subjects are between 19$\sim$77 years old. During the experiments, the subjects wore their daily attire with different fabrics.

Experimental Environment: We perform our experiments in 5 different rooms in a standard office building. The evaluation environment contains office furniture including desks, chairs, couches, and computers. The experiments are performed while other users are present in the room. The change in the experimental environment and the presence of other users had a negligible impact on the results because the FMCW radio described in Section 8.2 eliminates reflections from static objects (e.g., furniture) and isolates reflections from different humans [68].

Metrics: To evaluate EQ-Radio's heartbeat extraction algorithm, we use metrics that are common in emotion recognition:



(a) Scatterplot of IBI estimates for EQ-Radio vs. ECG

(b) CDF of error in IBI

(c) Error in emotion-related features

Figure 8-5: **Comparison of IBI Estimates Using EQ-Radio and a Commercial ECG Monitor.** The figure shows various metrics for evaluating EQ-Radio's heartbeat segmentation accuracy in comparison with an FDA-approved ECG monitor. Note that the CDF in (b) jumps at 4 ms intervals because the RF signal was sampled every 4 ms.

- *Inter-Beat-Interval (IBI):* The IBI measures the accuracy in identifying the boundaries of each individual beat.
- *Root Mean Square of Successive Differences (RMSSD):* This metric focuses on differences between successive beats. It is computed as $RMSSD = \sqrt{1/n \sum (IBI_{i+1} - IBI_i)^2}$, where n is the number of beats in the sum and i is a beat index. RMSSD is typically used as a measure of the parasympathetic nervous activity that controls the heart [243]. We calculate RMSSD for IBI sequences in a window of 2 minutes.
- *Standard Deviation of NN Intervals (SDNN):* The term NN-interval refers to the inter-beat-interval (IBI). Thus, SDNN measures the standard deviation of the beat length over a window of time. We use a window of 2 minutes.

Baseline: We obtain the ground truth for the above metrics using a commercial ECG monitor. We use the AD8232 evaluation board with a 3-lead ECG monitor to get the ECG signal. The synchronization between the FMCW signal and the ECG signal is accomplished by connecting both devices to a shared clock.

Accuracy in comparison to ECG We run experiments with 30 participants, collecting over 130,000 heart beats. Each subject is simultaneously monitored with EQ-Radio and the ECG device. We process the data to extract the above three metrics.

We first compare the IBIs estimated by EQ-Radio to the IBIs obtained from the ECG monitor. Figure 8-5(a) shows a scatter plot where the x and y coordinates are the IBIs derived from EQ-Radio and the ECG respectively. The color indicates the density of points in a specific region. Points on the diagonal have identical IBIs in EQ-Radio and ECG, while the distance to the diagonal is proportional to the error. It can be visually observed that all points are clustered around the diagonal, and hence EQ-Radio can estimate IBIs accurately irrespective of the their lengths.

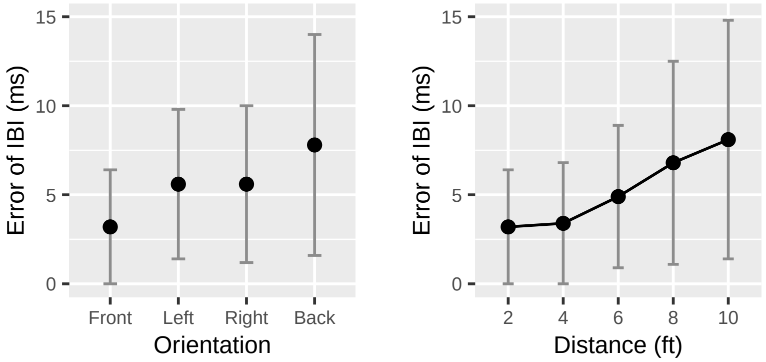
We quantitatively evaluate the errors in Figure 8-5(b), which shows a cumulative distribution function (CDF) of the difference between EQ-Radio's IBI estimate and the ECG-based IBI estimate for each beat. The CDF has jumps at 4ms intervals because the RF signal was sampled every 4ms.[5] The CDF shows that the 97th percentile error is 8ms. Our results further show that EQ-Radio's mean IBI estimation error is 3.2 ms. Since the average IBI in our experiments is 740 ms, on average, EQ-Radio estimates a beat length to within 0.43% of its correct value.

In Figure 8-5(c), we report results for beat variation metrics that are typically used in emotion recognition. The figure shows the CDF of errors in recovering the SDNN and RMSSD from RF reflections in comparison to contact-based ECG sensors. The plots show that the median error for each of these metrics is less than 2% and that even the 90^{th} percentile error is less than 8%. The high accuracy of these emotion-related metrics suggests that EQ-Radio's emotion recognition accuracy will be on par with contact-based techniques, as we indeed show in Section 8.5.3.

Accuracy for different orientations & distances

In the above experiments, the subject sat relatively close to EQ-Radio, at a distance of 3 to 4 feet, and was facing the device. It is desirable, however, to allow emotion recognition

[5] The actual sampling rate of our receiver is 1MHz. However, because each FMCW sweep takes 4ms, we obtain one phase measurement every 4ms. For a detailed explanation, please refer to [68].



(a) Error in IBI vs. orientation (b) Error in IBI vs. distance

Figure 8-6: **Error in IBI with Different Orientations and Distances.** (a) plots the error in IBI as a function of the user's orientation with respect to the device. (b) plots the error in IBI as a function of the distance between the user and the device.

even when the subject is further away or is not facing the device.

Thus, we evaluate EQ-Radio's beat segmentation accuracy as a function of orientation and distance. First, we fix the distance to 3 feet and repeat the above experiments for four different orientations: subject faces the device, subject has his back to the device, and the subject is facing left or right (perpendicular) to the device. We plot the median and standard deviation of EQ-Radio's IBI estimate for these four orientations in Figure 8-6(a). The figure shows that, across all orientations, the median error remains below 8ms (i.e., 1% of the beat length). As expected, however, the accuracy is highest when the user directly faces the device.

Next, we test EQ-Radio's beat segmentation accuracy as a function of its distance to the subject. We run experiments where the subject sits on a chair at different distances from the device. Figure 8-6(b) shows the median and standard deviation error in IBI estimate as a function of distance. Even at 10 feet, the median error is less than 8 ms (i.e., 1% of the beat length).

■ 8.5.3 Evaluation of Emotion Recognition

In this section, we investigate whether EQ-Radio can accurately classify a person's emotions based on RF reflections off her/his body. We also compare EQ-Radio's performance with more traditional emotion classification methods that rely on ECG signals or images.

Experimental Setup

Participants: We recruited 12 participants (6 females). Among them, 6 participants (3 females) have acting experience of 3∼7 years. People with acting experience are more skilled in emotion management, which helps in gathering high-quality emotion data and providing a reference group [224]. All subjects were compensated for their participation, and all experiments were approved by our IRB.

Experiment design: Obtaining high-quality data for emotion analysis is difficult, especially in terms of identifying the ground truth emotion [224]. Thus, it is crucial to design experiments carefully. We designed our experiments in accordance with previous work on emotion recognition using physiological signals [145, 224]. Specifically, before the experiment, the subjects individually prepare stimuli (e.g., personal memories, music, photos, and videos); during the experiment, the subject sits alone in one out of the 5 conference rooms and elicits a certain emotional state using the prepared stimuli. Some of these emotions are associated with small movements like laughing, crying, smiling, etc.[6] After the experiment, the subject reports the period during which she/he felt that type of emotion. Data collected during the corresponding period are labeled with the subject's reported emotion.

Throughout these experiments, each subject is monitored using three systems: 1) EQ-Radio, 2) the AD8232 ECG monitor, and 3) a video camera focused on the subject's face.

Ground Truth: As described above, subjects are instructed to evoke a particular emotion and report the period during which they felt that emotion. The subject's reported emotion is used to label the data from the corresponding period. These labels provide the ground truth for classification.

Baselines: We compare EQ-Radio's emotion classification to more traditional emotion recognition approaches based on ECG signals and image analysis. We describe the details of these systems in the corresponding sub-sections.

Metrics & Visualization: When tested on a particular data point, the classifier outputs a score for each of the considered emotional states. The data point is assigned the emotion that corresponds to the highest score. We measure *classification accuracy* as the percent of test data that is assigned the correct emotion.

[6] We note that the differentiation filter described in Section 8.3.1 mitigates such small movements. However, it cannot deal with larger body movements like walking. Though the FMCW radio we used can isolate signals from different users, as we show in Section 8.5.2, for better elicitation of emotional state, there is no other user in the room during this experiment.

We visualize the output of the classification as follows: Recall that the four emotions in our system can be represented in a 2D plane whose axes are *valence* and *arousal*. Each emotion occupies one of the four quadrants: Sadness (negative valence and negative arousal), Anger (negative valence and positive arousal), Pleasure (positive valence and negative arousal), and Joy (positive valence and positive arousal). Thus, we can visualize the classification result for a particular test data by showing it in the 2D valence-arousal space. If the point is classified correctly, it would fall in the correct quadrant.

For any data point, we can calculate the valence and arousal scores as:

$$S_{\text{valence}} = \max(S_{\text{joy}}, S_{\text{pleasure}}) - \max(S_{\text{sadness}}, S_{\text{anger}}),$$

$$S_{\text{arousal}} = \max(S_{\text{joy}}, S_{\text{anger}}) - \max(S_{\text{pleasure}}, S_{\text{sadness}}),$$

where S_{joy}, S_{pleasure}, S_{sadness}, and S_{anger} are the classification score output by the classifier for the four emotions. For example, consider a data point with the following scores $S_{\text{joy}} = 1$, $S_{\text{pleasure}} = 0$, $S_{\text{sadness}} = 0$, and $S_{\text{anger}} = 0$ –i.e., this data point is one unit of pure joy. Such data point falls on the diagonal in the upper right quadrant. A data point that has a high joy score but small scores for other emotions would still fall in the joy quadrant, but not exactly on the diagonal. (Check Figure 8-8 for an example.)

EQ-Radio's emotion recognition accuracy

To evaluate EQ-Radio's emotion classification accuracy, we collect 400 two-minute signal sequences from 12 subjects, 100 sequences for each emotion. We train two types of emotion classifiers: a person-dependent classifier, and a person-independent classifier. Each person-dependent classifier is trained and tested on data from a particular subject. Training and testing are done on mutually-exclusive data points using leave-one-out cross validation [244]. As for the person-independent classifier, it is trained on 11 subjects and tested on the remaining subject, and the process is repeated for different test subjects.

We first report the person-dependent classification results. Using the valence and arousal scores as coordinates, we visualize the results of person-dependent classification in Figure 8-7. Different types of points indicate the label of the data. We observe that emotions are well clustered and segregated, suggesting that these emotions are distinctly encoded in valence and arousal, and can be decoded from features captured by EQ-Radio. We also observe that the points tend to cluster along the diagonal and anti-diagonal, show-

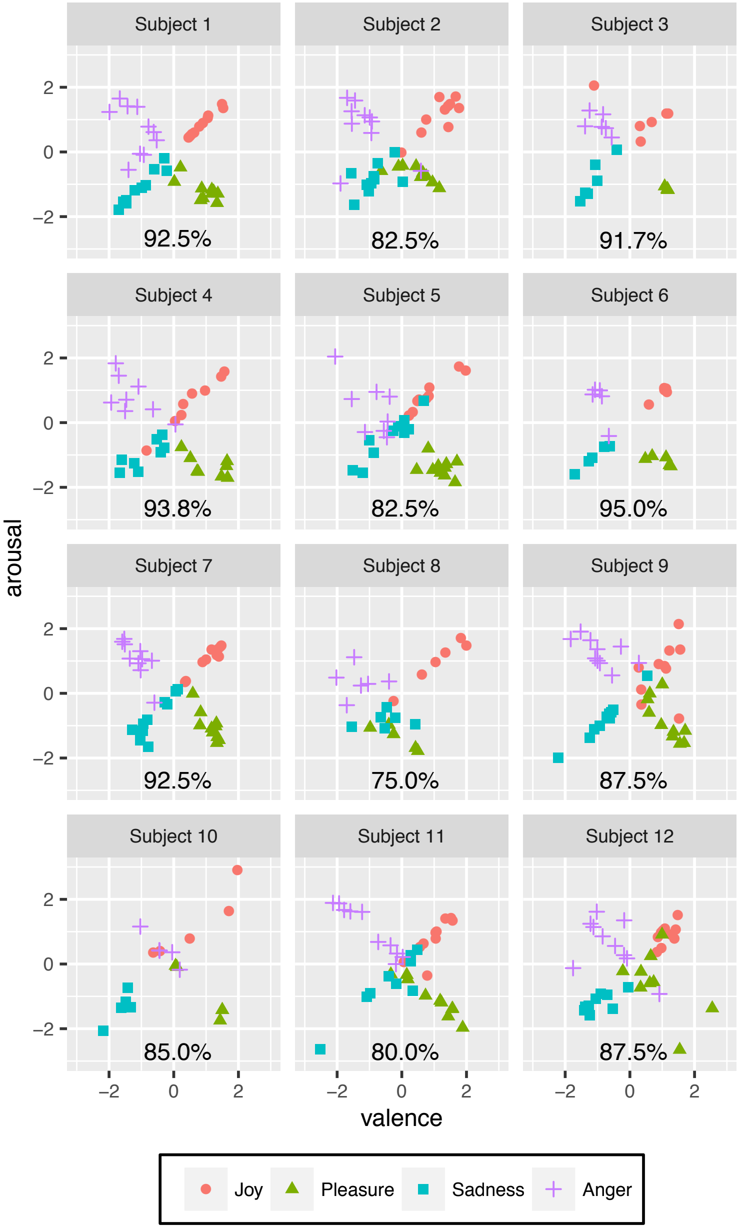


Figure 8-7: **Visualization of EQ-Radio's Person-dependent Classification Results.** The figure shows the person-dependent emotion-classification results for each of our 12 subjects. The x-axis in each of the scatter plots corresponds to the valence, and the y-axis corresponds to the arousal. For each data point, the label is our ground truth, and the coordinate is the classification result. At the bottom of each sub-figure, we show the classification accuracy for the corresponding subject.

ing that our classifiers have high confidence in the predictions. Finally, the accuracy of person-dependent classification for each subject is also shown in the figure with an overall average accuracy of 87.0%.

The results of person-independent emotion classification are visualized in Figure 8-8.

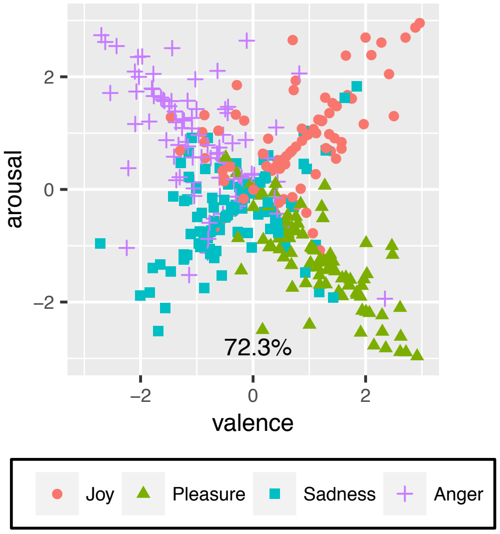


Figure 8-8: **Visualization of EQ-Radio's Person-independent Classification Results.** The figure shows the results of person-independent emotion-classification. The x-axis corresponds to valence, and the y-axis corresponds to arousal.

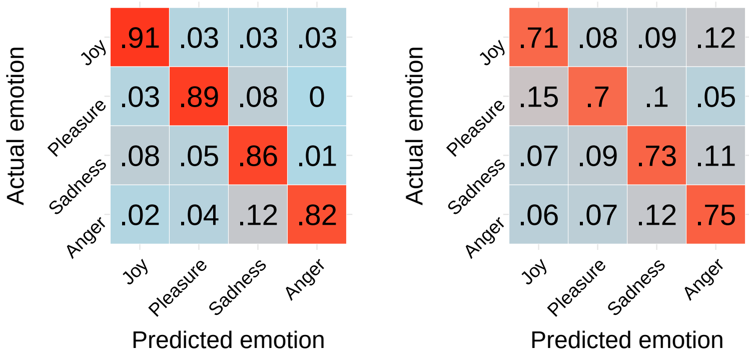
EQ-Radio is capable of recognizing a subject's emotion with an average accuracy of 72.3% purely based on data from other subjects, meaning that EQ-Radio succeeds in learning person-independent features for emotion recognition.

As expected, the accuracy of person-independent classification is lower than the accuracy of person-dependent classification. This is because person-independent emotion recognition is intrinsically more challenging since an emotional state is a rather subjective conscious experience that could be very different among different subjects. We note, however, that our accuracy results are consistent with the literature both for the case of person-dependent and person-independent emotion classifications [53]. Further, our results present the first demonstration of RF-based emotion classification.

To better understand the classification errors, we show the confusion matrix of both person-dependent and person-independent classification results in Figure 8-9. We find that EQ-Radio achieves comparable accuracy in recognizing the four types of emotions. We also observe that EQ-Radio typically makes fewer errors between emotion pairs that are different in both valence and arousal (i.e., joy vs. sadness and pleasure vs. anger).

Emotion recognition accuracy versus data source

It is widely known that gathering data that genuinely corresponds to a particular emotional state is crucial to recognizing emotions and that people with acting experience are



Actual \ Predicted	Joy	Pleasure	Sadness	Anger
Joy	.91	.03	.03	.03
Pleasure	.03	.89	.08	0
Sadness	.08	.05	.86	.01
Anger	.02	.04	.12	.82

(a) Person-dependent

Actual \ Predicted	Joy	Pleasure	Sadness	Anger
Joy	.71	.08	.09	.12
Pleasure	.15	.7	.1	.05
Sadness	.07	.09	.73	.11
Anger	.06	.07	.12	.75

(b) Person-independent

Figure 8-9: **Confusion Matrix of Person-dependent and Person-independent Classification Results.** The diagonal of each of these matrices shows the classification accuracy and the off-diagonal grid points show the confusion error.

better at emotion management. We would like to test whether there is a difference in the performance of EQ-Radio's algorithms in classifying the emotions of actors vs. non-actors, as well as in classifying the emotions of males vs. females. We evaluate the performance of a specific group of subjects in terms of mutual predictability/consistency, i.e., we predict the emotion label of a data point by training on data obtained from within the same group only. Figure 8-10 shows our results. These results show that our emotion recognition algorithm works for both actors and non-actors, and for both genders. However, the accuracy of this algorithm is higher for actors than non-actors and for females than males. This could suggest that actors/females have better emotion management skills or that they are indeed more emotional.

EQ-Radio versus ECG-based emotion recognition

In this section, we compare EQ-Radio's emotion classification accuracy with that of an ECG-based classifier. Note that both classifiers use the same set of features and decision making process. However, the ECG-based classifier uses heartbeat information directly extracted from the ECG monitor. In addition, we allow the ECG monitor to access the breathing signal from EQ-Radio and use EQ-Radio's breathing features. This mirrors today's emotion monitors which also use breathing data but require the subject to wear a chest band in order to extract that signal.

The results in Table 8-2 show that EQ-Radio achieves comparable accuracy to emotion recognition systems that use on-body sensors. Thus, by using EQ-Radio, one can eliminate

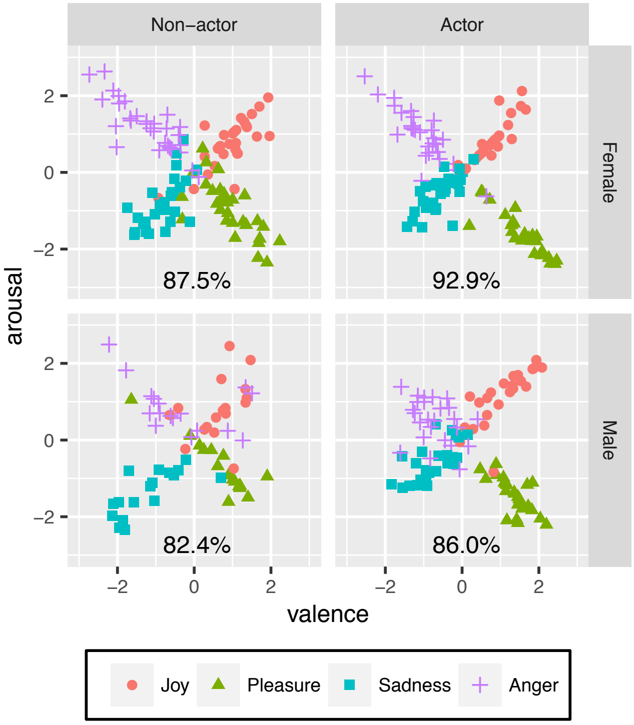


Figure 8-10: **Visualization of EQ-Radio's Group-dependent Classification Results.** The figure shows the results of EQ-Radio's classification within 4 different groups, defined by gender and acting experience. The x-axis corresponds to valence and the y-axis corresponds to arousal.

body sensors without jeopardizing the accuracy of emotion recognition based on physiological signals.

EQ-Radio versus vision-based emotion recognition

In order to compare the accuracy of EQ-Radio with vision-based emotion recognition systems, we use the Microsoft Project Oxford Emotion API to process the images of the subjects collected during the experiments, and analyze their emotions based on facial expressions. Since the Microsoft Emotion API and EQ-Radio use different emotion models, we use the following four emotions that both systems share for our comparison: joy/pleasure, sadness, anger, and neutral. For each data point, the Microsoft Emotion API outputs scores for eight emotions. We consider their scores for the above four shared emotions and use the label with highest score as their output.

Figure 8-11 compares the accuracy of EQ-Radio (both person-dependent and person-independent) with the Microsoft Emotion API. The figure shows that that the Microsoft Emotion API does not achieve high accuracy for the first three categories of emotions, but

Method	Person-dependent	Person-independent
EQ-Radio	87%	72.3%
ECG-based	88.2%	73.2%

Table 8-2: **Comparison with the ECG-based Method.** The table compares the accuracy of EQ-Radio's person-dependent and person-independent emotion classification accuracy with the emotion classification accuracy achieved using the ECG signals (combined with the extracted respiration features).

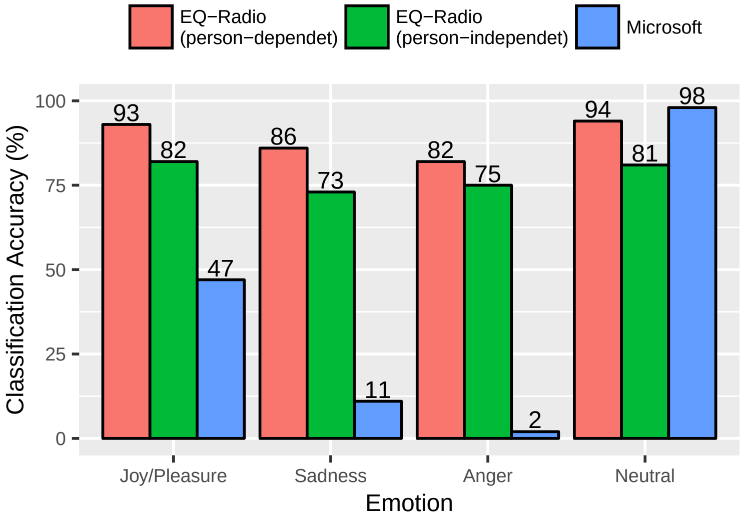


Figure 8-11: **Comparison of EQ-Radio with Image-based Emotion Recognition.** The figure shows the accuracies (on the y-axis) of EQ-Radio and Microsoft's Emotion API in differentiating among the four emotions (on the x-axis).

achieves very high accuracy for neutral state. This is because vision-based methods can recognize an emotion only when the person explicitly expresses it on her face, and fail to recognize the innermost emotions and hence they report such emotions as neutral. We also note that the Microsoft Emotion API has higher accuracy for positive emotions than negative ones. This is because positive emotions typically have more visible features (e.g., smiling), while negative emotions are visually closer to a neutral state.

Emotion recognition versus accurate beat segmentation

Finally, we would like to understand how tolerant emotion recognition is to errors in beat segmentation. We take the ground truth beats derived from the ECG monitor and add to them different levels of Gaussian noise. The Gaussian distribution has zero mean and its standard deviation varies between 0 and 60 milliseconds. We re-run the person-dependent emotion recognition classifier using these noisy beats. Figure 8-12 shows that small errors

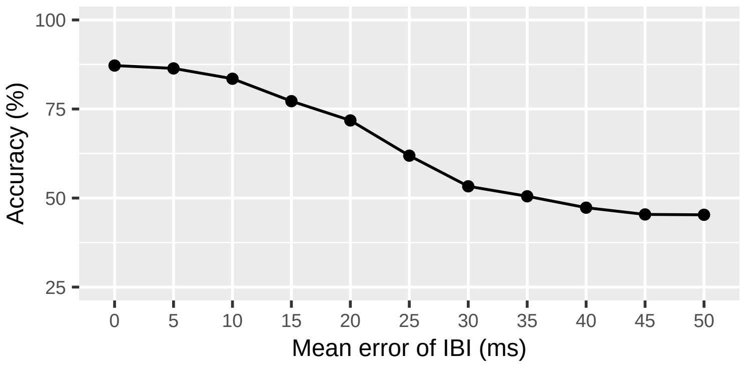


Figure 8-12: **Impact of Millisecond Errors in IBI on Emotion Recognition.** The figure shows that adding small errors to the IBI values (x-axis) significantly reduces the classification accuracy (y-axis). Given that we have four classes, a random guess would have 25% accuracy.

in estimating the beat lengths can lead to a large degradation in classification accuracy. In particular, an error of 30 milliseconds in inter-beat-interval can reduce the accuracy of emotion recognition by over 35%. This result emphasizes the importance of extracting the individual beats and delineating their boundaries at an accuracy of a few milliseconds.[7]

■ 8.6 Conclusion

In this section, we present a technology capable of recognizing a person's emotions by relying on wireless signals reflected off her/his body. We believe this marks an important step in the nascent field of emotion recognition. It also builds on a growing interest in the wireless systems' community in using RF signals for sensing, and as such, the work expands the scope of RF sensing to the domain of emotion recognition. Further, while this work has laid foundations for wireless emotion recognition, we envision that the accuracy of such systems will improve as wireless sensing technologies evolve and as the community incorporates more advanced machine learning mechanisms in the sensing process.

We also believe that the implications of this work extend beyond emotion recognition. Specifically, while we used the heartbeat extraction algorithm for determining the beat-to-beat intervals and exploited these intervals for emotion recognition, our algorithm recovers the entire human heartbeat from RF, and the heartbeat displays a very rich mor-

[7] Note that given that we have four classes, a random guess would have 25% accuracy. Adding small errors to the IBI values significantly reduces the classification accuracy. The accuracy converges to about 40% instead of 25% because the respiration features are left intact.

phology. We envision that this result paves way for exciting research on understanding the morphology of the heartbeat both in the context of emotion-recognition as well as in the context of non-invasive health monitoring and diagnosis.

CHAPTER 9
Conclusion

This thesis presents a data-driven approach to wireless sensing system design. By customizing machine learning models and algorithms for radio signals, the presented technologies are able to extract rich semantic information from radio signals despite their complex interaction with human bodies and the environment. We demonstrate the effectiveness of our approach and introduce two categories of new sensing capabilities – through-wall human sensing and contactless health monitoring. These systems detect humans through walls, track their movements, and recognize their actions. The proposed systems also passively and continuously monitor people's health – they capture people's vital signs and emotions, monitor their sleep and sleep stages, and detect and assess medication usage.

Our contributions span both wireless sensing and computer vision. From a wireless sensing perspective, we introduce a new approach that fundamentally improves the sensing capabilities of wireless systems. In particular, in contrast to traditional approaches that use signal processing algorithms with simple approximations of how radio signals interact with human bodies, our research learns these complex interactions from data with custom machine learning methods. As a result, our approaches is able to extract rich semantic information (e.g., human poses, body meshes, emotions, sleep stages, etc.) from radio signals. From a computer vision perspective, this dissertation introduces a new approach to deal with occlusion, which is a fundamental challenge for any vision system. Our approach could greatly improve the robustness and safety of modern vision systems by expanding them to work in the presence of occlusions and bad lighting conditions.

The work also has a broader impact on digital health and improving care and wellness. Our smart radio sensor could be deployed in the home of the elderly and chronically ill to monitor their sleep, vital signs, emotions, and activities. It could inform the caregiver of changes in health status and help doctors better understand diseases and monitor symptoms, and detect exacerbations at an earlier time. It could also be used in clinical trials to monitor patient's reaction to drugs, improve safety, and speed up the drug development process. In fact, some of the technologies presented in this dissertation have already been adopted and deployed in the real world with patients. In particular, our sleep monitor has been used in collaboration with University of Rochester Medical School to monitor sleep in Parkinson's patients [245], and in collaboration with the Washington's University Medical School to detect the impact of itch on sleep and sleep stages. More recently, these technologies have been used to remotely monitor COVID-19 patients [246].

While our technical work has not directly addressed privacy and security, we note that all experiments in this dissertation that involves human subjects have been reviewed and approved by the Institutional Review Board (IRB) of the Massachusetts Institute of Technology. Further, monitored subjects provided informed consent in accordance with IRB requirements. Additionally researchers who interacted with human subjects, including myself, have obtain all necessary training and certificates.

■ 9.1 Future Directions

This thesis represents only a first attempt at using wireless signals combined with machine learning to perform the sensing function. Looking forward, connecting wireless and IoT systems with other fields, including artificial intelligence and digital health, and working across software-hardware boundaries could make wireless sensing an indispensable part of people's lives. Below we highlight some of the future research directions.

Multi-Modal Sensing for Robotics: An exciting research avenue is to develop multi-modal sensing systems that integrate RF signals and IoT devices with other sensors such as cameras, LiDAR, and ultrasound. The research in this dissertation on RF sensing systems that see through walls and clouds could empower robots to sense through obstacles and occlusions. Exploring new cross-modal and multi-modal learning mechanisms could allow these systems to be trained with less or even zero human supervision. In addition, the future of robotic sensing could go beyond the sensors on the robot itself and leverage the increasing amount of IoT sensors embedded in the physical world. Since these IoT devices communicate with RF signals (e.g., Wi-Fi, Bluetooth, 5G), future research could re-purpose these signals to help robots understand the environment.

Digital Healthcare: Contactless human health sensing is a promising approach to improve healthcare. This dissertation introduced new RF sensing capabilities with implications for many disease areas. Future work could further investigate how such technologies can help address unmet healthcare needs. While powerful, wireless sensing technologies capture information that are not traditionally measured or used in clinical settings. Combining this information with traditional health data (X-rays, blood tests, medical records, etc.) through novel machine learning techniques would help develop new insights about patient health. It would also be important to validate these new solutions in clinical settings and integrating them into the new standards of care.

Exploiting Sensing for Wireless Communication at Higher Frequencies: New generations of wireless networks are moving to higher frequency bands, e.g., millimeter-wave and terahertz bands, which promise to significantly increase the network capacity. However, these high-frequency signals form highly directional beams, which require the transmitter to be aimed accurately toward the receiver. For example, in virtual reality, it would require the beam to accurately follow the moving player. Knowing how the user moves and capturing the human skeleton through obstacles could enable accurate beams towards receivers. While these high-frequency RF signals could bring in a new level of sensing resolution given larger bandwidth and denser antenna elements, they suffer from high path loss and cannot penetrate thick walls. Future research could combine RF signals of different frequencies to achieve high sensing resolution while maintaining the ability to see through walls and occlusions.

Environmental Health & Sustainability: Environmental health and sustainability are among the most critical and most challenging problems facing humanity. RF signals with large coverage areas (e.g., radio base stations and satellites) provide a unique opportunity for sensing and monitoring our planet. For example, can we use RF signals for the measurement of greenhouse gas from farms? Can we detect wildfires through smoke? While the research in this dissertation focused on indoor human sensing with RF signals, future research could explore the use of RF sensing in other contexts.

References

[1] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Dina Katabi, and Antonio Torralba. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

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

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

[4] Mingmin Zhao, Kreshnik Hoti, Hao Wang, Aniruddh Raghu, and Dina Katabi. Assessment of medication self-administration using artificial intelligence. *Nature medicine*, 27(4):727–735, 2021.

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

[6] Mingmin Zhao, Fadel Adib, and Dina Katabi. Emotion recognition using wireless signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, MobiCom, 2016.

[7] Fadel Adib and Dina Katabi. See through walls with wifi! In *Proceedings of the ACM SIGCOMM 2013*, pages 75–86, New York, NY, USA, 2013. ACM.

[8] Fadel Adib, Zachary Kabelac, Dina Katabi, and Robert C. Miller. 3D tracking via body radio reflections. In *Proceedings of the USENIX Conference on Networked Systems Design and Implementation*, NSDI, 2014.

[9] Fadel Adib, Zachary Kabelac, and Dina Katabi. Multi-person localization via RF body reflections. In *Proceedings of the USENIX Conference on Networked Systems Design and Implementation*, NSDI, 2015.

[10] Kiran Raj Joshi, Dinesh Bharadia, Manikanta Kotaru, and Sachin Katti. Wideo: Fine-grained device-free motion tracing using rf backscatter. In *Proceedings of the USENIX Conference on Networked Systems Design and Implementation*, NSDI, 2015.

[11] Ju Wang, Hongbo Jiang, Jie Xiong, Kyle Jamieson, Xiaojiang Chen, Dingyi Fang, and Binbin Xie. Lifs: Low human-effort, device-free localization with fine-grained subcarrier information. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pages 243–256, 2016.

[12] Xiang Li, Shengjie Li, Daqing Zhang, Jie Xiong, Yasha Wang, and Hong Mei. Dynamic-music: accurate device-free indoor localization. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 196–207, 2016.

[13] Yaxiong Xie, Jie Xiong, Mo Li, and Kyle Jamieson. md-track: Leveraging multi-dimensionality for passive indoor wi-fi tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–16, 2019.

[14] Wei Wang, Alex X Liu, and Muhammad Shahzad. Gait recognition using WiFi signals. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016.

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

[16] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, ICCV, 2015.

[17] Alejandro Newell and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *arXiv preprint arXiv:1611.05424*, 2016.

[18] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, ECCV, 2016.

[19] Wei Yang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *CVPR*, 2016.

[20] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, 2016.

[21] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *Proceedings of the European Conference on Computer Vision*, ECCV, 2016.

[22] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, 2017.

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

[24] Matteo Ruggero Ronchi and Pietro Perona. Benchmarking and error diagnosis in multi-instance pose estimation. In *Proceedings of the IEEE international conference on computer vision*, ICCV, 2017.

[25] Xiao Chu, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Structured feature learning for pose estimation. In *CVPR*, 2016.

[26] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3D human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[27] Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. Thin-slicing network: A deep structured model for pose estimation in videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[28] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

[29] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L. Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *CVPR*, 2017.

[30] Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, and Dina Katabi. Making the invisible visible: Action recognition through walls and occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[31] Lijie Fan, Tianhong Li, Rongyao Fang, Rumen Hristov, Yuan Yuan, and Dina Katabi. Learning longterm representations for person re-identification using radio signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10699–10709, 2020.

[32] Lijie Fan, Tianhong Li, Yuan Yuan, and Dina Katabi. In-home daily-life captioning using radio signals. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 105–123. Springer, 2020.

[33] 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 wifi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pages 1–14, 2020.

[34] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. mmmesh: towards 3d real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, pages 269–282, 2021.

[35] Yongkun Song, Tian Jin, Yongpeng Dai, Yongping Song, and Xiaolong Zhou. Through-wall human pose reconstruction via uwb mimo radar and 3d cnn. *Remote Sensing*, 13(2):241, 2021.

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

[37] Eleftheria Vaportzis, Maria Giatsi Clausen, , and Alan J. Gow. Older adults perceptions of technology and barriers to interacting with tablet computers: A focus group study. In *Frontiers in Psychology*, 2017.

[38] Nadir G. Abdelrahman, Raza Haque, Molly E. Polverento, Andrea Wendling, Courtney M. Goetz, and Bengt B. Arnetz. Brain health: Attitudes towards technology adoption in older adults. In *Healthcare*, 2021.

[39] Meera Viswanathan, Carol E Golin, Christine D Jones, Mahima Ashok, Susan J Blalock, Roberta CM Wines, Emmanuel JL Coker-Schwimmer, David L Rosen, Priyanka Sista, and Kathleen N Lohr. Interventions to improve adherence to self-administered medications for chronic diseases in the united states: a systematic review. *Annals of internal medicine*, 157(11):785–795, 2012.

[40] Aurel O Iuga and Maura J McGuire. Adherence and health care costs. *Risk management and healthcare policy*, 7:35, 2014.

[41] Catherine E Cooke, Helen Y Lee, Yvette P Tong, and Stuart T Haines. Persistence with injectable antidiabetic agents in members with type 2 diabetes in a commercial managed care organization. *Current medical research and opinion*, 26(1):231–238, 2010.

[42] M Peyrot, AH Barnett, LF Meneghini, and P-M Schumm-Draeger. Insulin adherence behaviours and barriers in the multinational global attitudes of patients and physicians in insulin therapy study. *Diabetic Medicine*, 29(5):682–689, 2012.

[43] Mathieu Molimard, Chantal Raherison, Severine Lignot, Aurelie Balestra, Stephanie Lamarque, Anais Chartier, Cecile Droz-Perroteau, Regis Lassalle, Nicholas Moore, and Pierre-Olivier Girodet. Chronic obstructive pulmonary disease exacerbation and inhaler device handling: real-life assessment of 2935 patients. *European Respiratory Journal*, 49(2), 2017.

[44] Elizabeth Selvin, Christina M Parrinello, Natalie Daya, and Richard M Bergenstal. Trends in insulin use and diabetes control in the us: 1988–1994 and 1999–2012. *Diabetes Care*, 39(3):e33–e35, 2016.

[45] National Asthma Council Australia. Inhaler technique for people with asthma or copd. https://www.nationalasthma.org.au/living-with-asthma/resources/health-professionals/information-paper/hp-inhaler-technique-for-people-with-asthma-or-copd.

[46] American Diabetes Association et al. Insulin administration. *Diabetes care*, 27(suppl 1):s106–s107, 2004.

[47] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[48] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.

[49] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[50] Daniel S Quintana, Adam J Guastella, Tim Outhred, Ian B Hickie, and Andrew H Kemp. Heart rate variability is associated with emotion recognition: direct evidence for a relationship between the autonomic nervous system and social cognition. *International Journal of Psychophysiology*, 86(2):168–172, 2012.

[51] Richard D Lane, Kateri McRae, Eric M Reiman, Kewei Chen, Geoffrey L Ahern, and Julian F Thayer. Neural correlates of heart rate variability during emotion. *Neuroimage*, 44(1):213–222, 2009.

[52] Rafael A Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *Affective Computing, IEEE Transactions on*, 1(1):18–37, 2010.

[53] S Jerritta, M Murugappan, R Nagarajan, and Khairunizam Wan. Physiological signals based human emotion recognition: a review. In *Signal Processing and its Applications (CSPA), 2011 IEEE 7th International Colloquium on*, pages 410–415. IEEE, 2011.

[54] 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*, pages 313–325, 2019.

[55] Yonglong Tian, Guang-He Lee, Hao He, Chen-Yu Hsu, and Dina Katabi. Rf-based fall monitoring using convolutional neural networks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–24, 2018.

[56] Kanil Patel, Kilian Rambach, Tristan Visentin, Daniel Rusev, Michael Pfeiffer, and Bin Yang. Deep learning-based object classification on automotive radar spectra. In *2019 IEEE Radar Conference (RadarConf)*, pages 1–6. IEEE, 2019.

[57] Peijun Zhao, Chris Xiaoxuan Lu, Jianan Wang, Changhao Chen, Wei Wang, Niki Trigoni, and Andrew Markham. mid: Tracking and identifying people with millimeter wave radar. In *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 33–40. IEEE, 2019.

[58] Han Zou, Jianfei Yang, Hari Prasanna Das, Huihan Liu, Yuxun Zhou, and Costas J Spanos. Wifi and vision multimodal learning for accurate and robust device-free human activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[59] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE Sensors Journal*, 20(17):10032–10044, 2020.

[60] Fida Mohammad Thoker and Juergen Gall. Cross-modal knowledge distillation for action recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 6–10. IEEE, 2019.

[61] Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel. Whole-home gesture recognition using wireless signals. In *Proceedings of the 19th Annual International Conference on Mobile computing & Networking*, MobiCom. ACM, 2013.

[62] Li Sun, Souvik Sen, Dimitrios Koutsonikolas, and Kyu-Han Kim. Widraw: Enabling hands-free drawing in the air on commodity wifi devices. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 77–89. ACM, 2015.

[63] Kamran Ali, Alex Xiao Liu, Wei Wang, and Muhammad Shahzad. Keystroke recognition using wifi signals. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 90–102. ACM, 2015.

[64] Teng Wei and Xinyu Zhang. mtrack: High-precision passive tracking using millimeter wave radios. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 117–129. ACM, 2015.

[65] Heba Abdelnasser, Moustafa Youssef, and Khaled A Harras. Wigest: A ubiquitous wifi-based gesture recognition system. In *Computer Communications (INFOCOM), 2015 IEEE Conference on*, pages 1472–1480. IEEE, 2015.

[66] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 65–76. ACM, 2015.

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

[68] Fadel Adib, Hongzi Mao, Zachary Kabelac, Dina Katabi, and Robert C Miller. Smart homes that monitor breathing and heart rate. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015.

[69] Amy D Droitcour, Olga Boric-Lubecke, and Gregory TA Kovacs. Signal-to-noise ratio in doppler radar system for heart and respiratory rate measurements. *Microwave Theory and Techniques, IEEE Transactions on*, 57(10):2498–2507, 2009.

[70] Amy D Droitcour, Olga Boric-Lubecke, Victor M Lubecke, Jenshan Lin, and Gregory TA Kovacs. Range correlation and I/Q performance benefits in single-chip silicon doppler radars for noncontact cardiopulmonary monitoring. *Microwave Theory and Techniques, IEEE Transactions on*, 52(3):838–848, 2004.

[71] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. Capturing the human figure through a wall. *ACM Transactions on Graphics*, 34(6):219, November 2015.

[72] Yanzi Zhu, Yibo Zhu, Ben Y Zhao, and Haitao Zheng. Reusing 60ghz radios for mobile radar imaging. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 103–116, 2015.

[73] Jie Xiong and Kyle Jamieson. Arraytrack: A fine-grained indoor location system. In *Proceedings of the USENIX Conference on Networked Systems Design and Implementation*, NSDI, 2013.

[74] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. Spotfi: Decimeter level localization using wifi. In *ACM SIGCOMM Computer Communication Review*, 2015.

[75] Deepak Vasisht, Swarun Kumar, and Dina Katabi. Decimeter-level localization with a single wifi access point. In *NSDI*, 2016.

[76] Pedro Melgarejo, Xinyu Zhang, Parameswaran Ramanathan, and David Chu. Leveraging directional antenna capabilities for fine-grained gesture recognition. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014.

[77] Rich Fletcher and Jing Han. Low-cost differential front-end for doppler radar vital sign monitoring. In *Microwave Symposium Digest, 2009. MTT'09. IEEE MTT-S International*, pages 1325–1328. IEEE, 2009.

[78] Neal Patwari, Lara Brewer, Quinn Tate, Ossi Kaltiokallio, and Maurizio Bocca. Breathfinding: A wireless network that monitors and locates breathing in a home. *Selected Topics in Signal Processing, IEEE Journal of*, 8(1):30–42, 2014.

[79] Ossi Kaltiokallio, Huseyin Yigitler, Riku Jantti, and Neal Patwari. Non-invasive respiration rate monitoring using a single cots tx-rx pair. In *Information Processing in Sensor Networks, IPSN-14 Proceedings of the 13th International Symposium on*, pages 59–69. IEEE, 2014.

[80] Alberto Zaffaroni, Philip De Chazal, Conor Heneghan, Patricia Boyle, Patricia Ronayne Mppm, and Walter T McNicholas. Sleepminder: an innovative contact-free device for the estimation of the apnoea-hypopnoea index. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 7091–9094. IEEE, 2009.

[81] Philip De Chazal, Emer O Hare, Niall Fox, and Conor Heneghan. Assessment of sleep/wake patterns using a non-contact biomotion sensor. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pages 514–517. IEEE, 2008.

[82] Wansuree Massagram, Victor M Lubecke, Anders Host-Madsen, and Olga Boric-Lubecke. Assessment of heart rate variability and respiratory sinus arrhythmia via doppler radar. *Microwave Theory and Techniques, IEEE Transactions on*, 57(10):2542–2549, 2009.

[83] Wei Hu, Zhangyan Zhao, Yunfeng Wang, Haiying Zhang, and Fujiang Lin. Noncontact accurate measurement of cardiopulmonary activity using a compact quadrature doppler radar sensor. *Biomedical Engineering, IEEE Transactions on*, 61(3):725–735, 2014.

[84] Olga Boric-Lubecke, Wansuree Massagram, Victor M Lubecke, Anders Host-Madsen, and Branka Jokanovic. Heart rate variability assessment using doppler radar with linear demodulation. In *Microwave Conference, 2008. EuMC 2008. 38th European*, pages 420–423. IEEE, 2008.

[85] T. Sakamoto, R. Imasaka, H. Taki, T. Sato, M. Yoshioka, K. Inoue, T. Fukuda, and H. Sakai. Feature-based correlation and topological similarity for interbeat interval estimation using ultra-wideband radar. *Biomedical Engineering, IEEE Transactions on*, 2015.

[86] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[87] Georgia Gkioxari, Bharath Hariharan, Ross Girshick, and Jitendra Malik. Using k-poselets for detecting people and localizing their keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, 2014.

[88] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, 2016.

[89] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards Accurate Multi-person Pose Estimation in the Wild. In *CVPR*, 2017.

[90] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012.

[91] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1):4–27, March 2010.

[92] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[93] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, ICCV, pages 1440–1448, 2015.

[94] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the International Conference on Neural Information Processing Systems*, NIPS, pages 91–99, 2015.

[95] Petr Beckmann and Andre Spizzichino. The scattering of electromagnetic waves from rough surfaces. *Norwood, MA, Artech House, Inc.*, 1987.

[96] Human3.6M Dataset. http://vision.imar.ro/human3.6m (accessed january 31, 2018), 2018.

[97] Yu Chen, Tae-Kyun Kim, and Roberto Cipolla. Inferring 3D shapes and deformations from single views. In *European Conference on Computer Vision*, pages 300–313. Springer, 2010.

[98] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1381–1388. IEEE, 2009.

[99] Nils Hasler, Hanno Ackermann, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1823–1830. IEEE, 2010.

[100] Jun Kai Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3D human body shape and pose prediction. In *BMVC*, volume 3, page 6, 2017.

[101] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[102] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016.

[103] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 3, 2017.

[104] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[105] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018.

[106] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019.

[107] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NIPS*, 2014.

[108] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[109] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *NIPS*, 2016.

[110] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[111] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, NIPS, 2016.

[112] Sanjay Purushotham, Wilka Carvalho, Tanachat Nilanon, and Yan Liu. Variational recurrent adversarial deep domain adaptation. *ICLR*, 2017.

[113] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *ICLR*, 2017.

[114] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. *ICLR*, 2017.

[115] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

[116] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. *ICML*, 2017.

[117] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *ICLR*, 2017.

[118] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *ICML*, 2015.

[119] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 2016.

[120] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. *ICCV*, 2015.

[121] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. *NIPS Workshop on Adversarial Training*, 2016.

[122] Sven Stegemann, J-P Baeyens, F Cerreta, E Chanie, A Löfgren, Mario Maio, G Schreier, and E Thesing-Bleck. Adherence measurement systems and technology for medications in older patient populations. *European Geriatric Medicine*, 3(4):254–260, 2012.

[123] Federico Lavorini, Christer Janson, Fulvio Braido, Georgios Stratelis, and Anders Løkke. What to consider before prescribing inhaled medications: a pragmatic approach for evaluating the current inhaler landscape. *Therapeutic advances in respiratory disease*, 13:1753466619884532, 2019.

[124] Murtadha Aldeer, Mehdi Javanmard, and Richard P Martin. A review of medication adherence monitoring technologies. *Applied System Innovation*, 1(2):14, 2018.

[125] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 1960.

[126] Anders H Frid, Gillian Kreugel, Giorgio Grassi, Serge Halimi, Debbie Hicks, Laurence J Hirsch, Mike J Smith, Regine Wellhoener, Bruce W Bode, Irl B Hirsch, et al. New insulin delivery recommendations. In *Mayo Clinic Proceedings*, volume 91, pages 1231–1255. Elsevier, 2016.

[127] Zachary Kabelac, Christopher G Tarolli, Christopher Snyder, Blake Feldman, Alistair Glidden, Chen-Yu Hsu, Rumen Hristov, E Ray Dorsey, and Dina Katabi. Passive monitoring at home: a pilot study in parkinson disease. *Digital biomarkers*, 3(1):22–30, 2019.

[128] Paula M Trief, Donald Cibula, Elaine Rodriguez, Bridget Akel, and Ruth S Weinstock. Incorrect insulin administration: a problem that warrants attention. *Clinical Diabetes*, 34(1):25–33, 2016.

[129] Allan Rechtschaffen and Anthony Kales. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. *US Government Printing Office, US Public Health Service*, 1968.

[130] Farideh Ebrahimi, Mohammad Mikaeili, Edson Estrada, and Homer Nazeran. Automatic sleep stage classification based on eeg signals by using neural networks and wavelet packet coefficients. *IEEE EMBC*, 2008.

[131] Luay Fraiwan, Khaldon Lweesy, Natheer Khasawneh, Heinrich Wenz, and Hartmut Dickhaus. Automated sleep stage identification system based on time–frequency analysis of a single eeg channel and random forest classifier. *Computer methods and programs in biomedicine*, 2012.

[132] Djordje Popovic, Michael Khoo, and Philip Westbrook. Automatic scoring of sleep stages and cortical arousals using two electrodes on the forehead: validation in healthy adults. *Journal of sleep research*, 2014.

[133] John R Shambroom, Stephan E Fábregas, and Jack Johnstone. Validation of an automated wireless system to monitor sleep in healthy adults. *Journal of sleep research*, 2012.

[134] Alexander Tataraidze, Lyudmila Korostovtseva, Lesya Anishchenko, Mikhail Bochkarev, and Yurii Sviryaev. Sleep architecture measurement based on cardiorespiratory parameters. *IEEE EMBC*, 2016.

[135] Xi Long, Jie Yang, Tim Weysen, Reinder Haakma, Jérôme Foussier, Pedro Fonseca, and Ronald M Aarts. Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging. *Physiological measurement*, 2014.

[136] Tian Hao, Guoliang Xing, and Gang Zhou. isleep: unobtrusive sleep quality monitoring using smartphones. *ACM SenSys*, 2013.

[137] Weixi Gu, Zheng Yang, Longfei Shangguan, Wei Sun, Kun Jin, and Yunhao Liu. Intelligent sleep stage mining service with smartphones. *ACM UbiComp*, 2014.

[138] Charles P Pollak, Warren W Tryon, Haikady Nagaraja, and Roger Dzwonczyk. How accurately does wrist actigraphy identify the states of sleep and wakefulness? *SLEEP-NEW YORK*, 2001.

[139] Genevieve Alelis, Ania Bobrowicz, and Chee Siang Ang. Exhibiting emotion: Capturing visitors' emotional responses to museum artefacts. In *Design, User Experience, and Usability. User Experience in Novel Technological Environments*, pages 429–438. Springer, 2013.

[140] Antonio Fernández-Caballero, José Miguel Latorre, José Manuel Pastor, and Alicia Fernández-Sotos. Improvement of the elderly quality of life and care through smart emotion regulation. In *Ambient Assisted Living and Daily Activities*, pages 348–355. Springer, 2014.

[141] Daniel McDuff, Rana El Kaliouby, Jeffrey F Cohn, and Rosalind W Picard. Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *Affective Computing, IEEE Transactions on*, 6(3):223–235, 2015.

[142] Mohammad Soleymani, Maja Pantic, and Thierry Pun. Multimodal emotion recognition in response to videos. *Affective Computing, IEEE Transactions on*, 3(2):211–223, 2012.

[143] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[144] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[145] Jonghwa Kim and Elisabeth André. Emotion recognition based on physiological changes in music listening. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(12):2067–2083, 2008.

[146] Bradley M Appelhans and Linda J Luecken. Heart rate variability as an index of regulated emotional responding. *Review of general psychology*, 10(3):229, 2006.

[147] Texas Instruments. Texas instruments. `http://www.ti.com/`.

[148] Walabot. Walabot. `https://walabot.com/`.

[149] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European conference on computer vision*, ECCV, 2014.

[150] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[151] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, January 2013.

[152] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 4489–4497. IEEE, 2015.

[153] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, 2010.

[154] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[155] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, ICCV, 2015.

[156] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[157] Yunze Zeng, Parth H Pathak, and Prasant Mohapatra. Wiwho: wifi-based person identification in smart spaces. In *Proceedings of the 15th International Conference on Information Processing in Sensor Networks*, page 4, 2016.

[158] Yuxi Wang, Kaishun Wu, and Lionel M Ni. Wifall: Device-free fall detection by wireless networks. *IEEE Transactions on Mobile Computing*, 16(2):581–594, 2017.

[159] Mark A Richards. *Fundamentals of radar signal processing*. Tata McGraw-Hill Education, 2005.

[160] Vicon Motion Systems. https://www.vicon.com/ (accessed january 31, 2018), 2018.

[161] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

[162] Lior Rokach and Oded Maimon. Clustering methods. In *Data mining and knowledge discovery handbook*. Springer, 2005.

[163] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[164] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[165] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR, 2014.

[166] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018.

[167] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[168] Zhoutong Zhang, Qiujia Li, Zhengjia Huang, Jiajun Wu, Josh Tenenbaum, and Bill Freeman. Shape and material from sound. In *Advances in Neural Information Processing Systems*, pages 1278–1288, 2017.

[169] Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward Adelson. Connecting look and feel: Associating the visual and tactile properties of physical materials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5580–5588, 2017.

[170] Amanda Berg, Jorgen Ahlberg, and Michael Felsberg. Generating visible spectrum images from thermal infrared. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1143–1152, 2018.

[171] Andrew G Stove. Linear fmcw radar techniques. In *IEE Proceedings F (Radar and Signal Processing)*, volume 139, pages 343–350. IET, 1992.

[172] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Kari Pulli. Short-range fmcw monopulse radar for hand-gesture sensing. In *2015 IEEE Radar Conference (RadarCon)*, pages 1491–1496. IEEE, 2015.

[173] Emerald Innovations. Emerald. https://www.emeraldinno.com/clinical.

[174] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[175] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-track: Efficient pose estimation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 350–359, 2018.

[176] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[177] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *ICLR*, 2017.

[178] Carnegie Mellon Graphics Lab. CMU Graphics Lab Motion Capture Database.

[179] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):220, 2014.

[180] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. Towards accurate markerless human shape and pose estimation over time. In *International Conference on 3D Vision (3DV)*, 2017.

[181] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406*, 2018.

[182] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[183] Cristian Sminchisescu and Bill Triggs. Covariance scaled sampling for monocular 3D body tracking. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[184] Wai Yin Lam and Paula Fresco. Medication adherence measures: an overview. *BioMed research international*, 2015, 2015.

[185] Leah L Zullig, Walid F Gellad, Jivan Moaddeb, Matthew J Crowley, William Shrank, Bradi B Granger, Christopher B Granger, Troy Trygstad, Larry Z Liu, and Hayden B Bosworth. Improving diabetes medication adherence: successful, scalable interventions. *Patient preference and adherence*, 9:139, 2015.

[186] Federico Lavorini, Antoine Magnan, Jean Christophe Dubus, Thomas Voshaar, Lorenzo Corbetta, Marielle Broeders, Richard Dekhuijzen, Joaquin Sanchis, Jose L Viejo, Peter Barnes, et al. Effect of incorrect use of dry powder inhalers on management of patients with asthma and copd. *Respiratory medicine*, 102(4):593–604, 2008.

[187] John Urquhart. Role of patient compliance in clinical pharmacokinetics. *Clinical pharmacokinetics*, 27(3):202–215, 1994.

[188] NR Samaranayake and BMY Cheung. Medication safety in hospitals: avoiding medication errors in the medication use process. *Advanced Pharmacoepidemiology Drug Safety*, 2:2167–1052, 2013.

[189] Paul Kelly, Simon J Marshall, Hannah Badland, Jacqueline Kerr, Melody Oliver, Aiden R Doherty, and Charlie Foster. An ethical framework for automated, wearable cameras in health behavior research. *American journal of preventive medicine*, 44(3):314–319, 2013.

[190] Teresa H Truong, Trang T Nguyen, Becky L Armor, and Jamie R Farley. Errors in the administration technique of insulin pen devices: a result of insufficient education. *Diabetes Therapy*, 8(2):221–226, 2017.

[191] Geralyn Spollett, Steven V Edelman, Patricia Mehner, Claudia Walter, and Alfred Penfornis. Improvement of insulin injection technique: examination of current issues and recommendations. *The Diabetes Educator*, 42(4):379–394, 2016.

[192] Anna Murphy. How to help patients optimise their inhaler technique. *The Pharmaceutical Journal*, 297(7891):11–21, 2016.

[193] Violaine Giraud, François-André Allaert, and Nicolas Roche. Inhaler technique and asthma: feasability and acceptability of training by pharmacists. *Respiratory medicine*, 105(12):1815–1822, 2011.

[194] Songül Göriş, Sultan Taşci, and Ferhan Elmali. The effects of training on inhaler technique and quality of life in patients with copd. *Journal of aerosol medicine and pulmonary drug delivery*, 26(6):336–344, 2013.

[195] Giuseppe Lippi and Brandon Michael Henry. Chronic obstructive pulmonary disease is associated with severe coronavirus disease 2019 (covid-19). *Respiratory medicine*, 167:105941, 2020.

[196] Weina Guo, Mingyue Li, Yalan Dong, Haifeng Zhou, Zili Zhang, Chunxia Tian, Renjie Qin, Haijun Wang, Yin Shen, Keye Du, et al. Diabetes is a risk factor for the progression and prognosis of covid-19. *Diabetes/metabolism research and reviews*, 36(7):e3319, 2020.

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

[198] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[199] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[200] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997.

[201] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[202] Robert G Newcombe. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine*, 17(8):857–872, 1998.

[203] Andrea S Melani, Marco Bonavia, Vincenzo Cilenti, Cristina Cinti, Marco Lodi, Paola Martucci, Maria Serra, Nicola Scichilone, Piersante Sestini, Maria Aliani, et al. Inhaler mishandling remains common in real life and is associated with reduced disease control. *Respiratory medicine*, 105(6):930–938, 2011.

[204] Beth L Laube, Hettie M Janssens, Frans HC de Jongh, Sunalene G Devadason, Rajiv Dhand, Patrice Diot, Mark L Everard, Ildiko Horvath, Paolo Navalesi, Thomas Voshaar, et al. What the pulmonary specialist should know about the new inhalation therapies, 2011.

[205] C Victor Spain, Jonathon J Wright, Rebecca M Hahn, Ashley Wivel, and Alan A Martin. Self-reported barriers to adherence and persistence to treatment with injectable medications for type 2 diabetes. *Clinical therapeutics*, 38(7):1653–1664, 2016.

[206] D Price, S Bosnic-Anticevich, A Briggs, Henry Chrystyn, C Rand, G Scheuch, J Bousquet, Inhaler Error Steering Committee, et al. Inhaler competence in asthma: common errors, barriers to use and recommended solutions. *Respiratory medicine*, 107(1):37–46, 2013.

[207] National Institute of Health. Sleep disorders. http://www.ninds.nih.gov/disorders/brain_basics/understanding_sleep.html.

[208] Mary A Carskadon and Allan Rechtschaffen. Monitoring and staging human sleep. *Principles and practice of sleep medicine*, 2000.

[209] Ellen Herbst et al. Adaptation effects to sleep studies in participants with and without chronic posttraumatic stress disorder. *Psychophysiology*, 2010.

[210] A Zaffaroni, L Gahan, L Collins, E O'hare, C Heneghan, C Garcia, I Fietze, and T Penzel. Automated sleep staging classification using a non-contact biomotion sensor. *Journal of Sleep Research*, 2014.

[211] Alexander Tataraidze, Lyudmila Korostovtseva, Lesya Anishchenko, Mikhail Bochkarev, Yurii Sviryaev, and Sergey Ivashov. Bioradiolocation-based sleep stage classification. *IEEE EMBC*, 2016.

[212] Thiago LT da Silveira, Alice J Kozakevicius, and Cesar R Rodrigues. Single-channel eeg sleep stage classification based on a streamlined set of statistical features in wavelet domain. *Medical & biological engineering & computing*, 2016.

[213] Brendan P Lucey, Jennifer S Mcleland, Cristina D Toedebusch, Jill Boyd, John C Morris, Eric C Landsness, Kelvin Yamada, and David M Holtzman. Comparison of a single-channel eeg sleep study to polysomnography. *Journal of sleep research*, 2016.

[214] Purav C Shah, Eric Yudelevich, Frank Genese, Miguel Martillo, Iazsmin B Ventura, Katherine Fuhrmann, Marie Mortel, Daniel Levendowski, Charlisa D Gibson, Pius Ochieng, et al. Can disrupted sleep affect mortality in the mechanically ventilated critically ill? *A state of unrest: Sleep/SDB in the ICU and hospital*, 2016.

[215] Richard Landis and Gary Koch. The measurement of observer agreement for categorical data. *Biometrics*, 1977.

[216] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008.

[217] Mary A Carskadon, William C Dement, et al. Normal human sleep: an overview. *Principles and practice of sleep medicine*, 2005.

[218] Deba Pratim Saha, Thomas L Martin, and R Benjamin Knapp. Towards incorporating affective feedback into context-aware intelligent environments. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 49–55. IEEE, 2015.

[219] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18(1):32–80, 2001.

[220] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, et al. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, pages 1–13, 2015.

[221] Rosalind W Picard. Affective computing: from laughter to ieee. *Affective Computing, IEEE Transactions on*, 1(1):11–17, 2010.

[222] Hillary Anger Elfenbein and Nalini Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203, 2002.

[223] Klaus R Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1):227–256, 2003.

[224] Rosalind W Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(10):1175–1191, 2001.

[225] Foteini Agrafioti, Dimitrios Hatzinakos, and Adam K Anderson. Ecg pattern analysis for emotion detection. *Affective Computing, IEEE Transactions on*, 3(1):102–115, 2012.

[226] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013.

[227] Todd B Kashdan, Anjali Mishra, William E Breen, and Jeffrey J Froh. Gender differences in gratitude: Examining appraisals, narratives, the willingness to express emotions, and changes in psychological needs. *Journal of personality*, 77(3):691–730, 2009.

[228] Sylvia D Kreibig. Autonomic nervous system activity in emotion: A review. *Biological psychology*, 84(3):394–421, 2010.

[229] Microsoft Research. Microsoft project Oxford emotion API. https://www.projectoxford.ai/emotion.

[230] Peter J Lang. The emotion probe: studies of motivation and attention. *American psychologist*, 50(5):372, 1995.

[231] Laura Anitori, Ardjan de Jong, and Frans Nennie. Fmcw radar for life-sign detection. In *Radar Conference, 2009 IEEE*, pages 1–6. IEEE, 2009.

[232] Octavian Postolache, Pedro Silva Girão, Gabriela Postolache, and Joaquim Gabriel. Cardio-respiratory and daily activity monitor based on fmcw doppler radar embedded in a wheelchair. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 1917–1920. IEEE, 2011.

[233] David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge university press, 2005.

[234] Pavel Holoborodko. Noise robust differentiators for second derivative estimation. http://www.holoborodko.com/pavel/downloads/NoiseRobustSecondDerivative.

[235] Sky McKinley and Megan Levine. Cubic spline interpolation. *College of the Redwoods*, 45(1):1049–1060, 1998.

[236] Andrew D Wiens, Mozziyar Etemadi, Shuvo Roy, Liviu Klein, and Omer T Inan. Toward continuous, noninvasive assessment of ventricular function and hemodynamics: Wearable ballistocardiography. *Biomedical and Health Informatics, IEEE Journal of*, 19(4):1435–1442, 2015.

[237] Rajendra Acharya, Paul Joseph, Natarajan Kannathal, Choo Min Lim, and Jasjit Suri. Heart rate variability: a review. *Medical and biological engineering and computing*, 44(12):1031–1051, 2006.

[238] Peter Walter Kamen, Henry Krum, and Andrew Maxwell Tonkin. Poincare plot of heart rate variability allows quantitative display of parasympathetic nervous activity in humans. *Clinical science*, 91(2):201–208, 1996.

[239] Douglas E Lake, Joshua S Richman, M Pamela Griffin, and J Randall Moorman. Sample entropy analysis of neonatal heart rate variability. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 283(3):R789–R797, 2002.

[240] Thomas Penzel, Jan W Kantelhardt, Ludger Grote, Jörg-Hermann Peter, and Armin Bunde. Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea. *Biomedical Engineering, IEEE Transactions on*, 50(10):1143–1151, 2003.

[241] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[242] Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 1-norm support vector machines. *Advances in neural information processing systems*, 16(1):49–56, 2004.

[243] Juan Sztajzel et al. Heart rate variability: a noninvasive electrocardiographic method to measure the autonomic nervous system. *Swiss medical weekly*, 134:514–522, 2004.

[244] Pierre A Devijver and Josef Kittler. *Pattern recognition: A statistical approach*, volume 761. Prentice-Hall London, 1982.

[245] C Tarolli, Z Kabelac, T Myers, E Waddell, H Rahul, R Hristov, P Auinger, T Nordahl, E Dorsey, T Ellis, et al. A day in the life of parkinson's: Using passive monitoring to characterize the disease at home: 1433. *Movement Disorders*, 35, 2020.

[246] Guo Zhang, Ipsit V Vahia, Yingcheng Liu, Yuzhe Yang, Rose May, Hailey V Cray, William McGrory, and Dina Katabi. Contactless in-home monitoring of the long-term respiratory and behavioral phenotypes in older adults with covid-19: A case series. *Frontiers in Psychiatry*, 12, 2021.