# IMU2Doppler: Cross-Modal Domain Adaptation for Doppler-based Activity Recognition Using IMU Data

SEJAL BHALLA, University of Toronto, Canada
MAYANK GOEL, Carnegie Mellon University, USA
RUSHIL KHURANA, Carnegie Mellon University, USA

The proliferation of sensors powered by state-of-the-art machine learning techniques can now infer context, recognize activities and enable interactions. A key component required to build these automated sensing systems is labeled training data. However, the cost of collecting and labeling new data impedes our ability to deploy new sensors to recognize human activities. We tackle this challenge using domain adaptation *i.e.,* using existing labeled data in a different domain to aid the training of a machine learning model for a new sensor. In this paper, we use off-the-shelf smartwatch IMU datasets to train an activity recognition system for mmWave radar sensor with minimally labeled data. We demonstrate that despite the lack of extensive datasets for mmWave radar, we are able to use our domain adaptation approach to build an activity recognition system that classifies between 10 activities with an accuracy of 70% with only 15 seconds of labeled doppler data. We also present results for a range of available labeled data (10 - 30 seconds) and show that our approach outperforms the baseline in every single scenario. We take our approach a step further and show that multiple IMU datasets can be combined together to act as a single source for our domain adaptation approach. Lastly, we discuss the limitations of our work and how it can impact future research directions.

CCS Concepts: • **Computing methodologies** → **Transfer learning**; • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: domain adaptation, doppler sensor, imu, activity recognition

**ACM Reference Format:**
Sejal Bhalla, Mayank Goel, and Rushil Khurana. 2021. IMU2Doppler: Cross-Modal Domain Adaptation for Doppler-based Activity Recognition Using IMU Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 145 (December 2021), 20 pages. https://doi.org/10.1145/3494994



## 1 INTRODUCTION

Researchers and developers often rely on sensors in smartphones [15, 26], smartwatches [39, 60], cameras [4, 25, 51], and even microphones [43, 59] to infer context, recognize user activities, and adapt to the user's needs. Recently, we have seen many activity recognition systems that rely on Doppler Effect-based mmWave radars to measure activity movements [33, 44, 45]. An advantage of a mmWave radar is its ability to characterize fine-grained motion. It has the ability to capture micro-motion dynamics of subtle activities (*e.g.,* hand activities such as brushing, eating etc.) captured via the *micro-Doppler Effect* [11]. A mmWave radar-based activity recognition system also offers a higher degree of privacy preservation compared to other popular ambient sensors such as cameras or microphones.

Authors' addresses: Sejal Bhalla, sejal@cs.toronto.edu, University of Toronto, Toronto, Ontario, Canada; Mayank Goel, mayankgoel@cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Rushil Khurana, rushil@cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2021 Association for Computing Machinery.
2474-9567/2021/12-ART145 $15.00
https://doi.org/10.1145/3494994

These activity recognition and sensing systems are typically built using machine learning models that need labeled, *in-situ* sensor data for training. These training labels typically rely on manual annotation or user intervention to segment and label specific activities performed by users. This process introduces time and resource constraints that impedes our ability to quickly deploy and use new doppler sensors. Moreover, the ground truth collection (cameras, user intervention etc.) tends to be intrusive and may be unsuitable in scenarios with elevated privacy constraints. While mass production of new hardware is an outstanding challenge [27], **data collection and labeling cost** is one of the biggest challenges to use these new sensors to build an activity recognition system. These systems need to work well out-of-the-box with no or very-little in-situ calibration. Ideally, the machine learning models would not need any *in-situ* training, and ultimately facilitate easier deployability.

One method to overcome the challenge of data labeling cost is automated domain adaptation. Such approaches rely on successful knowledge transfer from labeled data collected in one domain and use it to assist the training of a model in a target domain with no (or limited) data of its own. Here, one popular approach has been to use videos as the source domain [1, 8]. Videos provide a rich source of information with a considerable feature space. Moreover, the extensive library of labeled video datasets make it an attractive choice for a source domain. However, using videos as the source domain requires the full body of a human to be visible in the source videos. The approach cannot handle occlusion or partial capture of the body. This limitation significantly reduces the available video datasets that can be reliably used for domain adaptation.

In this paper, we present *IMU2Doppler* and evaluate the use of off-the-shelf inertial measurement units (IMU) datasets as the source domain to build an activity recognition model for the mmWave radar sensor. IMU data does not share the same limitations as videos. The uninhibited signal that captures the motion performed as a part of the activity is rotation and environment invariant which makes it a good candidate as a source. Additionally, IMU retains some of the advantages of video datasets *i.e.,* prior works in activity recognition have extensively collected IMU data for a gamut of activities and made it publicly available.

We demonstrate that IMU2Dopper can map the doppler data (input to the untrained ML model) to a latent feature representation of the pre-trained IMU model. In addition to this representation of the IMU model, we use minimally labeled (akin to a calibration step) doppler data to classify 10 activities of daily living. This novel approach allows us to recognize these activities with an accuracy of **70% with only 15 seconds of labeled data** from the mmWave radar sensor. We acknowledge that this is not the performance we should expect from a real world system. However, IMU2Doppler provides an out-of-the-box model that can benefit from a quick personalization and calibration step. Our contribution lies in facilitating rapid development of a 'good enough' base model that can then be used with other techniques such as active learning [42] or meta-labeling [14] that personalize to the user's environment and improve over time without a need for significant data labeling.

In this work, we also demonstrate that we can combine multiple IMU datasets recorded in completely different environments with different users as a unified source of training data. Typically, using multiple sources is a significant challenge for domain adaptation due to the domain shift that exists across the sources. Zhao *et al.* have summarized the numerous challenges of multi-source domain adaptation [63]. However, we show that our approach is resilient to such issues. In fact, when we combine the training data from two IMU datasets, IMU2Doppler demonstrated a small increase of 1% in recognition performance. From a practical perspective, it means that not only any publicly shared IMU data can be used, but a user who wishes to record a completely new activity may incur a one-time-cost, use an app on their smartwatch to collect IMU data for that activity, and personalize the machine learning model. Moreover, if the user chooses to share their data of this new activity, other users can leverage it to train their doppler sensor without incurring the same time and resource penalty.

In summary, our contributions are as follows:

(1) An activity recognition system for 10 different activities using mmWave radar. Prior work has shown the use of mmWave radar to capture gross movements. We include and expand the set of activities to include subtle activities such as brushing teeth, folding laundry *etc.*

(2) A novel multi-class heterogeneous domain adaptation approach that learns a feature mapping between inertial sensors worn on a user's wrist and a mmWave radar sensor placed in the environment. It means that our approach is viewpoint and translation invariant.

(3) While we are not the first to use domain adaptation, our work is the first approach that uses off-the-shelf IMU dataset as the source domain to train a doppler sensor. It means that the source data was not only collected on different users as the target domain, but also at a different time. We also show that we can use muliple datasets and combine them as a single source to achieve the same results.

## 2 RELATED WORK

In this section, we first briefly cover recent advances in activity recognition using doppler. Prior work in this area shows the promise of the mmWave radar as a sensor for activity recognition. Next, we expand upon prior work in domain adaptation with a focus on heterogeneity. This is a relatively new space with exciting prospects. We discuss strategies adopted by prior work and how our work builds upon some of these ideas.

### 2.1 Doppler-based Activity Recognition

Radar sensors exhibit numerous advantages such as non-intrusiveness, high distance range, deep penetration, accessibility, inability to sense personally identifiable information and high signal fidelity [9, 10, 23], which make them an appealing solution for human activity recognition. In the past, they have been used for vital signs monitoring such as heart rate and breathing [31], gait patterns analysis [53], gesture sensing [22], person tracking and identification [36], and distinguishing emotions [62]. The superior range resolution of mmWave Radar sensors, which operate in the frequency range of 30GHz and 300GHz, has further enabled the recognition of fine-grained human activities [33, 45, 50, 58]. For instance, Singh et al. [44] used a voxelised representation of sparse point-clouds from mmWave radar to detect five activities. Zhang et al. [61] converted the point cloud data into micro-Doppler spectrograms before using a CNN to classify multiple human actions in real-time. Micro-doppler spectrograms present an effective way to visualize both doppler and micro-doppler shifts. Over the last decade, they have been extensively used for human activity recognition using radar in different contexts [7, 18, 28, 57].

Most of the prior works have focused on large body movements only such as walking and jogging. And even though some of these prior works have collected data for those large body movement based activities, the datasets that are publicly available are sparse and do not contain enough activity labels. Therefore, despite the promise of the doppler sensor, its utility is impeded by the data collection and labeling cost.

Next, we discuss how prior works have tackled the challenge of data labeling using domain adaptation. We examine techniques not only for the doppler sensor but broadly for other modalities as well. We outline how takeaways from some of these techniques ground our approach to build a doppler based activity recognition system from domain adaptation of IMU data.

## 2.2 Heterogeneous Domain Adaptation

Most prior work on domain adaptation assumes that data of different domains are of the same dimensionality or are drawn from the same feature space [5, 35, 65]. Specifically for Human Activity Recognition (HAR), previous works have proposed a range of techniques including self-supervision based unsupervised transfer learning [40], few shot learning [19], substructure-level matching based domain adaptation [34], adaptive spatial-temporal transfer learning (ASTTL) [37] and adversarial domain adaptation frameworks [64]. However, the assumption of homogeneity may not hold for many applications. Consequently, recent work has witnessed a rise in heterogeneous domain adaptation (HDA) techniques, which tackle the incongruity of source and target feature spaces by mapping features into a common and closer subspace [47, 55], or exploiting the correlations between features [6], or directly transforming data from one domain to the other [21, 49]. Although these approaches have shown promising results, they still suffer from challenges. While mapping features into a predefined subspace may lead to the loss of shareable information, feature translators which attempt to synthesize target data that follows source domain distribution (or vice-versa) are domain-specific and often difficult to be constructed in real-world applications. Moreover, most existing HDA methods simply learn multiple binary classifiers by adopting a one-vs-rest strategy to achieve multi-class classification [13, 17, 52]. This hinders the full exploration of the underlying structure among multiple classes in the target domain. While there has been exploration of domain adaptation for human activity recognition, we specifically focus on cross-domain adaptation approaches. Our work in particular looks at using different source and target domains. An additional layer of heterogeneity comes into play when the source and target domains belong to different modalities. Cross-modal domain adaptation approaches have succeeded in transferring knowledge between modalities like vision and sound [3], text and vision [2], vision to inertial data [29, 38] and inertial signals to video [55]. These techniques rely on using a higher dimensional modality to train an activity recognition model in a lower dimensional target domain. Even though these approaches work well, they are limited by their need for paired, synchronous instances in both domains (*e.g.,* [55]). Despite this limitation, there are some key takeaways from these proposed techniques. Most importantly, it is possible to robustly transfer knowledge between two different modalities/domains by learning a shared latent representation. In fact, some of these works were even able to use lower dimensional modalities such as IMU to knowledge transfer onto a higher dimensional modality such as videos.

Next, we focus specifically on what are perhaps the closest prior works to our problem. The following techniques have used different modalities to train the doppler sensor for learning human activities. Vid2Doppler [1] and Cai et al.'s work in RF sensing [8] uses videos, detects and tracks humans in them, reconstruct a 3D mesh and use them to generate a synthetic signal that can be used to train the doppler sensor. In fact, these approaches have been shown to work robustly and accurately without the need for *labeled* paired synchronous data. However, they have a severe limitation. The use and reconstruction of 3D human pose means that the videos that can be used as a source need to have full human body visible without any occlusion. This significantly reduces the size of publicly available labeled datasets that can be used by these two approaches.

Another very successful approach to reduce the data labeling cost for the doppler sensor uses audio as a source modality to teach the doppler sensor [48]. This approach converts the doppler spectograms into pseudo-audio representations using a GAN and then uses a pre-existing sound classifier to classify activities. This approach is an improvement over other approaches since it is neither limited by signal occlusion issues, nor does it require paired synchronous samples. However, the system still requires a larger amount of initial data to build a model that can convert the doppler spectogram into its pseudo-representation. They used a dataset of 1109 spectograms across six activities where each spectogram with each sample collected over a period of 5 seconds. Again, the data labeling cost for a wide range of activities inhibits the use of this approach.

Based on our learnings from prior work, it is clear that generating synthetic data or pseudo-representations may perform better in some scenarios, but it has severe limitations. Therefore, in our work, we build on the

idea of a shared latent feature subspace that shares the knowledge of the source domain while also preserving the target domain characteristics. We achieve the same using a minimal, asynchronously labeled target dataset which is modeled by a multi-objective optimization learning approach that simultaneously constrains the domain confusion and multi-class classification loss, thus overcoming the majority of the challenges outlined in this section.

## 3 ALGORITHM

IMU2Doppler is a transfer learning-assisted ambient sensing system that uses mmWave radar sensors to detect and distinguish between a set of activities of daily living with minimal labeled data. To account for the lack of labeled radar data, we implement a multi-objective optimisation technique that uses domain adaptation. It uses a neural network pre-trained on inertial measurement data from multiple datasets specifically curated for the task of activity recognition. Below, we describe our sensing principle and algorithm in detail.

### 3.1 Sensing Principle

Millimeter-wave (mmWave) radar sensors transmit pulses of electromagnetic energy and receive reflections when obstructed by rigid targets in the environment. By exploiting the Doppler Effect, it is possible to measure certain motion characteristics of the target like its relative velocity, angle of arrival and distance to the radar system. While the Doppler effect arises from the bulk motion of the target, micro-motion dynamics of the target or its structure such as vibration, rotation, tumbling and coning motions induce the *micro-Doppler Effect* [11]. For instance, in case of a moving person, the arms and legs act as independent elements in motion [41]. Since the intensity of the micro-Doppler effect is dependent on the velocity and direction of the motion, individual movements of the target with discernible motion characteristics produce distinct micro-Doppler signatures, which can be used for human activity recognition [7, 61].

We collect the synthetic aperture radar (SAR) data from the doppler sensor and used the azimuth-range-doppler algorithm to parse the continuous data. As shown in Figure 1, the images corresponding to different activities represent distinct patterns. These patterns can be modeled and recognised by appropriate learning algorithms, as described in the following sections.

### 3.2 Knowledge Transfer

Machine learning algorithms show exceptional predictive power in a range of HAR tasks [24, 25, 30, 32] but require an abundance of annotated data. Although such labeled data exists for a number of sensing modalities, the newfound promise of doppler radar sensing is limited by the lack of a sufficiently large labeled dataset. To solve this problem, we use transfer learning, specifically domain adaptation, wherein we can leverage neural networks trained on a sufficiently large dataset of a different but related modality (source domain) to accelerate the learning of micro-doppler signatures (target domain).

The accelerometer data captured by a wearable inertial measurement units (IMU) characterizes similar motion characteristics as that of a doppler sensor. It captures an environment and position invariant snapshot of the motion of human movement. We postulate that this characteristic of IMU makes it a suitable candidate for source domain. Besides heterogeneous domain adaptation across two different modalities, we also use off-the-shelf datasets to demonstrate that the same events do not need to be recorded synchronously for knowledge transfer across different modalities.

For knowledge transfer, we propose a supervised, cross-modal domain adaptation approach that maps the input of the untrained doppler model to the shared latent feature representation of the pre-trained IMU model. Further, to preserve the information about the target domain or doppler data, we adopt multi-task learning to simultaneously minimise the domain discrepancy (between the latent representation of the two modalities) along

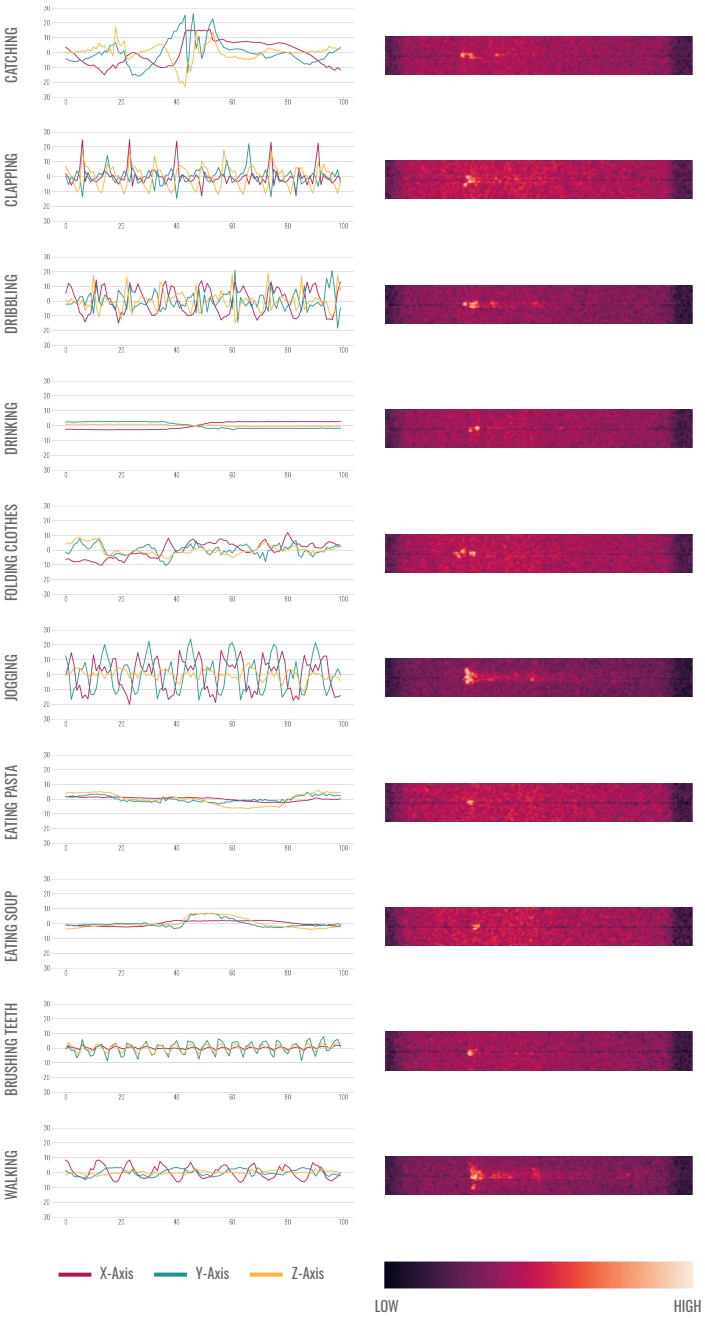



Fig. 1. Corresponding Doppler and IMU signals for various activities.

with the classification loss (between the predicted and actual target label). This multi-objective optimisation ensures that the underlying structure of the target data is retained even in the latent feature subspace. The rationale for our approach is rooted in the observation that task-specific and domain-invariant semantic features can be better associated with the higher layers (close to the output side) of a network [16]. This observation allows us to opt different neural architectures that are best suited for each modality, with the constraint of having identical fully connected layers that are responsible for producing the shared latent feature representation.

### 3.3 IMU: Data Processing and Neural Architecture

We use the "WISDM Smartphone and Smartwatch Activity and Biometrics Dataset"[54] for training an activity recognition classifier with inertial data. The dataset was collected from the accelerometer and gyroscope sensors on both the smartwatch and smartphone of a total of 51 users. It consists of 18 unique activities, ranging from basic ambulation like walking and jogging to other activities of daily living like eating and drinking. For the purpose of our work, we chose a subset of 10 activities for evaluation, as listed in Figure 1. We chose these activities based on their suitability for detection with doppler sensor. For example we did not include activities such as kicking a soccer ball or two other different eating related activities (sandwich, chips). We also excluded activities that do not include any motion such as sitting.

The four streams of data, namely phone accelerometer, phone gyroscope, watch accelerometer and watch gyroscope, are each recorded at a sampling rate of 20 Hz. For our purpose, we consider only the smartwatch accelerometer data since smartphone sensors (placed in the user's pocket) fail to capture hand-oriented/upper body movements. Further, we didn't observe a significant increase in the performance of $M_{IMU}$ on adding smartwatch gyroscope data. Thus, to reduce training overhead and eliminate the requirement of another stream of data for performing the subsequent domain adaptation, we limited our evaluation to smartwatch accelerometer data. We segment the raw watch acceleration data using a sliding window of size 5 secs and an overlap of 2.5 secs. The extracted tri-axial frames are reshaped into 3-channel windows with a length of 100 samples (input size: $100 \times 3$) that are ready for classification.

*3.3.1 IMU Model Selection.* To assess the discriminability of the activities in the source domain, we evaluate the performance of a set of deep neural networks including 1D CNN (5 Convolutional Units, each consisting of a Convolutional layer, a Batch Normalisation layer and a Max Pooling layer), LSTM (2 LSTM layers, Units: [128, 256]) and Bidirectional-LSTM (1 Bi-LSTM layer, Units: 128). One fully-connected layer (Units: 128) and the final output layer were added at the end of each model. The networks were trained from scratch with Adam optimizer (Learning Rate: 0.01) coupled with a learning rate decay of 0.1 (to check for the saturation of validation loss). We use the Categorical Crossentropy loss function to optimise the outputs of the final layer, which uses the Softmax activation to classify activities. We used Keras [12] and Python to implement and train these models.

Table 1. Classification results of different models on a subset of WISDM Dataset (10 activities)

| Model | Trainable Parameters | Accuracy ± SD |
|---|---|---|
| 1D CNN | 453,002 | 79.16 ± 3.35 |
| LSTM | 496,010 | 80.67 ± 2.92 |
| **Bi-LSTM** | 169,354 | **83.34 ± 4.23** |

We followed a subject-independent scheme for evaluation and split the dataset into 5 folds of 10 subjects each. Each train-test split resulted in approximately 28.4K training instances and 7.8K test instances. Table 1 provides the classification performance of all the models along with their total number of trainable parameters. We found

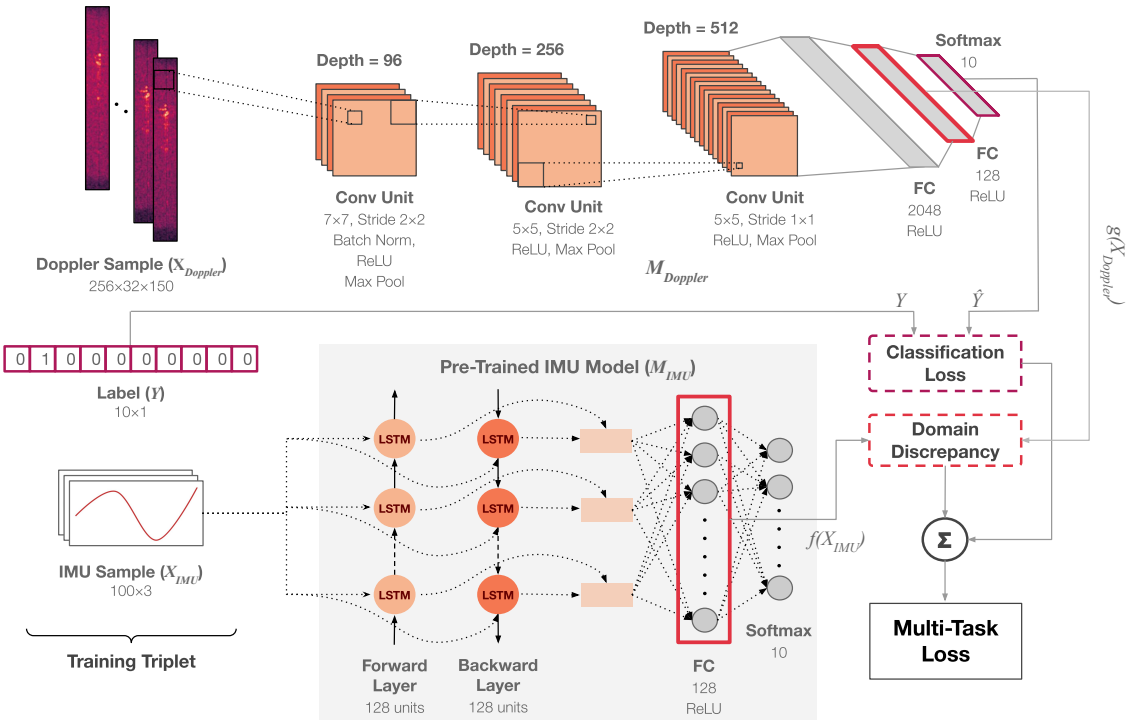



Fig. 2. Training schematic for cross-modal domain adaptation with IMU data as the souce domain and doppler data as the target domain.

the bidirectional LSTM classifier to produce the best accuracy owing to the superiority of Recurrent Neural Networks (RNNs) like LSTMs and Bi-LSTMs in modeling long-range temporal dependencies and. Moreover, since a bidirectional-LSTM layer consists of two LSTM layers that operate on the original and reversed copy of the data in parallel, it preserves the information from both the future and the past, thus outperforming an LSTM. Hence, for all further experiments, we use the Bi-LSTM as the neural architecture for the source domain.

## 3.4 End-to-end Learning Algorithm

The pipeline begins with training a Bi-LSTM classifier, $M_{IMU}$, on the entire IMU dataset. $M_{IMU}$ acts as the pre-trained model used for domain adaptation. The goal is to train a new model, $M_{Doppler}$, for classifying the limited labeled doppler data. Each doppler sample, $X_{Doppler}$, is paired with a random IMU sample, $X_{IMU}$, having the same activity label $Y$, to form training triplets of the form $(X_{IMU}, X_{Doppler}, Y)$. To extract a latent feature representation, we consider the output of the second-last fully connected layer, which has an identical configuration in both $M_{IMU}$ and $M_{Doppler}$. We utilize the output of this layer to minimize the divergence between the source and target domain. Apart from the standard Mean-Squared Error (MSE), the most commonly used divergence measures are Maximum Mean Discrepancy (MMD) [20] and Correlation Alignment (CORAL) [46]. While MMD is a hypothesis test which compares the means of the features after mapping them to Reproducing Kernel Hilbert Space (RKHS), CORAL attempts to align the second-order statistics of the source and target distributions. On comparing the domain adaptation performance of the three metrics, MSE proved to be the most successful in minimizing the domain divergence for our task (see Table 3).

Let the sequential transformation of all the layers before the final layer be denoted by $f(.)$ and $g(.)$ for $M_{IMU}$ and $M_{Doppler}$ respectively. Essentially, f(.) and g(.) map the input sensor data to the output of the pre-final layer of MIMU and MDoppler respectively. The learning objective of $M_{Doppler}$ is to optimise the weighted sum of the mean-squared error between the latent representations, i.e. $|g(X_{Doppler}) - f(X_{IMU})|^2$, and the categorical cross-entropy loss between the predicted softmax values, $\hat{Y}$, and actual label, $Y$. Mathematically, we define our objective function $\mathcal{L}$ as follows:

$$\mathcal{L}(X_{IMU}, X_{Doppler}, Y) = \alpha \times |g(X_{Doppler}) - f(X_{IMU})|^2 + \beta \times - \sum_i Y^{(i)} log \hat{Y}^{(i)}$$

Here, $Y^{(i)}$ and $\hat{Y}^{(i)}$ denote the value of the $i^{th}$ class in the actual and predicted one-hot encoded labels respectively. Adam optimiser (Learning Rate: 0.001), coupled with a learning rate decay of 0.1 (to check for validation loss saturation) is used to optimise $\mathcal{L}$. Empirically, we found the value of $\alpha = 1.3$ and $\beta = 0.7$ to produce the best performing classifier (see Table 4). To further accelerate the learning, we initialise the final layer of $M_{Doppler}$ with weights of the corresponding layer in the pre-trained $M_{IMU}$. In this way, the knowledge in $M_{IMU}$, in the form of learned parameter values and input-output mapping, is effectively transferred to $M_{Doppler}$. The entire training schematic is visualised in Figure 2.

### 3.5 Doppler: Data Processing and Neural Architecture

We apply the azimuth-range-doppler algorithm on our collected doppler dataset to get rolling spectrograms consisting of 256 frequency bins and 32 time steps (representing nearly 0.01s of data). We construct sequences of these spectrograms by using a sliding window of 5s with a step size of 1s. When stacked together, each window consists of 150 (5s × 30 Hz) frames of 256 × 32 spectrograms (final input size: 150 × 256 × 32).

To compare the results of our domain adaptation approach and determine the best neural architecture for the target domain, i.e. doppler data, we trained and evaluated different models chosen from state-of-the-art deep learning architectures that are generally adapted in a wide range of applications. We compared the performance of a 3-layer 2D CNN, a 5-layer 2D CNN, a CNN-LSTM and a CNN-Bi LSTM, on our processed dataset of spectrogram sequences. Standalone LSTMs and Bi-LSTMs can't be considered since the dataset comprises sequences of 2D images, which need to be condensed into sequences of 1D vectors before they can be processed by an LSTM layer. For this purpose, we added a CNN encoder before the first LSTM layer in order to extract the spatial information from each image while modeling the temporal dependencies of a sequence. Thus, both CNN-LSTM and CNN-Bi LSTM consist of a 3-layer time distributed 2D-CNN, followed by 2 LSTM layers (Units: [128, 256]) and 2 Bi-LSTM layers (Units: [128, 256]) respectively. On the other hand, the 5-layer and 3-layer 2D-CNNs just consist of 5 and 3 Convolutional Units (convolution layer and a max pooling layer) respectively. Following the same structure as the IMU model, each model is connected to a 128-unit fully-connected layer, followed by the output layer with Softmax activation for classification. As shown in Section 5.1, the 3-layer 2D-CNN proved to be the optimal neural architecture for modeling our dataset.

## 4 DATA COLLECTION

### 4.1 Participants and Apparatus

We collected data from 9 participants (6 males, 3 females), ranging in age from 20-32 (Mean: 26.3, SD: 3.8). The data was collected in a lab space roughly 5.2 x 6.5 x 2.8 m. We used TI's AWR1642 doppler radar sensor (Figure 3) to record SAR data at a sampling rate of 30 Hz. The sensor was placed at a distance of approximately 2m from the participants. All activities were recorded using a laptop and ground truth was collected with an accompanying video camera.

All activities except clapping, jogging and walking required additional apparatus. The food was packaged in the same takeout containers for each participant. All participants used manual toothbrushes from the same brand.

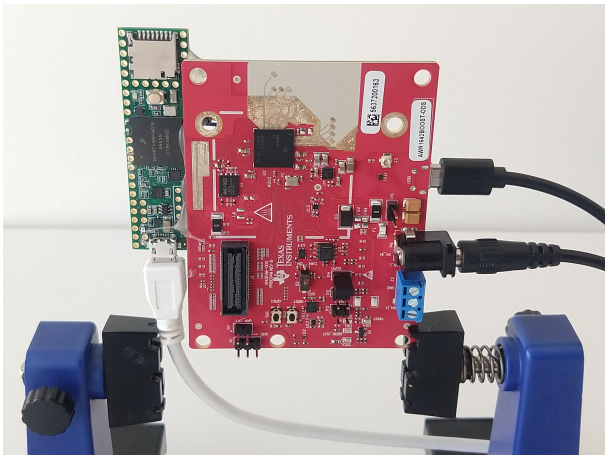


Fig. 3. We used TI's AWR1642 Radar sensor for our data collection.

We provided the participants with a tennis ball and a basketball for the catching and dribbling tasks. We did not control the apparatus for drinking and folding clothes. The participants were given different sized cups out of convenience; and the participants typically folded their own clothes (e.g. jackets). Our data collection may differ from how the WISDM dataset was initially collected. There is no way to exactly replicate the same procedure, we opted to not control intra-class variability for these actions. Instead, we rely on the largely periodic nature of all these activities (time period < window size) such that each data window captures similar movement with a shift in time. This would lead to finding similar features across samples in a given class.

### 4.2 Experimental Design and Protocol

Before beginning the data collection, the researcher introduced the protocol and briefed the participants about the activities. We conducted an extensive within-subject study in which we recorded 1500s of data (10 activities $\times$ 150 seconds) from each participant. For each activity, we divided the recording into 10 sessions of 15s each. We created a custom app with a start/stop button. The app has the ability to record the name of an activity and the session number alongside the doppler data. This allowed us to quickly assign an apppropriate label to each session.

## 5 RESULTS

In this section, we evaluate the performance of the proposed cross-modal domain adaptation approach via extensive experiments on our collected dataset comprising 10 different subjects. Since our dataset is balanced, we have chosen accuracy as our evaluation metric throughout. However, we have reported F1 scores in Table 5 for completion.

### 5.1 Doppler Model Selection/Baselines

We consider a subset of the doppler data, consisting of three subjects, for determining an appropriate baseline. We built three per-user classifiers, one for each subject, for each type of model using a leave-n-sessions-out scheme (n = 9). Here, in each fold, we train the model on one session of data (training set), calibrate the hyperparameters on another session (validation set) and evaluate the performance on the remaining eight sessions (test set). Thus, for this experiment, we obtain a total of 110 training instances (10 activities $\times$ 1 session $\times$ 11 instances/session;

Table 2. Baseline Results for different models across 10 activities and 3 subjects

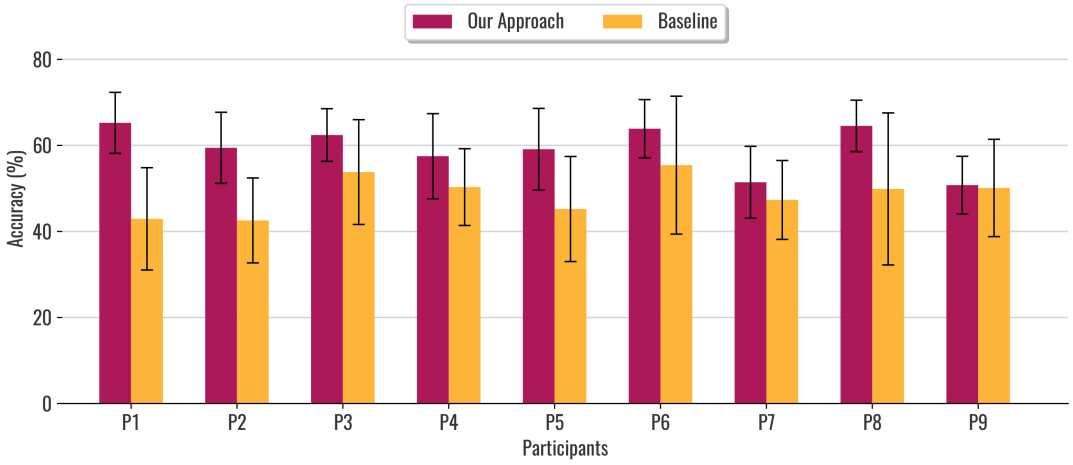
| Model | Input Size | Trainable Parameters | Accuracy ± SD |
|---|---|---|---|
| **3-Layer 2D CNN** | $N \times 256 \times 32 \times 150$ | 11,154,922 | **76.55 ± 5.25** |
| 5-Layer 2D CNN | $N \times 256 \times 32 \times 150$ | 15,874,538 | 68.12 ± 0.01 |
| CNN-LSTM | $N \times 150 \times 256 \times 32 \times 1$ | 4,765,504 | 52.57 ± 3.01 |
| CNN-Bi LSTM | $N \times 150 \times 256 \times 32 \times 1$ | 9,698,624 | 53.94 ± 22.53 |

each session is 15s long per activity), 110 validation instances and 880 test instances. Table 2 summarizes the classification accuracies of all models, the number of associated trainable parameters and the required shape of the input data. Despite the limited training data, the 3-layer CNN produces an accuracy of 76.55%, thereby outperforming the rest. In fact, we can generalise that for our task, CNNs perform significantly better than the CRNNs (Convolutional-Recurrent Neural Networks; CNN-LSTM and CNN-Bi LSTM). The superiority of CNNs can be attributed to the way in which the input data is modeled by the two architectures. While both use convolutional layers as feature extractors, the CNNs interpret the entire sequence as a multi-channel image (stacked spectrograms), unlike the CRNNs, which treat each frame of the sequence individually before fusing the extracted features and passing them through an LSTM. The former allows a more comprehensive representation of the sequence by systematically organizing the temporal information as spatial neighbours. The latter, on the other hand, diminishes the local intra-frame temporal dependencies. Lastly, additional convolutional layers in the 5 layer network make the model unnecessarily complex for a small dataset, thus leading to overfitting. Therefore, we find a 3-layer CNN to be the most suitable for effectively modeling doppler spectrograms. We deploy our domain adaptation approach on the same to compare against the best-performing baseline.

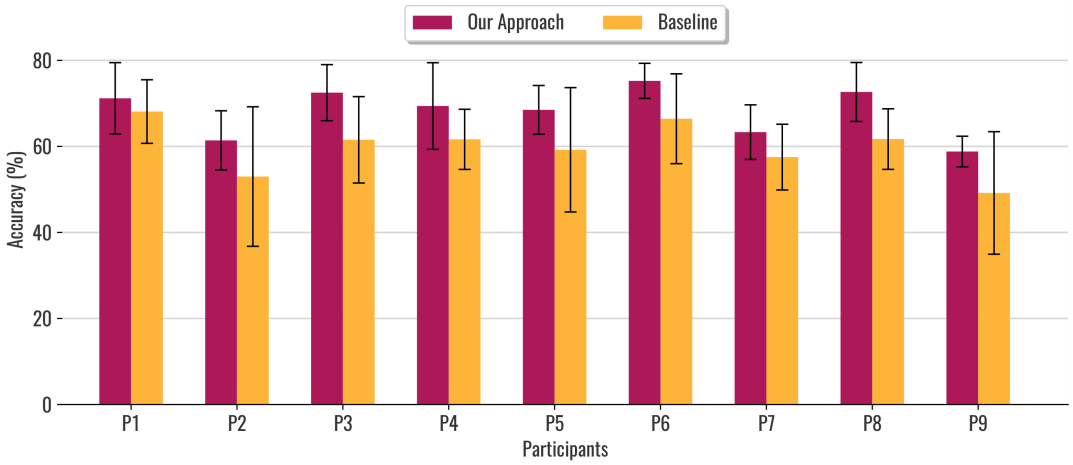
## 5.2 Domain Adaptation Results

Finally, we evaluate the performance of our proposed approach under varying conditions. We primarily vary the amount of labeled data used in the learning procedure with the objective of navigating the tradeoff between minimizing the amount of annotated data learned by the model and increasing the resultant performance. Starting from 10 seconds of data per class, we examine different sizes of annotated data uptil 30 seconds. With increments of 5s, we obtain a total of 5 durations: 10s, 15s (or 14s), 20s, 25s, and 30s. Each of these durations entail different combinations of training and validation size (see Table 5), represented by (T, V), where T denotes the training size per activity (in seconds) and V denotes the validation size per activity (in seconds). For instance, a model can be exposed to 20s of labeled data in two ways: (15s, 5s) or (10s, 10s). Further, if an entire session is not consumed in the training or validation set, we discard the remainder to prevent information leakage. This ensures that no two windows belonging to the same session are present in two different sets.

We adopt a leave-n-sessions-out scheme to train the models, where $n$ represents the number of sessions that are not a part of the training set. $n$ can take different values depending on the size of the training set. For example, if we consider a training size that is greater than the session size (15s), say 20s, we will require 2 sessions for training. This leaves us with 1 session for validation and 7 sessions for the test set ($n$ = 8), thus leading to a total of ${}^{10}C_8$ combinations of train-validation-test sets. On the other hand, with a training size of 10s, one session would suffice for training ($n$ = 9) and we'll obtain ${}^{10}C_9$ train-validation-test sets. Each set was trained for a maximum of 500 epochs with Adam optimizer (Learning Rate: 0.01) coupled with a learning rate decay of 0.1 and early stopping on the validation set with a patience of 100. The average of the results across ${}^{10}C_n$ runs is reported for each (T, V) configuration.




(a) (5s, 5s)




(b) (10s, 5s)

Fig. 4. Per-user comparison of our approach and the baseline under two labeled data distributions consisting of 5s and 10s of training data per class respectively, combined with 5s of validation data per class. The error bar indicates the variation (Standard Deviation) across different folds.

Before proceeding with domain adaptation, we performed a comparative analysis of the domain divergence metrics and the parametric distribution of the loss coefficients. We evaluated the performance of three metrics, namely MSE, MMD and CORAL, that attempt to measure and optimise the divergence between the source and target domains (see Section 3.4 for more details). We also analysed the distribution of $\alpha$ and $\beta$, which are the coefficients of the domain discrepancy loss and classification loss respectively. For this purpose, we chose their values such that each ($\alpha$, $\beta$) pair sums up to 2.0 and each value lies in the range of [0.5, 1.5] to check for redundancy

Table 3. Domain Adaptation Results for different domain divergence metrics (loss functions) across 10 activities and 5 participants.

| | **MMD** | **CORAL** | **MSE** |
|---|---|---|---|
| **Accuracy ± SD** | 64.27 ± 5.65 | 67.00 ± 5.05 | **68.73 ± 4.43** |

Table 4. Domain Adaptation Results for different values of loss coefficients ($\alpha$, $\beta$) across 10 activities and 5 participants.

| **($\alpha$, $\beta$)** | **Accuracy ± SD** |
|---|---|
| (0.5, 1.5) | 67.88 ± 4.69 |
| (0.6, 1.4) | 65.61 ± 5.36 |
| (0.7, 1.3) | 67.21 ± 3.96 |
| (0.8, 1.2) | 65.60 ± 4.47 |
| (0.9, 1.1) | 64.31 ± 2.98 |
| (1.0, 1.0) | 64.62 ± 4.65 |
| (1.1, 0.9) | 65.91 ± 4.31 |
| (1.2, 0.8) | 65.17 ± 3.23 |
| (1.3, 0.7) | **68.73 ± 4.43** |
| (1.4, 0.6) | 66.38 ± 4.06 |
| (1.5, 0.5) | 66.15 ± 2.53 |

Table 5. Classification accuracy and F1 score of our proposed domain adaptation approach and baseline for different amounts of training and validation data, averaged across 9 subjects.

| **Amount of Labeled Data** | **Configuration (Training, Validation)** | **Our Approach** | | **Baseline** | |
|---|---|---|---|---|---|
| | | **Accuracy** | **F1** | **Accuracy** | **F1** |
| 10s | (5s, 5s) | 59.36 | 0.5839 | 48.61 | 0.4757 |
| 15s | (7s, 7s) | 70.00 | 0.6847 | 64.16 | 0.6267 |
| | (10s, 5s) | 68.18 | 0.6639 | 59.81 | 0.5787 |
| 20s | (10s, 10s) | 72.68 | 0.7178 | 68.01 | 0.6702 |
| | (15s, 5s) | 71.80 | 0.7054 | 66.67 | 0.6465 |
| 25s | (15s, 10s) | 74.58 | 0.7319 | 70.03 | 0.6856 |
| | (20s, 5s) | 74.46 | 0.7330 | 69.28 | 0.6738 |
| 30s | (15s, 15s) | 75.68 | 0.7512 | 72.55 | 0.7143 |
| | (20s, 10s) | 77.15 | 0.7603 | 73.18 | 0.7165 |

(scaling a loss by a factor less than 0.5 would significantly diminish its contribution to the overall loss). Both the experiments were carried out for 5 participants and an arbitrary (T, V) configuration, (10s, 5s) in this case. As shown in Table 3 and Table 4, MSE loss and the ($\alpha$, $\beta$) value of (1.3, 0.7) produce the best performing classifier, which is used for all further experiments.

Different (T, V) configurations imply varying numbers of training and validation instances. After following the window segmentation procedure described in Section 3.5, we obtain 1, 3, 6, 11, 12, and 17 instances from a total of 5s, 7s, 10s, 15s, 20s, and 25s of data respectively. To account for the limited training samples in case of (5s, 5s)

and (7s, 7s), we augment the doppler dataset by pairing each doppler sample with three IMU samples in order to get three training triplets, $(X_{Doppler}, X_{IMU\_a}, Y)$, $(X_{Doppler}, X_{IMU\_b}, Y)$, and $(X_{Doppler}, X_{IMU\_c}, Y)$. As a result, we have 3 and 9 instances per activity for (5s, 5s) and (7s, 7s) respectively.

We initialized the final layer of $M_{Doppler}$ with the weights of the corresponding yer of $M_{IMU}$. When freezing them throughout training, we obtain an average accuracy of 64.65% for the (10s, 5s) configuration across all participants. However, tuning the parameters while training leads to a higher overall accuracy in a subject-dependent model. Table 5 summarizes the domain adaptation results for different (T, V) configurations and their corresponding baseline results. It shows convincing evidence that micro-doppler based human activity classifiers can learn from the knowledge of a pre-trained IMU model. All configurations show a jump of at least 3% from the baseline, with maximum difference observed in the lower training and validation sizes. The (5s, 5s), (7s, 7s) and (10s, 5s) configurations show an improvement of approximately 10%, 6% and 9% over the baseline respectively. To verify the statistical significance of these results, we conducted a *paired-samples t-test* to compare the activity recognition performance with and without domain adaptation for all configurations. **We found a significant increase in the performance after applying the proposed domain adaptation approach with $p < 0.05$ and Cohen's $d > 1.0$ for each (T, V) configuration.**

A deeper look at the class-wise performance of the activities for the (10s, 5s) configuration (see Figure 5 and Table 6) shows that our model confuses the most between eating soup, eating pasta and drinking, thus classifying them with the least accuracy. These activities can be broadly represented by a similar hand-to-mouth gesture, thus showing limited distinction in both the source and the target domain. Nevertheless, this result is quite encouraging as the difference in recognition performance of our proposed approach from the baseline is considerable in spite of the heterogeneity of the source and target domains. As anticipated, the classification accuracy increases with an increase in the amount of annotated data used for training. The classifier trained on 20s and validated on 10s of data per activity produces the best average accuracy of 77.15% (max participant accuracy: 85.47%), highlighting the role of proportionately distributing our minimally labeled data into training and validation.

The results of these experiments substantially support the feasibility and effectiveness of the proposed supervised domain adaptation approach. Demonstrated over a range of locomotion and other complex daily activities, domain transformed micro-doppler representations are seen to better capture motion information in comparison with the original micro-doppler spectrograms.

Table 6. Class-wise performance of our approach and the baseline for the (10s, 5s) configuration. The table represents the combined results of all participants.

| Activity | F1-Score | |
|---|---|---|
| | Baseline | Our Approach |
| Catching | 0.7201 | 0.8035 |
| Clapping | 0.6445 | 0.7214 |
| Dribbling | 0.8062 | 0.8615 |
| Drinking | 0.4109 | 0.4735 |
| Folding Clothes | 0.7926 | 0.8679 |
| Jogging | 0.7444 | 0.8327 |
| Eating Pasta | 0.3640 | 0.4438 |
| Eating Soup | 0.3453 | 0.5089 |
| Brushing | 0.4502 | 0.5418 |

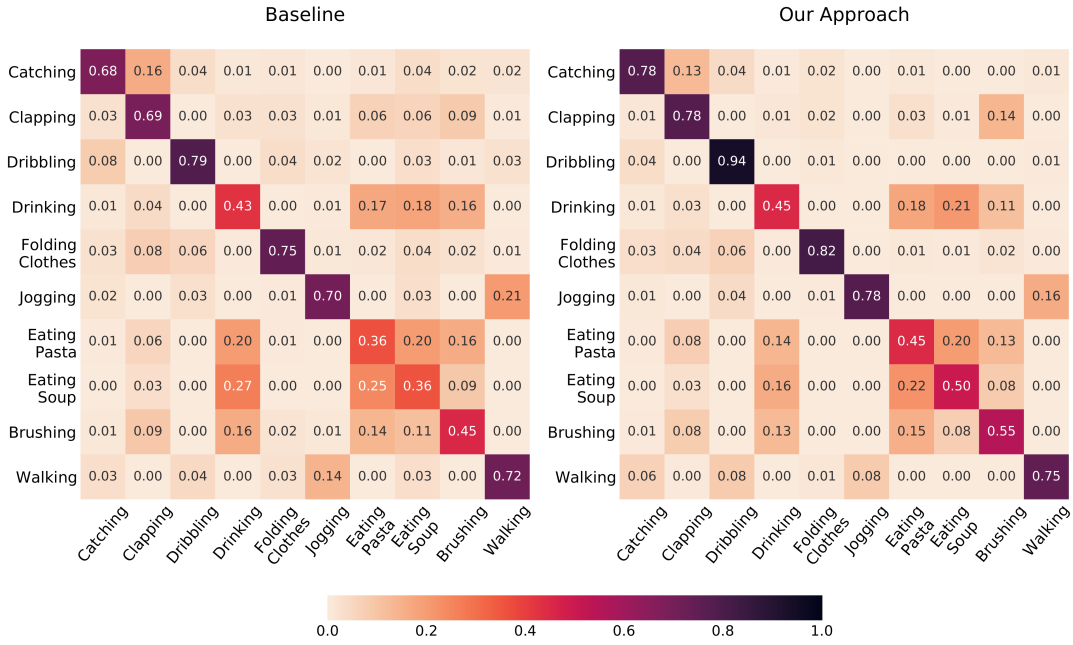



Fig. 5. Confusion matrices for a 3-layer 2D CNN trained with 10s of training and 5s of validation data, without and with domain adaptation. The figure represents the combined results of all participants.

## 5.3 Domain Adaptation Using Multiple Datasets Combined as a Single Source

Although our approach relies on the transfer of higher-level domain-invariant features, we verify the same by distilling information from multiple datasets combined into a single source domain. We constructed a new IMU dataset by replacing the data of two activities in our current dataset, namely *walking* and *jogging*, by corresponding samples drawn from the Wearable Activity Recognition Dataset (WARD) [56]. WARD consists of sequences of 13 human actions (including walking and jogging) collected from 20 participants by a network of 5 sensors placed at different body positions (including the wrist), each carrying a triaxial accelerometer and a biaxial gyroscope. In order to maintain class balance in the proposed dataset, we randomly selected 20 participants from our current dataset before combining it with the wrist accelerometer data from WARD. Due to the inconsistency in the sensor specifications, participants and other external factors, we normalised the mixed dataset to account for the incompatible data distributions across activities from the two datasets. This was done by normalizing (min-max scaling) the two datasets individually so that all values would be comparable (since they lie in the range of 0 to 1) and unaffected by the minimum and maximum of the other dataset.

We trained a Bidirectional LSTM, which proved to be the best classifier for IMU data, from scratch for the mixed dataset. Adopting the same training procedure as followed in Section 3.3.1, we achieved an accuracy of 80.36%, which is at par with the results for a homogeneous IMU dataset. Using this model for extracting learned latent feature representations, we evaluated the performance of our domain adaptation approach with a training and validation size of 10s and 5s, respectively. With an average **per-user accuracy of 69.37%** (see Figure 6), the results not only indicate invariance to heterogeneity in the source domain but also show a marginal increase in comparison to the previous results together with a high accuracy/classifiability for the classes belonging to the minority dataset (walking, jogging). Thus, the results of this experiment highlight the potential of leveraging

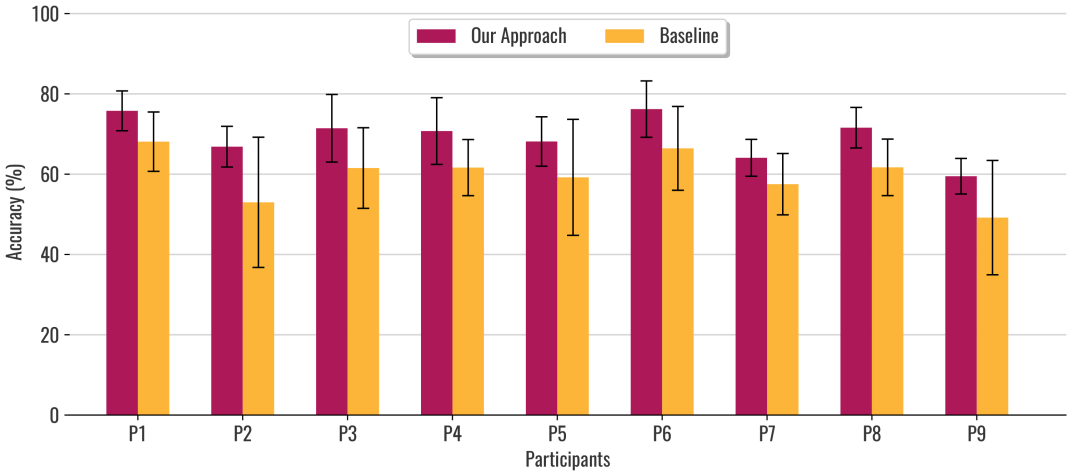



Fig. 6. Per-user comparison of baseline and proposed domain adaptation approach with a heterogeneous source domain comprising of two datasets. The training and validation data contain 15s of labeled data per class (10s, 5s). The error bar indicates the variation (Standard Deviation) across different folds.

the sizeable collection of IMU datasets that cover a multitude of human actions, to build a more comprehensive human activity classifier for doppler data.

## 6 LIMITATIONS AND DISCUSSION

In this section, we discuss some key limitations of our work and reflect on how it may impact the usability and deployability of our approach. We also discuss how our work may contribute to future research directions.

### 6.1 Classifier Accuracy for Real World Use

Our work demonstrates success in domain adaptation and is able to outperform the baseline consistently. However, even with 30 seconds of labeled data, our approach is able to classify these 10 activities with an accuracy of 77.15% (compared to 73.18% with baseline). We acknowledge that this is not sufficient for a system to be deployed in the real world. However, we believe that our approach can be used in combination with other strategies such as meta-labeling [14] and active learning [42] designed to improve the classifier accuracy and robustness over time. Such approaches typically require a 'good enough' base model that can be used to make initial, out-of-the-box predictions. However, building that base model is also not easy without significant labeled data. Our approach can assist in rapidly building these base models to facilitate these techniques that can learn and improve over time without introducing significant data labeling cost.

### 6.2 Limited Activities in Source Domain

Our work shows that we can use existing off-the-shelf IMU datasets to train a mmWave radar sensor. While our approach is robust, the multi-task learning method can only be leveraged to train the mmWave radar with activities that are distinctively recognizable in the source domain. Since all eating-related activities (eating chips, eating sandwich, eating pasta, eating soup) display high similarity in both the source and target domain, we chose a representative set, consisting of eating pasta and eating soup, for evaluation. Moreover, activities like climbing stairs and kicking a ball are out of scope for ambient sensing modalities since the user may not be in

the frame of sensing at all times. Thus, our solution is not a catch-all but despite this limitation, a vast body of prior IMU work means that our approach can be a catalyst to improve deployability of mmWave radar sensor for activity recognition. In fact, our work can potentially leverage prior work to convert the extensive video datasets into virtual IMU streams [29] and then use those virtual IMU streams to train the doppler sensor to recognize a wide gamut of activities.

### 6.3 Controlled Environment for User Study

Despite promising results, one key limitation of our work is that the study was conducted in a controlled environment. The users were free to perform the actions/activities as they normally would but they were recorded in a largely static environment. There were no other motions except the primary user in the field of view of the mmWave radar sensor. The source domain (IMU) is impervious to this challenge, but the doppler sensor captures a wide range of motions in the environment. This limitation needs to be overcome before our work can be deployed widely. Fortunately, newer doppler radar sensors are bundled with person tracking algorithms [1] that can be leveraged to sample the doppler from the primary user. Secondly, our work can still be used in scenarios where only a single user with (mostly) static background would be expected. For example, a small office, single-owned apartment or a home gym.

## 7 CONCLUSION

A fundamental challenge of scaling up any machine-learning system, especially activity recognition systems has been collecting and labeling the data required to train a model. Every few years there is a new sensor in the market that shows promise either due to the signal it is able to capture or advancements in the software and compute capabilities (*e.g.,* surge of computer vision in recent years). In this paper, we tackle the challenge of data collection and labeling with the new and promising mmWave radar sensor. We showcase that we can use existing IMU datasets to learn a latent feature representation that can be used by the mmWave radar sensor to classify between 10 activities with minimal data labeling of its own data (10 seconds).

Our approach not only demonstrates successful heterogeneous domain adaptation, but importantly also works with off-the-shelf datasets. From a real world perspective, it means that not only existing IMU datasets can be used to train the mmWave radar sensor, we can catalogue and label a library of activities recorded using IMU-ladden smartwatches which can be then be used to train sensors such as the mmWave radar. This is an improvement over simply collecting and labeling doppler data because: (1) mmWave radar sensors are not widely adopted or used which makes it hard to do a large data collection; and (2) it is harder to collect the ground truth required for doppler data as it would potentially require cameras (and video coders) or dedicated user time in front of the doppler for direct labeling. On the other hand, smartwatches are popular with a large user base and they have the capability to passively sense, record and label activities with minimal user disruption.

## REFERENCES

[1] Karan Ahuja, Yue Jiang, Mayank Goel, and Chris Harrison. 2021. Vid2Doppler: Synthesizing Doppler Radar Data from Videos for Training Privacy-Preserving Activity Recognition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 292, 10 pages. https://doi.org/10.1145/3411764.3445138

[2] Yusuf Aytar, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2018. Cross-Modal Scene Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 10 (Oct. 2018), 2303–2314. https://doi.org/10.1109/TPAMI.2017.2753232

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems* 29 (2016), 892–900.

[1] https://www.ti.com/tool/TIDEP-01000

[4] Abdelkareem Bedri, Diana Li, Rushil Khurana, Kunal Bhuwalka, and Mayank Goel. 2020. Fitbyte: Automatic diet monitoring in unconstrained situations using multimodal sensing on eyeglasses. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[5] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Mach Learn* 79, 1-2 (May 2010), 151–175. https://doi.org/10.1007/s10994-009-5152-4

[6] John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain Adaptation with Structural Correspondence Learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)*. Association for Computational Linguistics, USA, 120–128. event-place: Sydney, Australia.

[7] Bahri Cagliyan and Sevgi Zubeyde Gurbuz. 2015. Micro-Doppler-Based Human Activity Classification Using the Mote-Scale BumbleBee Radar. *IEEE Geosci. Remote Sensing Lett.* 12, 10 (Oct. 2015), 2135–2139. https://doi.org/10.1109/LGRS.2015.2452946

[8] Hong Cai, Belal Korany, Chitra R Karanam, and Yasamin Mostofi. 2020. Teaching RF to Sense without RF Training Measurements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22.

[9] Haiquan Chen and Wenbin Ye. 2020. Classification of Human Activity Based on Radar Signal Using 1-D Convolutional Neural Network. *IEEE Geosci. Remote Sensing Lett.* 17, 7 (July 2020), 1178–1182. https://doi.org/10.1109/LGRS.2019.2942097

[10] Qingchao Chen, Bo Tan, Karl Woodbridge, and Kevin Chetty. 2018. Doppler Based Detection of Multiple Targets in Passive Wi-Fi Radar Using Underdetermined Blind Source Separation. In *2018 International Conference on Radar (RADAR)*. IEEE, Brisbane, QLD, 1–6. https://doi.org/10.1109/RADAR.2018.8557324

[11] V.C. Chen, Fayin Li, Shen-Shyang Ho, and H. Wechsler. 2006. Micro-doppler effect in radar: phenomenon, model, and simulation study. *IEEE Trans. Aerosp. Electron. Syst.* 42, 1 (Jan. 2006), 2–21. https://doi.org/10.1109/TAES.2006.1603402

[12] Francois Chollet et al. 2015. *Keras*. https://github.com/fchollet/keras

[13] Wenyuan Dai, Yuqiang Chen, Gui-Rong Xue, Qiang Yang, and Yong Yu. 2008. Translated Learning: Transfer Learning across Different Feature Spaces. In *Proceedings of the 21st International Conference on Neural Information Processing Systems (NIPS'08)*. Curran Associates Inc., Red Hook, NY, USA, 353–360. event-place: Vancouver, British Columbia, Canada.

[14] Marcos Lopez De Prado. 2018. *Advances in financial machine learning*. John Wiley & Sons.

[15] Gulustan Dogan, Iremnaz Cay, Sinem Sena Ertas, Şeref Recep Keskin, Nouran Alotaibi, and Elif Sahin. 2020. Where are you? Human activity recognition with smartphone sensor data. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 301–304.

[16] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14)*. JMLR.org, I–647–I–655. event-place: Beijing, China.

[17] Lixin Duan, Dong Xu, and Ivor W. Tsang. 2012. Learning with Augmented Features for Heterogeneous Domain Adaptation. In *Proceedings of the 29th International Coference on International Conference on Machine Learning (ICML'12)*. Omnipress, Madison, WI, USA, 667–674. event-place: Edinburgh, Scotland.

[18] Dustin P. Fairchild and Ram M. Narayanan. 2016. Multistatic micro-doppler radar for determining target orientation and activity classification. *IEEE Trans. Aerosp. Electron. Syst.* 52, 1 (Feb. 2016), 512–521. https://doi.org/10.1109/TAES.2015.130595

[19] Siwei Feng and Marco F. Duarte. 2019. Few-shot learning-based human activity recognition. *Expert Systems with Applications* 138 (2019), 112782. https://doi.org/10.1016/j.eswa.2019.06.070

[20] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. 2006. A Kernel Method for the Two-Sample-Problem. *Arxiv preprint arXiv:0805.2368*, 513–520.

[21] Maayan Harel and Shie Mannor. 2011. Learning from Multiple Outlooks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML'11)*. Omnipress, Madison, WI, USA, 401–408. event-place: Bellevue, Washington, USA.

[22] Souvik Hazra and Avik Santra. 2019. Short-Range Radar-Based Gesture Recognition System Using 3D CNN With Triplet Loss. *IEEE Access* 7 (2019), 125623–125633. https://doi.org/10.1109/ACCESS.2019.2938725

[23] Yuan He, Yang Yang, Yue Lang, Danyang Huang, Xiaojun Jing, and Chunping Hou. 2018. Deep Learning based Human Activity Classification in Radar Micro-Doppler Image. In *2018 15th European Radar Conference (EuRAD)*. IEEE, Madrid, 230–233. https://doi.org/10.23919/EuRAD.2018.8546615

[24] Shian-Ru Ke, Hoang Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. 2013. A Review on Video-Based Human Activity Recognition. *Computers* 2, 2 (June 2013), 88–131. https://doi.org/10.3390/computers2020088

[25] Rushil Khurana, Karan Ahuja, Zac Yu, Jennifer Mankoff, Chris Harrison, and Mayank Goel. 2018. GymCam: Detecting, recognizing and tracking simultaneous exercises in unconstrained scenes. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–17.

[26] Rushil Khurana and Mayank Goel. 2020. Eyes on the Road: Detecting Phone Usage by Drivers Using On-Device Cameras. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.

[27] Rushil Khurana and Steve Hodges. 2020. Beyond the Prototype: Understanding the Challenge of Scaling Hardware Device Production. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.

[28] Youngwook Kim and Taesup Moon. 2016. Human Detection and Activity Classification Based on Micro-Doppler Signatures Using Deep Convolutional Neural Networks. *IEEE Geosci. Remote Sensing Lett.* 13, 1 (Jan. 2016), 8–12. https://doi.org/10.1109/LGRS.2015.2491329

[29] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. 2020. IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.

[30] Oscar D. Lara and Miguel A. Labrador. 2013. A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Commun. Surv. Tutorials* 15, 3 (2013), 1192–1209. https://doi.org/10.1109/SURV.2012.110112.00192

[31] Wenda Li, Bo Tan, and Robert Piechocki. 2018. Passive Radar for Opportunistic Monitoring in E-Health Applications. *IEEE J. Transl. Eng. Health Med.* 6 (2018), 1–10. https://doi.org/10.1109/JTEHM.2018.2791609

[32] Dawei Liang and Edison Thomaz. 2019. Audio-Based Activities of Daily Living (ADL) Recognition with Large-Scale Acoustic Embeddings from Online Videos. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1 (March 2019), 1–18. https://doi.org/10.1145/3314404

[33] Yier Lin, Julien Le Kernec, Shufan Yang, Francesco Fioranelli, Olivier Romain, and Zhiqin Zhao. 2018. Human Activity Classification With Radar: Optimization and Noise Robustness With Iterative Convolutional Neural Networks Followed With Random Forests. *IEEE Sensors J.* 18, 23 (Dec. 2018), 9669–9681. https://doi.org/10.1109/JSEN.2018.2872849

[34] Wang Lu, Yiqiang Chen, Jindong Wang, and Xin Qin. 2021. Cross-domain activity recognition via substructural optimal transport. *Neurocomputing* 454 (2021), 65–75. https://doi.org/10.1016/j.neucom.2021.04.124

[35] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* 22, 10 (Oct. 2010), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

[36] Jacopo Pegoraro, Francesca Meneghello, and Michele Rossi. 2021. Multiperson Continuous Tracking and Identification From mm-Wave Micro-Doppler Signatures. *IEEE Trans. Geosci. Remote Sensing* 59, 4 (April 2021), 2994–3009. https://doi.org/10.1109/TGRS.2020.3019915

[37] Xin Qin, Yiqiang Chen, Jindong Wang, and Chaohui Yu. 2019. Cross-Dataset Activity Recognition via Adaptive Spatial-Temporal Transfer Learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 148 (Dec. 2019), 25 pages. https://doi.org/10.1145/3369818

[38] Valentin Radu and Maximilian Henne. 2019. Vision2sensor: Knowledge transfer across sensing modalities for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–21.

[39] Gabriel Reyes, Jason Wu, Nikita Juneja, Maxim Goldshtein, W Keith Edwards, Gregory D Abowd, and Thad Starner. 2018. Synchrowatch: One-handed synchronous smartwatch gestures using correlation and magnetic sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–26.

[40] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-Task Self-Supervised Learning for Human Activity Detection. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 61 (June 2019), 30 pages. https://doi.org/10.1145/3328932

[41] Linda Senigagliesi, Gianluca Ciattaglia, Adelmo De Santis, and Ennio Gambi. 2020. People Walking Classification Using Automotive Radar. *Electronics* 9, 4 (March 2020), 588. https://doi.org/10.3390/electronics9040588

[42] Burr Settles. 2009. Active learning literature survey. (2009).

[43] Liu Sicong, Zhou Zimu, Du Junzhao, Shangguan Longfei, Jun Han, and Xin Wang. 2017. Ubiear: Bringing location-independent sound awareness to the hard-of-hearing people with smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–21.

[44] Akash Deep Singh, Sandeep Singh Sandha, Luis Garcia, and Mani Srivastava. 2019. RadHAR: Human Activity Recognition from Point Clouds Generated through a Millimeter-wave Radar. In *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems - mmNets'19*. ACM Press, Los Cabos, Mexico, 51–56. https://doi.org/10.1145/3349624.3356768

[45] Thomas Stadelmayer, Markus Stadelmayer, Avik Santra, Robert Weigel, and Fabian Lurz. 2020. Human Activity Classification Using mm-Wave FMCW Radar by Improved Representation Learning. In *Proceedings of the 4th ACM Workshop on Millimeter-Wave Networks and Sensing Systems*. ACM, London United Kingdom, 1–6. https://doi.org/10.1145/3412060.3418430

[46] Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of Frustratingly Easy Domain Adaptation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (Phoenix, Arizona) *(AAAI'16)*. AAAI Press, 2058–2065.

[47] Ben Tan, Yu Zhang, Sinno Jialin Pan, and Qiang Yang. 2017. Distant Domain Transfer Learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17)*. AAAI Press, 2604–2610. event-place: San Francisco, California, USA.

[48] Kimberly T. Tran, Lewis D. Griffin, Kevin Chetty, and Shelly Vishwakarma. 2020. Transfer Learning from Audio Deep Learning Models for Micro-Doppler Activity Recognition. In *2020 IEEE International Radar Conference (RADAR)*. IEEE, Washington, DC, USA, 584–589. https://doi.org/10.1109/RADAR42522.2020.9114643

[49] Yao-Hung Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang. 2016. Learning Cross-Domain Landmarks for Heterogeneous Domain Adaptation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, 5081–5090. https://doi.org/10.1109/CVPR.2016.549

[50] Prachi Vaishnav and Avik Santra. 2020. Continuous Human Activity Classification With Unscented Kalman Filter Tracking Using FMCW Radar. *IEEE Sens. Lett.* 4, 5 (May 2020), 1–4. https://doi.org/10.1109/LSENS.2020.2991367

[51] Philipp Voigt, Matthias Budde, Erik Pescara, Manato Fujimoto, Keiichi Yasumoto, and Michael Beigl. 2018. Feasibility of human activity recognition using wearable depth cameras. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*. 92–95.

[52] Chang Wang and Sridhar Mahadevan. 2011. Heterogeneous Domain Adaptation Using Manifold Alignment. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two (IJCAI'11)*. AAAI Press, 1541–1546. event-place: Barcelona, Catalonia, Spain.

[53] Fang Wang, Marjorie Skubic, Marilyn Rantz, and Paul E. Cuddihy. 2014. Quantitative Gait Measurement With Pulse-Doppler Radar for Passive In-Home Gait Assessment. *IEEE Trans. Biomed. Eng.* 61, 9 (Sept. 2014), 2434–2443. https://doi.org/10.1109/TBME.2014.2319333

[54] Gary M. Weiss, Kenichi Yoneda, and Thaier Hayajneh. 2019. Smartphone and Smartwatch-Based Biometrics Using Activities of Daily Living. *IEEE Access* 7 (2019), 133190–133202. https://doi.org/10.1109/ACCESS.2019.2940729

[55] Tianwei Xing, Sandeep Singh Sandha, Bharathan Balaji, Supriyo Chakraborty, and Mani Srivastava. 2018. Enabling Edge Devices that Learn from Each Other: Cross Modal Training for Activity Recognition. In *Proceedings of the 1st International Workshop on Edge Systems, Analytics and Networking*. ACM, Munich Germany, 37–42. https://doi.org/10.1145/3213344.3213351

[56] Allen Y Yang, Roozbeh Jafari, S Shankar Sastry, and Ruzena Bajcsy. 2009. Distributed recognition of human actions using wearable motion sensor networks. *Journal of Ambient Intelligence and Smart Environments* 1, 2 (2009), 103–115.

[57] Youngwook Kim and Hao Ling. 2009. Human Activity Classification Based on Micro-Doppler Signatures Using a Support Vector Machine. *IEEE Trans. Geosci. Remote Sensing* 47, 5 (May 2009), 1328–1337. https://doi.org/10.1109/TGRS.2009.2012849

[58] Matthew Zenaldin and Ram M. Narayanan. 2016. Radar micro-Doppler based human activity classification for indoor and outdoor environments, Kenneth I. Ranney and Armin Doerry (Eds.). Baltimore, Maryland, United States, 98291B. https://doi.org/10.1117/12.2228397

[59] Cheng Zhang, Qiuyue Xue, Anandghan Waghmare, Sumeet Jain, Yiming Pu, Sinan Hersek, Kent Lyons, Kenneth A Cunefare, Omer T Inan, and Gregory D Abowd. 2017. Soundtrak: Continuous 3d tracking of a finger using active acoustics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–25.

[60] Cheng Zhang, Junrui Yang, Caleb Southern, Thad E Starner, and Gregory D Abowd. 2016. WatchOut: extending interactions on a smartwatch with inertial sensing. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers*. 136–143.

[61] Renyuan Zhang and Siyang Cao. 2019. Real-Time Human Motion Behavior Detection via CNN Using mmWave Radar. *IEEE Sens. Lett.* 3, 2 (Feb. 2019), 1–4. https://doi.org/10.1109/LSENS.2018.2889060

[62] Mingmin Zhao, Fadel Adib, and Dina Katabi. 2016. Emotion recognition using wireless signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, New York City New York, 95–108. https://doi.org/10.1145/2973750.2973762

[63] Sicheng Zhao, Bo Li, Pengfei Xu, and Kurt Keutzer. 2020. Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv preprint arXiv:2002.12169* (2020).

[64] Zhijun Zhou, Yingtian Zhang, Xiaojing Yu, Panlong Yang, Xiang-Yang Li, Jing Zhao, and Hao Zhou. 2020. XHAR: Deep Domain Adaptation for Human Activity Recognition with Smart Devices. In *2020 17th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. 1–9. https://doi.org/10.1109/SECON48991.2020.9158431

[65] Yin Zhu, Yuqiang Chen, Zhongqi Lu, Sinno Jialin Pan, Gui-Rong Xue, Yong Yu, and Qiang Yang. 2011. Heterogeneous transfer learning for image classification. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.