DeepMV: Multi-View Deep Learning for Device-Free Human Activity Recognition

HONGFEI XUE*, State University of New York at Buffalo, USA
WENJUN JIANG*, State University of New York at Buffalo, USA
CHENGLIN MIAO, State University of New York at Buffalo, USA
FENGLONG MA, Pennsylvania State University, USA
SHIYANG WANG, State University of New York at Buffalo, USA
YE YUAN, JD Intelligent Cities Research, China
SHUOCHAO YAO, University of Illinois Urbana-Champaign, USA
AIDONG ZHANG, University of Virginia, USA
LU SU†, State University of New York at Buffalo, USA

Recently, significant efforts are made to explore device-free human activity recognition techniques that utilize the information collected by existing indoor wireless infrastructures without the need for the monitored subject to carry a dedicated device. Most of the existing work, however, focuses their attention on the analysis of the signal received by a single device. In practice, there are usually multiple devices "observing" the same subject. Each of these devices can be regarded as an information source and provides us an unique "view" of the observed subject. Intuitively, if we can combine the complementary information carried by the multiple views, we will be able to improve the activity recognition accuracy. Towards this end, we propose DeepMV, a unified multi-view deep learning framework, to learn informative representations of heterogeneous device-free data. DeepMV can combine different views' information weighted by the quality of their data and extract commonness shared across different environments to improve the recognition performance. To evaluate the proposed DeepMV model, we set up a testbed using commercialized WiFi and acoustic devices. Experiment results show that DeepMV can effectively recognize activities and outperform the state-of-the-art human activity recognition methods.

CCS Concepts: • **Networks** → *Wireless access points, base stations and infrastructure*; • **Human-centered computing** → *Interaction techniques.*

Additional Key Words and Phrases: Human Activity Recognition, Device Free, Deep Learning, Multi-View Learning

*The first two authors contributed equally to this work.
†Lu Su is the corresponding author.

Authors' addresses: Hongfei Xue, State University of New York at Buffalo, Buffalo, NY, USA, hongfeix@buffalo.edu; Wenjun Jiang, State University of New York at Buffalo, Buffalo, NY, USA, wenjunji@buffalo.edu; Chenglin Miao, State University of New York at Buffalo, Buffalo, NY, USA, cmiao@buffalo.edu; Fenglong Ma, Pennsylvania State University, University Park, PA, USA, fenglong@psu.edu; Shiyang Wang, State University of New York at Buffalo, Buffalo, NY, USA, shiyangw@buffalo.edu; Ye Yuan, JD Intelligent Cities Research, Beijing, China, yuanye48@jd.com; Shuochao Yao, University of Illinois Urbana-Champaign, Urbana, IL, USA, syao9@illinois.edu; Aidong Zhang, University of Virginia, Charlottesville, VA, USA, aidong@virginia.edu; Lu Su, State University of New York at Buffalo, Buffalo, NY, USA, lusu@buffalo.edu.

2474-9567/2020/3-ART34 $15.00
https://doi.org/10.1145/3380980

ACM Reference Format:
Hongfei Xue, Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shiyang Wang, Ye Yuan, Shuochao Yao, Aidong Zhang, and Lu Su. 2020. DeepMV: Multi-View Deep Learning for Device-Free Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 34 (March 2020), 26 pages. https://doi.org/10.1145/3380980

1 INTRODUCTION

In order to overcome the limitation of traditional wearable device based human activity recognition approaches which may bring extra burden and discomfort to the monitored subject, significant efforts are recently made to explore **device-free human activity recognition** techniques that utilize the information collected by existing indoor wireless infrastructures without the need for the monitored subject to carry a dedicated device. These approaches, though different in various aspects, share the same idea: by extracting and analyzing information carried by the wireless signal transmitted between a pair of wireless devices (e.g., smartphone, laptop, WiFi access point), we can infer the activities of a person located between the sender and receiver, since his/her activities would make changes to the transmission pattern of the wireless signals.

Fig. 1. Real-world Scenario of Wireless Environment.

Most of the existing work, however, focuses their attention on the analysis of the information provided by a single sender-receiver pair. In practice, there are usually multiple information sources available for the task of activity recognition. Figure 1 shows a real-world human activity recognition scenario. As can be seen, in the room, there are multiple devices, from smartphone and iPad to the laptop, and even to the TV and the printer, which can receive various wireless signals, from the WiFi packets broadcast by the access point, to acoustic or even light waves. Each of these devices can be regarded as an information source and provides us a unique "view" of the observed subject. Intuitively, if we can combine the complementary information carried by the multiple views, we will be able to improve the activity recognition accuracy.

However, to unleash the power of multi-view information, we have to address a series of challenges. First, different sources may provide *heterogeneous data*. On one hand, different types of signals (e.g., WiFi signal, ultrasound or acoustic wave, and visible light) may be collected concurrently for the recognition of same activities. On the other hand, different sender-receiver pairs may have different packet exchanging patterns (e.g., transmission rate, signal strength), and this will add further heterogeneity to the information extracted from different devices. Second, *different sources may carry different amount of information*, due to various reasons such as the quality of hardware, the distance and angle to the observed subject, background noise, as well as the

Fig. 2. Illustration of Human Activities Used to Evaluate the Performance of DeepMV.

setting of the room. An ideal activity recognition approach should be able to capture the variance of information quality among the sources and rely on more informative ones. Third, the wireless signals arriving at the receiving devices usually *carry substantial information that is specific to the environment where the activities are recorded and the human subject who conducts the activities*. On one hand, the signals, when being transmitted, may be penetrating, reflected, and diffracted by the media (e.g., air, glass) and objects (e.g., wall, furniture) in the ambient environment. On the other hand, different human subjects with different ages, genders, heights, weights, and body shapes affect the signals in different ways, even if they are doing the same activity. As a result, an activity recognition model that is trained on a specific subject in a specific environment will typically not work well when being applied to predict another subject's activities that are recorded in a different environment.

To tackle the above challenges, we propose to adopt deep learning techniques, which have been proved to be effective on noisy and heterogeneous big data. In this paper, we propose a multi-view deep learning framework, named DeepMV, to recognize human activities from heterogeneous device-free data sources. Specifically, we utilize a CNN-based module to preserve the unique characteristics of each view while uniform the dimensionality of heterogeneous inputs. A Hierarchically-Weighted-Combination module is developed to estimate the quality of information contributed by each view and combine multi-view features in a weighted manner. We also construct a *adversarial network* that can *remove the environment and subject specific information* contained in the activity data and *extract environment/subject-independent features* shared by the data collected on different subjects under different environments. Taking advantage of the multi-view structure of wireless infrastructures, the proposed DeepMV is able to not only make full use of data collected from different views with different levels of quality, but also characterize different patterns of the data across different environments.

In order to evaluate the proposed DeepMV framework, we conduct extensive real-world experiments. In particular, we deploy in three rooms both WiFi and acoustic transmitters as well as receivers. The transmitter continuously emits signals to the receivers, while a user is doing one of the nine activities (shown in Figure 2) in each room. The collected WiFi and acoustic signals, after being preprocessed, are fed to our proposed DeepMV

model. Experimental results demonstrate that our model outperforms the state-of-the-art algorithms significantly, which illustrates the effectiveness of the proposed DeepMV model.

We summarize the contributions of this paper as follows:

- We identify the opportunities as well as challenges in device-free human activity recognition with heterogeneous multi-view data.
- We propose DeepMV, a unified multi-view deep learning framework, to learn informative representations of heterogeneous device-free data. DeepMV can combine different views' information weighted by the quality of their data and extract environment/subject independent information to improve the recognition performance.
- We set up a testbed using COTS (i.e., commercial off-the-shelf) WiFi and acoustic devices, and collect real-world activity data. We empirically show that the proposed DeepMV model can effectively recognize activities and outperform the state-of-the-art human activity recognition methods on the collected dataset.

2 SYSTEM OVERVIEW

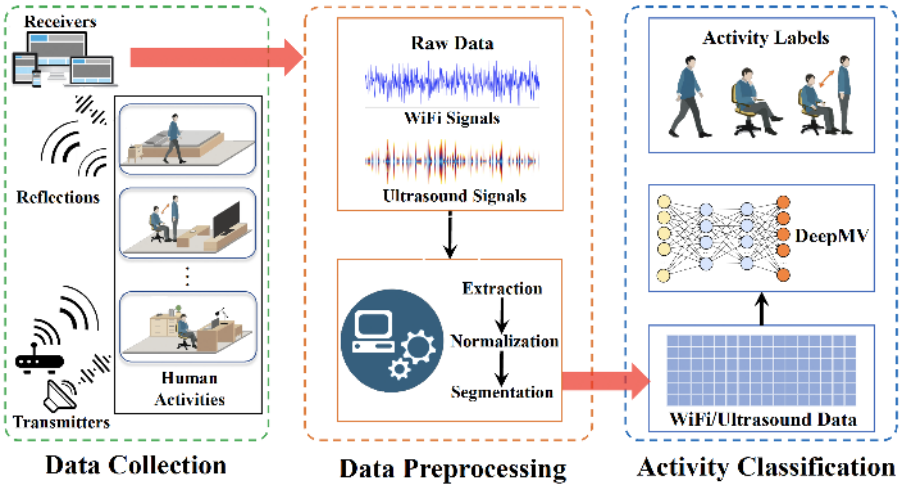

Fig. 3. The Overview of DeepMV Human Activity Recognition System.

DeepMV is a multi-view device-free human activity recognition system which takes advantage of the superior representation capability of deep learning techniques. The DeepMV system takes the raw sensing data (e.g., WiFi/Ultrasound signals) as the input and outputs inferred activities of the monitored subject. Figure 3 provides an overview of the DeepMV system. As can be seen, DeepMV consists of three major components: (1) data collection, (2) data preprocessing, and (3) activity classification.

2.1 Data Collection

In a device-free activity recognition scenario, transmitters are continuously sending signals to the space where the subject is taking daily activities that affect the propagation pattern of the signals received by the receivers. The major function of the data collection component is to collect various raw signals from heterogeneous wireless devices and forward them to the data preprocessing component. In this paper, we take into account the human activities performed in different environments (e.g., different rooms). In each environment, we deploy off-the-shelf

WiFi access point and iPad as transmitters to broadcast WiFi and acoustic (ultrasound) signals respectively. At the same time, PCs and smartphones are used as receivers to capture those signals correspondingly. In real practice, multiple PCs and smartphones are usually placed in different places of an indoor environment, providing multi-view information for the recognition of human activities.

2.2 Data Preprocessing

The function of this component is to generate the data that can be directly fed to the proposed deep learning model. After being collected, the raw signals are stored in some specifically formatted files (e.g., wav format for audio data) by the receivers. The raw data are first extracted and aligned from those files. Then to remove noise, we take several procedures to preprocess the extracted data, such as resizing, Fourier transform, normalization, standardization, spectrogram generation, etc. After that, to generate a dataset with a predefined data size, we also segment the data into non-overlapping pieces. The generated dataset is then fed to the proposed deep learning model for activity classification.

2.3 Activity Classification

Since the processed activity data are still very complex, i.e, high-dimensional, noisy and heterogeneous, traditional machine learning models can not well capture the underlying patterns of these data. To address this challenge, we make use of deep learning techniques which have been proved effective for extracting representations from complex data. In particular, we propose a multi-view deep learning framework that can not only incorporate the quality of data collected from different views, but also remove the specific information in each **domain** (defined as a pair of environment and human subject) and extract commonness shared across different domains. The details of the proposed DeepMV model are described in Section 3. With the processed heterogeneous activity data, our model is able to recognize the human activities recorded under unseen environments and significantly improve the recognition performance by learning informative representations of different activities.

3 METHODOLOGY

In this section, we introduce our DeepMV model, a unified multi-view environment-independent deep learning framework for human activity recognition with heterogeneous device-free data inputs. The architecture of the proposed model is illustrated in Figure 4. In our model, the input contains activity data from heterogeneous sources (e.g., wireless devices) and only a part of them are manually labeled. The goal of our model is to combine these heterogeneous data not only to recognize human activities but also make itself adaptive to different environments, especially for the environments without any provided activity labels.

To achieve this goal, the proposed model is enabled to simultaneously learn both discriminative features from heterogeneous data and transferable features for various environments. To obtain the discriminative features, we first propose a multi-view feature extractor, which consists of two modules: a **View-Representation module** to learn channel level vector representations and a **Hierarchically-Weighted-Combination module** to selectively integrate different views' representations into a latent vector (i.e., the global representation vector) in a hierarchical manner. The integrated feature vector is then utilized by an **Activity Recognizer**, which is designed to maximize the accuracy of activity prediction, and a **Domain Discriminator**, whose goal is to infer the domain label, in other words, to identify the human subject and the environment associated with the activity data. In order to learn transferable features for different environments, we make the feature extractor play a minimax game against the domain discriminator. Eventually, the domain discriminator will be cheated by the feature extractor. As a result, the domain-specific features are removed and the common environment-independent features are remained. In the following subsections, we will elaborate these components, respectively.

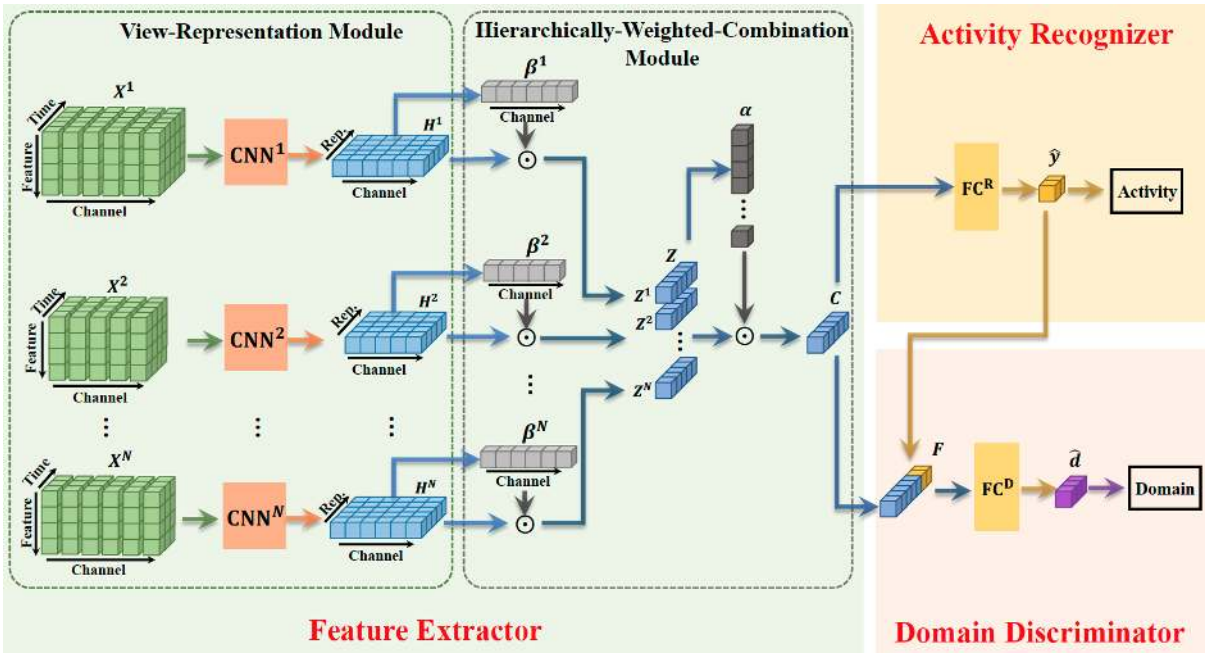

Fig. 4. The Illustration of DeepMV Model.

3.1 View-Representation Module

In practice, the collected multi-view device-free data can be represented as a collection of heterogeneous continuous time series consisting of several non-uniformly sampled signals. In order to *preserve the unique characteristics of each view while uniform the dimensionality of different views' inputs*, we present a CNN-based module to learn view-representations from the processed heterogeneous data. The efficiency and effectiveness of CNN make it an ideal building block for our activity recognition framework.

In our paper, the processed heterogeneous data for each view has three dimensions: the time dimension, the channel dimension and the feature dimension. Taking the WiFi data as an example, the time dimension represents the data collected from different time points, the channel dimension represents the data in different subcarriers, and different features in the data can be the real part, imaginary part, magnitude or FFT of the original complex data. In general, the main purpose of our design for the View-Representation Module is to carefully select the settings (including the number of layers, filter size, padding size, stride size, etc.) of the CNN block for each view to map the data in all dimensions into representation vectors with the same size, except for the channel dimension (because the numbers of channels in different views may be different). So in the designed CNN block, the convolutional layers with sets of learnable filter banks are the most important parts. The batch normalization layers are used for learning stable vector representations, which are followed by *ReLU* and dropout layers.

Specifically, we assume that there are N views of input data $\{\mathbf{X}^1, \mathbf{X}^2, ..., \mathbf{X}^N\}$, and the v-th view has S^v channels. We denote the i-th channel of the v-th view as $\mathbf{X}_i^v$. Then we can obtain its hidden representation $\mathbf{H}_i^v \in \mathcal{R}^{R\times 1}$, where R is the size of the generated representation vector, as follows:

$$\mathbf{H}_i^v = \mathbf{CNN}^v(\mathbf{X}_i^v; \mathbf{\Theta}^v), \tag{1}$$

where $\mathbf{CNN}^v()$ is the multi-layer CNN block to apply on each channel of the v-th view, and $\mathbf{\Theta}^v$ is the corresponding parameters of CNN block. $\mathbf{H}^v = [\mathbf{H}_1^v, \mathbf{H}_2^v, ..., \mathbf{H}_{S^v}^v]$ is the representation matrix of the data in the v-th view. For different views, the corresponding CNN parameters are different, but the CNN parameters for different channels in one view are the same.

3.2 Hierarchically-Weighted-Combination Module

Different views may carry different amount of information, due to various reasons, such as the type of the sensing signal, the quality of hardware, the distance and angle to the observed object, as well as the ambient noise and setting. Additionally, the amount of information in different channels of each view may also be different. This is mainly because different channels usually carry the information from different aspects (e.g., frequencies). An ideal activity recognition approach should be able to capture the variance of information quality among both different views and different channels, and rely on more informative ones to achieve better performance. Towards this end, we propose a Hierarchically-Weighted-Combination module to estimate the quality of information (referred to as *quality weights*) contributed by different channels and views, and combine the information from different channels and views in a hierarchically weighted manner.

The basic idea of our Hierarchically-Weighted-Combination module is derived based on the attention mechanism [19, 24, 40, 68, 87]. The attention mechanism is a weighted aggregation method that is widely used for the application of machine translation [10], computer vision [29, 60], and disease prediction [52, 93]. However, traditional attention mechanism is based on the assumption that *only a few views are related to the task goal*. As a result, it tends to assign close-to-zero weight to most of the views. In the scenario of device-free human activity recognition, however, such an assumption does not hold, since *a significant portion of the views may provide informative observations*.

To address this challenge, we borrow the idea of weight-calculation strategy from [94], and design a smoothing operation and a sharpening operation to not only avoid thoroughly neglecting information from any channels or views, but also fully distinguish them. Specifically, in the Hierarchically-Weighted-Combination module, we first estimate the quality of each channel and then use the weighted combination mechanism to generate the **view representation vector** (i.e., $\mathbf{Z}^v$) for each view. In a similar way, we measure the quality score of each view, and combine different views to generate the **global representation vector** (i.e., $\mathbf{C}$) in a weighted manner.

View Representation Vector Calculation. Given the representation vector $\mathbf{H}^v$ of the v-th view, the quality score of the i-th channel in the v-th view (i.e., e_i^v) can be calculated using the following formula:

$$e_i^v = \mathbf{w}_e^{v\top} \mathbf{H}_i^v + b_e^v \,, \tag{2}$$

where $\mathbf{w}_e^v \in \mathcal{R}^{R\times 1}$ and $b_e^v \in \mathcal{R}^{1\times 1}$ are the parameters to be learned for the v-th view.

Conventionally, a softmax-based normalization method is usually applied on obtained attention energies (i.e., weights). However, this operation tends to assign most of the elements with close-to-zero scores, which is not desired in our multi-view scenario. To fully utilize the information from multiple views and preserve the score distribution in each channel, we first smooth the quality score values and then rescale the values based on their corresponding channels. More formally, given the quality score e_i^v for the i-th channel of the v-th view, we narrow it using the *sigmoid* function and rescale it as follow:

$$\hat{e}_i^v = \text{sigmoid}(e_i^v) \,/\, \sum_{i=1}^{S^v} \text{sigmoid}(e_i^v) \,. \tag{3}$$

After calculating the smoothed score $\hat{e}_i^v$, we then need to sharpen the calculated scores for different channels and prevent the close-to-zero scores. The final quality score of the i-th channel in the v-th view (i.e., β_i^v) is calculated

as follows:

$$\beta_i^v = \exp(\gamma \hat{e}_i^v) \,/\, \sum_{i=1}^{S^v} \exp(\gamma \hat{e}_i^v)\,, \tag{4}$$

where γ is the predefined positive sharpening factor to avoid aggregating multiple focus [19]. The final weights vector for each view is denoted as $\beta^v = [\beta_1^v, \beta_2^v, ..., \beta_{S^v}^v]$. Based on Eq.(3), we know that the quality score $\hat{e}_i^v$ is constrained in the range from 0 to 1. So in Eq.(4), the minimal value of $\exp(\gamma \hat{e}_i^v)$ is greater than or equal to 1 (γ is a positive number), which prevents the close-to-zero attention weights. Then the view-wise representation vector $\mathbf{Z}^v$ can be calculated as:

$$\mathbf{Z}^{\mathbf{v}} = \sum_{i=1}^{S^n} \beta_i^v \odot \mathbf{H}_i^v\,, \tag{5}$$

where $\odot$ is the multiplication between a scalar and a vector. Finally, by stacking the obtained vectors from all the views, we can get the view representation matrix $\mathbf{Z} \in \mathcal{R}^{N \times R}$.

Global Representation Vector Calculation. Similar to the quality score calculation for each channel, we calculate the quality score for each view as follows:

$$E^v = \mathbf{w}_E^\top \mathbf{Z}^v + b_E\,, \tag{6}$$

where E^v is the quality score for the v-th view, and $\mathbf{w}_E$ and b_E are the corresponding parameters. Then we can smooth the quality score as:

$$\hat{E}^v = \text{sigmoid}(E^v) \,/\, \sum_{v=1}^{N} \text{sigmoid}(E^v)\,, \tag{7}$$

where $\hat{E}^v$ is the smoothed score. The final sharpened quality score α^v for the v-th view is calculated as:

$$\alpha^v = \exp(\gamma \hat{E}^v) \,/\, \sum_{v=1}^{N} \exp(\gamma \hat{E}^v)\,, \tag{8}$$

where γ is the same predefined sharpening factor as in Eq.(4). Finally, the global representation vector $\mathbf{C} \in \mathcal{R}^{R \times 1}$ can be calculated as:

$$\mathbf{C} = \sum_{v=1}^{N} \alpha^v \odot \mathbf{Z}^v\,. \tag{9}$$

3.3 Activity Recognizer

Using the output of the Hierarchically-Weighted-Combination module (i.e., $\mathbf{C}$), we can predict the label of the input human activity data. To achieve this goal, we map the global representation vector $\mathbf{C}$ into the latent space of human activity by using a fully connected feedforward neural network. The network we use contains stacked fully-connected layers. Each of these layers is followed by a *ReLU* activation function and a dropout layer to introduce nonlinearity. Then a linear transformation is leveraged to project the network output to the number of human activities and is followed by a $Softmax$ function to predict the probability of activities. We denote the designed fully connected feedforward neural network with the $Softmax$ layer as follows:

$$\hat{\mathbf{y}} = \mathbf{FC}^R(\mathbf{C}; \mathbf{\Theta}^R), \tag{10}$$

where $\mathbf{\Theta}^R$ is its parameter set and $\hat{\mathbf{y}} \in \mathcal{R}^{M \times 1}$ is the predicted probability distribution of one sample. And M is the number of different activities.

Since our input data include both labeled and unlabeled activities, we denote the predicted probability of activities as $\hat{\mathbf{y}} = [\hat{\mathbf{y}}^l, \hat{\mathbf{y}}^u]$, where $\hat{\mathbf{y}}^l$ and $\hat{\mathbf{y}}^u$ are the predicted probabilities of labeled data and unlabeled data, respectively.

We maximize the accuracy of activity prediction through minimizing the cross entropy loss between the prediction of the labeled data and their ground truths as follows:

$$L^R = -\frac{1}{n^l}\sum_{k=1}^{n^l}\sum_{m=1}^{M}\mathbf{y}_m^l \log(\hat{\mathbf{y}}_m^l)\,, \tag{11}$$

where n^l is the number of data samples with labels.

However, the activity recognizer alone is not sufficient to learn a good classifier because the data from different environments contain environment-specific information and there are no labeled data for a significant portion of the environments. As a result, the activity recognizer may be misguided by the environment-specific information of those labeled data and fail to learn the features that are common for all the environments.

3.4 Domain Discriminator

In order to learn the common features in different environments, we employ a domain adaptation technique called unsupervised adversarial training [26, 27], which utilizes the unlabeled data to project the data from different environments into an environment-independent latent space.

To achieve this, we design a domain discriminator whose goal is to maximize the accuracy of domain label prediction. In other words, it is to identify the human subject and the environment associated with the activity data.

The input of the domain discriminator is the concatenation of global representation vector (i.e., C) and the prediction of the activity recognizer (i.e., $\hat{\mathbf{y}}$) as follows:

$$\mathbf{F} = \mathbf{C} \oplus \hat{\mathbf{y}}, \tag{12}$$

where $\oplus$ is the concatenation operation and $\mathbf{F} \in \mathcal{R}^{(R+M)\times 1}$ is the domain representation vector. The domain discriminator takes both $\mathbf{C}$ and $\hat{\mathbf{y}}$ as the input in order to retain the information relevant to the activities in the extracted features [97].

The domain discriminator is implemented with a similar fully connected feedforward neural network architecture as the activity recognizer, which projects $\mathbf{F}$ into domain distributions $\hat{\mathbf{d}}$, as follows:

$$\hat{\mathbf{d}} = \mathbf{FC}^D(\mathbf{F}; \mathbf{\Theta}^D), \tag{13}$$

where $\mathbf{FC}^D$ is the designed neural network similar to $\mathbf{FC}^R$, and $\mathbf{\Theta}^D$ denotes its parameters.

The domain discriminator maximizes the domain label prediction through minimizing the loss between the domain distributions and true domain labels as follows:

$$L^D = -\frac{1}{n}\sum_{k=1}^{n}\sum_{g=1}^{G}\mathbf{d}_g \log(\hat{\mathbf{d}}_g), \tag{14}$$

where G denotes the number of domains, and $n = n^l + n^u$ is the total number of data samples contains both the labeled and the unlabeled. This calculation of cross entropy is also similar to the activity recognizer.

However, if the accuracy of the domain label prediction is high, it means the features we extract contain significant domain-specific information on the labeled domains so that they may not be good to predict the unlabeled activities. To solve this problem, we make the feature extractor learn features that boost the activity recognizer but cheat the domain discriminator through minimizing the following loss:

$$L = L^R - L^D. \tag{15}$$

Such a minimax game between the feature extractor and the domain discriminator will minimize the maximum accuracy the domain discriminator can achieve and finally learn the common environment-independent features.

4 EXPERIMENTS

In this section, we first introduce the state-of-the-art human activity recognition approaches as baselines. We then describe the experiment setups, including hardware/software settings, human activity design, data preprocessing and model settings. After that, we show the experimental results and analyze the learned weights on the homogeneous WiFi dataset and compare our model with several state-of-the-art human activity recognition algorithms. Finally, we evaluate our method on heterogeneous dataset containing both WiFi and ultrasound signals.

4.1 Baselines

To fairly evaluate the performance of the proposed DeepMV model, we use the following models as baselines:

SVM [20]. Support Vector Machine (SVM) is a widely adopted supervised machine learning model. There are some studies [78, 82, 98] employing SVM model for human activity recognition task. Since standard linear SVM model is a binary classifier, we use one-vs-all SVM for our multi-class classification task. In the experiments, we flatten data from all views into a single feature vector and use PCA [83] to reduce the feature dimension to 256. Then we feed the data to the SVM model.

RF. Random Forest (RF) is a frequently used learning method for classification tasks by ensembling result of multiple decision trees [36]. The algorithm can often produce a good prediction result without tuning too many parameters. Similar to the settings of SVM, we first use PCA [83] to reduce the feature dimension to 256 and then feed the data into the RF model in the experiments.

DeepSense [89]. DeepSense is the state-of-the-art deep learning model for classification of multi-sensor data. The architecture of DeepSense includes three layers of local CNN, three layers of global CNN and two layers of GRU. In our experiments, we follow the settings of the original paper. Specifically, on each convolutional layer, the number of filters is 64 and the size of filters is set to 3×3. In addition, dropout and batch norm technologies are also used. The DeepSense model has mainly two limitations compared to our DeepMV model: (1) the DeepSense model does not take the quality of different views into consideration but simply concatenates all the view representations; (2) the DeepSense model is more sensitive to environment changing because the model doesn't consider any domain information.

QualityDeepSense [90]. QualityDeepSense is a deep learning model that incorporates the sensor-temporal attention mechanism based on DeepSense [89]. The architecture of QualityDeepSense is the same as the DeepSense model, except that there are a sensor attention layer among input sensor and a temporal attention on the top of recurrent layers. Even though the authors claim that the QualityDeepSense can automatically balance the contribution of sensor inputs over time by their sensing qualities, the model still fails to give out an explicit sensor quality score. Additionally, QualityDeepSense is sensitive to environment changing since it does not take the domain information into consideration.

EI [42]. EI is the state-of-the-art environment-independent deep learning model for device-free activity recognition. The architecture of EI includes three layers of CNN with 64, 128, and 256 filters respectively followed by two fully connected layers of 256 neurons. In addition, non-linear activation functions, batch norm and max-pooling are used. Different from our model, EI is incapable of discriminating between different views and thus applying the same CNN for all the views.

DeepMV Variants. There are four modules in the proposed DeepMV model, including View-Representation module (referred to as VR), Hierarchically-Weighted-Combination module (referred to as HWC), activity recognizer and domain discriminator. Thus, we propose three simplified models as baselines:

- **VR+Avg**. we use View-Representation module to obtain the representations of channels, and then average all the channel representations. Moreover, a fully connected layer is employed to reduce its dimension. Finally, we use the reduced representation to make predictions.

- **VR+HWC**. In this baseline, we use the View-Representation module and the Hierarchically-Weighted-Combination module together to recognize human activities.

Table 1. Descriptions of Studied Human Activities.

Activity Name	Details
Wiping the table	The subject sits in front of a table and wipes it.
Typing	The subject sits in front of a table and types on a keyboard.
Writing	The subject sits in front of a table and writes on a piece of paper.
Rotating the chair	The subject sits on a swivel chair turning around between $-45°$ and $45°$.
Moving the chair	The subject moves a chair in the room.
Walking	The subject walks around in the room.
Cleaning the floor	The subject uses a mop to clean the floor.
Running in place	The subject runs on a spot in the room.
NULL	The subject does some undefined activities in the room.

4.2 Experiment Setups

The setups of our experiments include human activity design, hardware/software setup, data preprocessing and settings of the proposed DeepMV model.

4.2.1 Human Activity Design. As shown in Figure 2, we consider 9 different human activities in the experiments. The details of these activities are described in Table 1. We employ 8 volunteers (including both men and women) as the subjects and collect data from 3 different rooms. Each subject is asked to repeat the 9 activities in each room for 4 rounds and in each round, we let each subject take each type of activity for 51 seconds.

4.2.2 Hardware/Software Setup. In our experiments, we collect two kinds of signals: WiFi and Ultrasound. To collect **WiFi signals**, we use a TP-Link AC3150 Wireless WiFi Gigabit Router (Archer C3150 V1) to send packets to different receivers at a constant packet transmission rate, i.e., 200 packets per second, which is a reasonable transmission rate in practical wireless communication scenarios. Each receiver is configured with Intel Wireless Link 5300 NIC, Ubuntu 11.04 LTS with 2.2.36 kernel, and Linux 802.11n CSI extraction toolkit provided in [32]. For both 2.4 GHz and 5 GHz radio bands, the Linux 802.11n CSI extraction toolkit can report the CSI matrices of 30 sub-carriers. Figure 5 pictures the experiment environment, where the positions of transmitters and receivers are marked.

In order to collect **ultrasound signals**, we use an Apple iPad mini 4 as the sound generator, which transmits near-ultrasound (i.e., 20 KHz) signals towards the subject. Since the sampling rate of the MICs on smartphones can reach as high as 44.1 KHz, we can use smartphones as receivers. In our experiments, three Huawei Nexus 6P's are used as receivers to record the ultrasound signals reflected by the body of the subject. These receivers are deployed at different positions in the room as illustrated in Figure 5.

4.2.3 Data Preprocessing. Since the proposed DeepMV model is able to deal with heterogeneous data, in the experiments, we include two kinds of data: WiFi signals and Ultrasound signals. For different kinds of signals, we employ different preprocessing approaches.

- **WiFi signals**. In the experiment, we use the amplitude information of CSI. We first interpolate the CSI measurements to uniform sampling period to deal with packet loss and delay. Then the data are normalized to the mean zero and the standard deviation one. After that, we use a Hampel filter to eliminate the outliers. We

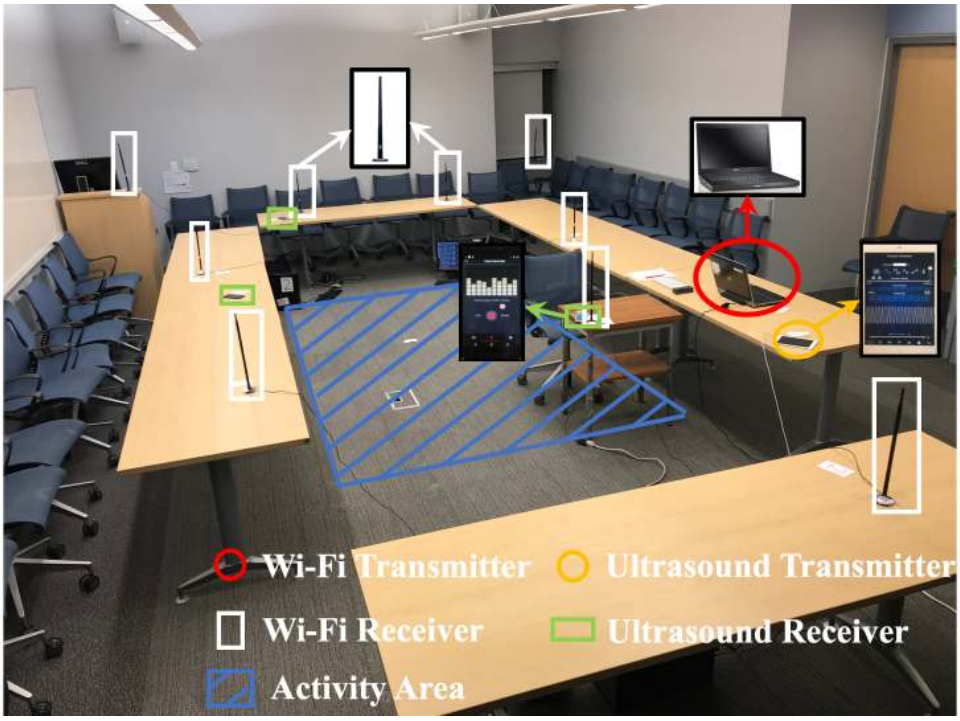

Fig. 5. Experiment Setup of DeepMV Human Activity Recognition System.

also segment the data without overlap via a window of 600 points, which corresponds to 3 seconds. We combine each segment with the FFT of it as the input of the deep learning model.

• **Ultrasound signals**. The transmitted ultrasound is a sinusoidal signal. When the signal is reflected from the human body, the movement of the human will increase or decrease the propagation distance of the signal, which causes a phase shift on the received signal.

We utilize the method in [80] to extract the phase information. Assume that the transmitted signal is $T(t) = A\cos(2\pi ft)$, the received signal can be represented by $R(t) = A'\cos(2\pi ft - 2\pi fd/c)$, where A and A' are the amplitude of the transmitted and received signal respectively, f is the frequency, c is the speed of sound, d is the propagation distance of the signal. $2\pi fd/c$ is the phase lag caused by the propagation delay. If we multiply the received signal with $\cos(2\pi ft)$, the result is:

$$\begin{aligned} &A'\cos(2\pi ft - 2\pi fd/c) \times \cos(2\pi ft) \\ &= \frac{A'}{2}(\cos(-2\pi fd/c) + \cos(4\pi ft - 2\pi fd/c)). \end{aligned} \tag{16}$$

After filtering the signal with a low pass filter with frequency $f' << f$, we remove the second term and the rest term $\frac{A'}{2}(\cos(-2\pi fd/c)$ does not change over time and is decided by the propagation distance d. In a similar way, we multiply the received signal with $-\sin(2\pi ft)$ and get $\frac{A'}{2}(\sin(-2\pi fd/c))$. The signals $\frac{A'}{2}(\cos(-2\pi fd/c)$ and $\frac{A'}{2}(\sin(-2\pi fd/c))$ are downsampled to 344 Hz and we segment these signals without overlap via a window of 1033 points (about 3 seconds). Next, we calculate the spectrogram of each segment, divide the frequency dimension of the spectrogram into even frequency bands, and call each frequency band a channel.

4.2.4 Model Settings. We validate the proposed DeepMV model on the dataset containing two kind of signals, WiFi signals and ultrasound signals. In the experiment, we stacked 6 blocks of CNN in View-Representation Module. The major difference of CNN parameter settings between two types of signals is the filter size. We list their CNN filter parameters in Table 2. The filter number for both signals are 64. There is no pooling and no padding in the convolutional layers. The representation length R is set to 256. The shapening factor γ in the Hierarchically-Weighted-Combination module of the model is set to 2. For activity recognizer and domain discriminator part, one layer of fully connected neural network are utilized. The sizes of layer is simply set to 64. In addition to aforementioned parameters, we adopt ReLU [53] as the activation function in each layer, and use dropout technique[66], with dropout rate set as 0.8 for both fully connected network and convolutional neural network.

Table 2. CNN Parameter Settings.

Layer Index	**WiFi Data**		**Ultrasound Data**	
	Filter	Stride	Filter	Stride
1	1×12	1×4	3×4	1×2
2	1×7	1×3	3×4	1×2
3	1×5	1×2	1×4	1×2
4	1×5	1×2	1×3	1×1
5	1×5	1×2	1×3	1×1
6	1×3	1×1	1×3	1×1

During the training process, ADAM optimization algorithm [44] is used to optimize the parameters. The learning rate is $1e-4$, and the batch size is 72. We use the data collected in the first and the second rooms for training and the data collected in the third rooms for testing. We use the accuracy score as our performance metric. We implement the proposed DeepMV model using Tensorflow [1]. The training process is done locally using NVIDIA Titan Xp GPU.

4.3 Experiments on the Homogeneous WiFi Data

The homogeneous WiFi data are collected from three different rooms, and in each room, we set up 9 WiFi views. In this section, we first conduct experiments on the homogeneous WiFi dataset to show the effectiveness of the proposed DeepMV model for human activity recognition task, and then analyze the weights learned by the proposed model.

4.3.1 Performance Validation. In this experiment, we select two rooms and use the data collected from them as the training dataset. The data collected from the third room are treated as the testing set. Table 3 shows the accuracy of all the approaches on the WiFi dataset. We can observe that the performance of traditional classification approaches SVM and RF is much worse than that of deep learning models, which demonstrates the effectiveness of deep learning models for HAR task.

The four deep learning baselines including DeepSense, QualityDeepSense, VR+Avg and VR+HWC all use different CNNs for different views. However, when combining the representations of these views, they use different ways. DeepSense and VR+Avg are not as good as VR+HWC and QualityDeepSense. This is mainly because DeepSense and VR+Avg treats all the views and channels equally, while VR+HWC and QualityDeepSense can automatically distinguish the qualities of different views or channels, and the results can benefit from combining the views or channels according to their quality.

Table 3. Performance on the WiFi Dataset.

Model	Accuracy
SVM	0.117
RF	0.225
EI	0.660
DeepSense	0.648
QualityDeepSense	0.693
VR+Avg	0.615
VR+HWC	0.716
DeepMV	**0.837**

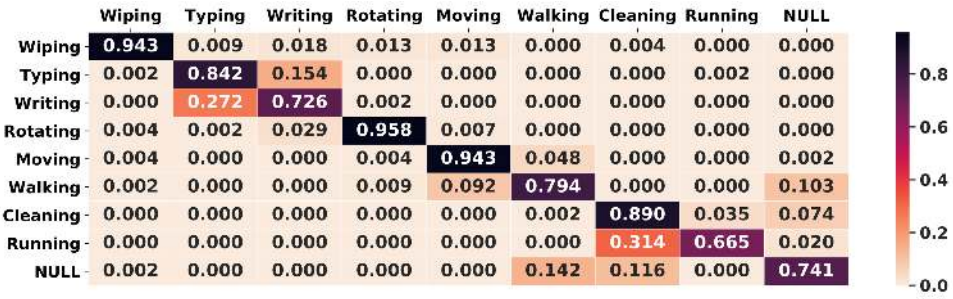

Fig. 6. Confusion Matrix Learned by DeepMV on the WiFi Dataset.

As WiFi signals are sensitive to the environment [42], the aforementioned four methods all suffer from a common issue that the learned models may be biased by the data in the training set. So these methods cannot perform well on the WiFi data collected from a different room. By extracting features that are common across different domains, EI outperforms VR+Avg and DeepSense. However, the improvement is not significant because EI is not designed for the multi-view scenario. Specifically, it uses the same CNN for all the views, which can not capture the variance of information quality across different views.

Our proposed DeepMV improves the feature extractor of EI with a View-Representation module and a Hierarchically-Weighted-Combination module. The former enables the model to extract the unique characteristics of each view, and the later smartly combines the views according to their importance with regard to the prediction task. With these two modules, DeepMV achieves significant improvement over EI on the multi-view human activity recognition tasks.

The proposed DeepMV assumes that different views may have different degrees of contributions for the HAR task. To validate this assumption, in this subsection, we apply DeepMV to each individual WiFi view. Table 4 lists the accuracy of each view. We can observe that 1) the accuracy of different views is different, which clearly confirms our assumption; and 2) even the highest accuracy (i.e., 0.699 on the third view) among all the views is much lower than DeepMV's accuracy on multi-view data as shown in Table 5. The results justify the necessity of combining the information from multiple views for the HAR task.

Figure 6 shows the confusion matrix learned by the proposed DeepMV model on the homogeneous WiFi dataset. We can observe that among the three fine-grained activities (wiping the table, typing and writing), wiping the table is the most distinguishable due to its special pattern. And typing and writing are more likely to be

mistakenly classified to each other. The reason is that these two fine-grained activities are quite similar to each other. For the five coarse-grained activities (rotating the chair, moving the chair, walking, cleaning the floor and running in place), rotating the chair and moving the chair are two most distinguishable activities, which both achieve over 0.94 accuracy. Since moving the chair and walking are very similar, the two activities are sometimes mistakenly classified to each other. Since cleaning the floor and running are conducted in the similar place, so the running activity may be classified as cleaning the floor. In this experiments, we also add a NULL class for the activities. In the NULL class, the activities are not pre-defined and some of them are composed of the activities involving walking and arm movements. So the NULL class activities may be classified to walking and cleaning the floor.

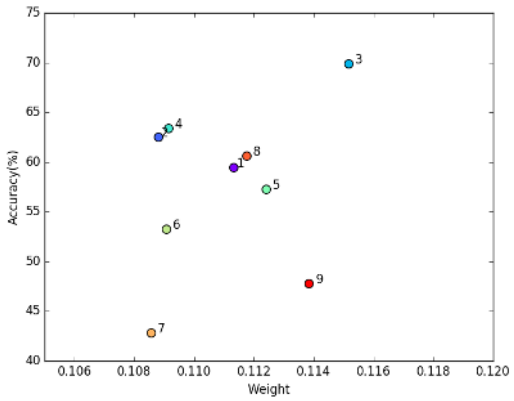

Fig. 7. The Relationship between Weights and Accuracy on the Homogeneous WiFi Dataset.

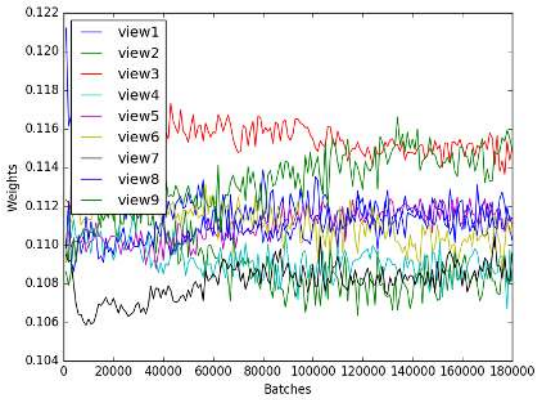

Fig. 8. The Trend of Weight Convergence on the Homogeneous WiFi Dataset.

4.3.2 Learned Weight Analysis. The advantage of DeepMV is its ability to interpret the importance of each view through analyzing the learned weights. In this section, we analyze the weight of each view learned by DeepMV on the WiFi dataset. The relationship between the learned weight and accuracy is shown in Figure 7. We can observe that the learned weights are positively correlated with the performance of individual views, which shows that the quality of each view on the prediction task is automatically captured. In fact, it proves that DeepMV can provide a high-level interpretability for the final prediction.

Additionally, we also quantitatively show the convergence of the weights learned by DeepMV in Figure 8. We can observe that at the beginning, all the views have similar weights. As the number of iterations increases, the weight on each view starts to be different. This shows that DeepMV can learn different weights on different views, and at the same time, the learned weights can converge to relatively stable values.

Table 4. Accuracy of Each View on the WiFi and Acoustic Dataset.

View Index	1	2	3	4	5	6	7	8	9	10	11	12
WiFi	0.595	0.626	0.699	0.634	0.573	0.533	0.429	0.607	0.478	-	-	-
Acoustic	-	-	-	-	-	-	-	-	-	0.295	0.554	0.520

4.4 Experiments on the Heterogeneous Data

To further demonstrate DeepMV's advantages, we conduct experiments on the heterogeneous dataset containing both WiFi data and acoustic data. Specifically, in addition to nine WiFi views in each room, we also collect three acoustic views.

Table 5. Performance on the Heterogeneous Dataset.

Models	Accuracy
SVM	0.220
RF	0.233
VR+Avg	0.703
VR+HWC	0.723
DeepMV	**0.879**

4.4.1 Performance Validation. Similar to the experiments on the homogeneous data, we use two rooms' data as the training set, and one room's data as the testing set.

For DeepSense, QualityDeepSense and EI, they require that the CNNs used for extracting feature representations from each view have the same architecture. However, in the heterogeneous dataset, WiFi data and acoustic data have different dimensions and cannot fit one CNN. So we do not consider DeepSense, QualityDeepSense and EI in the following experiments. Table 5 shows the results on the heterogeneous dataset. Similar to that in the homogeneous dataset, the performance of deep learning based models is much better than that of traditional machine learning models.

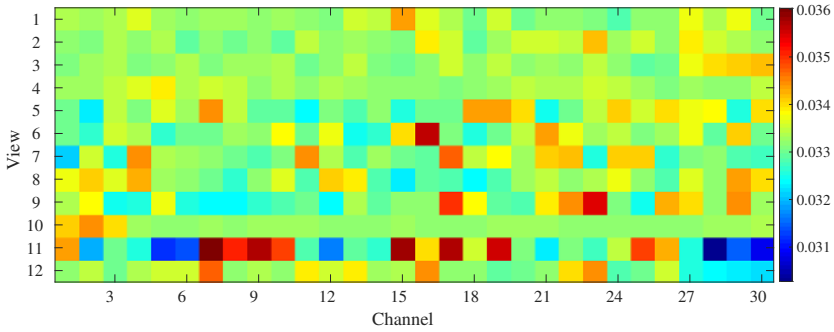

Fig. 9. The Heatmap of the Weight on Each View and Channel.

In order to further explore the relationship between different views and channels, we compare the derived weight on each view and channel (i.e., the product of β and its corresponding α) in Figure 9. The first nine views represent the WiFi views, and the last three views represent the acoustic views. For the WiFi views, the channels are their sub-carriers. While for the acoustic views, the channels are their frequency bands. The warm color represents high weight and the cool color represents low weight. From the figure we can observe that not only the weights of the WiFi views, but also the weights of the acoustic views are quite different. The results show that the proposed View-Representation module can successfully capture the differences among the views, which boosts

the performance of VR+Avg, VR+HWC and DeepMV. Furthermore, for a single view, the weights on the channels are also not evenly distributed. Through combining the information from views and channels in a hierarchically weighted manner, VR+HWC achieves better performance than VR+Avg, which does not differentiate the quality of different views and channels.

Similar to the experiments on the homogeneous dataset, the proposed DeepMV achieves significant improvement over the performance of baseline methods as well as the performance of single acoustic views (Table 4) on the heterogeneous dataset, which shows that the proposed method is robust not only in different environments but also in dealing with complex heterogeneous data.

Fig. 10. Confusion Matrix Learned by DeepMV on the Heterogeneous Dataset.

Figure 10 shows the confusion matrix learned by the proposed DeepMV model on heterogeneous dataset. We can observe that the proposed DeepMV can achieve good performance for each activity. Similar to the conclusion in the homogeneous experiments, the fine-grained activities typing and writing are still relatively difficult to distinguish. The coarse-grained activities cleaning the floor and running in place are often mistakenly classified to each other. And the NULL class is easily classified to walking and cleaning the floor.

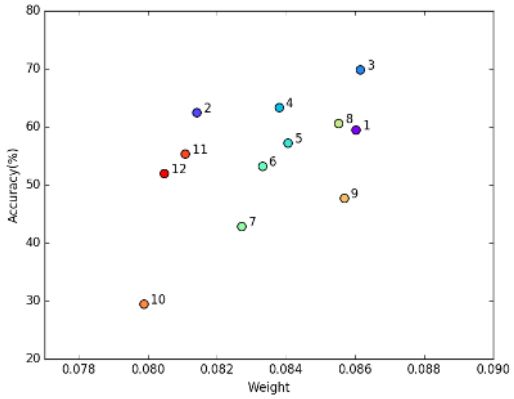

Fig. 11. The Relationship between Weights and Accuracy on the Heterogeneous Dataset.

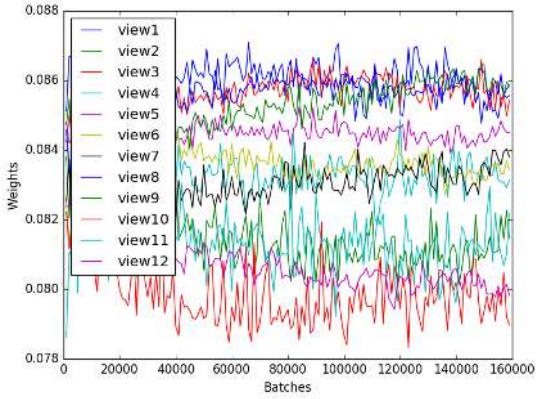

Fig. 12. The Trend of Weight Convergence on the Heterogeneous Dataset.

4.4.2 Learned Weight Analysis. In this section, we analyze the weight of each view learned by DeepMV from the heterogeneous dataset of both WiFi and acoustic data. The performance of each WiFi and acoustic view is

shown in Table 4. Then we show the relationship between the learned weight and the performance of each single view in Figure 11. It is not difficult to find that the positive correlation between the learned weights and the accuracy, which we find in the homogeneous WiFi dataset, also exists in the acoustic views as well as in the heterogeneous views. This proves that the design of DeepMV is transferable across modalities. For comparison, we also demonstrate the convergence of the learned weights with regard to iterations on the heterogeneous dataset in Figure 12. Similar patterns as that in Figure 8 can be observed.

4.4.3 Experiments on Different Transceiver Pair Numbers. In this experiment, we study the effect of transceiver pair numbers on the performance of our system. Here we consider three cases in which the transceiver pair numbers are set as 4 (3 WiFi views and 1 acoustic view), 8 (6 WiFi views and 2 acoustic views) and 12 (9 WiFi views and 3 acoustic views), respectively. For the first case, we deploy all the 4 receivers on only one side of the subject. For the second case, the receivers are deployed on two sides of the subject. On each side, there are 4 receivers (3 WiFi views and 1 acoustic view) and they are along the wall of the room. For the third case, the 12 receivers are deployed around the subject. The classification accuracy for the three cases are reported in Table 6, from which we can observe that the larger the transceiver pair number, the better the performance of our model. This is mainly because the information from different directions is complementary to each other.

Table 6. Performance on varying transceiver pair numbers.

Number of Transceiver Pairs	**4**	**8**	**12**
DeepMV	0.766	0.785	0.876

4.4.4 Experiments on Different Transceiver Locations. We also study the influence of the transceiver locations to our system without the re-training of our model. In this experiment, the model is trained using the data collected from the above described three rooms. Then, we randomly change the transceiver locations in one of the room, and ask 5 subjects to perform the aforementioned activities in this room. The data collected in this room are used as testing data. Table 7 reports the classification accuracy for the newly collected testing data. We can see that our proposed DeepMV model can achieve an accuracy of 0.619 (for 9 activities) while the accuracy of the baseline method is only 0.553. The results show that our proposed model can still achieve good performance and outperform the baseline without retraining the model when the deployment of WiFi transceivers is changed.

Table 7. Performance on random locations of transmitters and receivers

Models	**VR+HWC**	**DeepMV**
Accuracy	0.553	**0.619**

4.4.5 Experiments on Different Data Segmentation Lengths. Next, we study the effect of different segmentation lengths on the performance of our system. In this experiment, we consider three different segmentation lengths, i.e., 0.5s, 1.5s and 3.0s, and report the classification accuracy in Table 8. The results show that our proposed DeepMV model performs better than the baseline model in all cases. In addition, we can see that the baseline model is not sensitive to the segmentation lengths and it has similar performance in all cases. However, our model can achieve much higher accuracy when the segmentation length is large. This is mainly because that with a larger segmentation length, it is easier for our model to identify and extract the domain independent information from the data, and further improve the model performance.

Table 8. Performance on varying segmentation lengths.

Length of Segmentation	0.5s	1.5s	3.0s
VR+HWC	0.696	0.690	0.723
DeepMV	0.702	0.760	0.879

4.4.6 Experiments on Unseen Subject. To demonstrate the generalization of our proposed DeepMV model, we also evaluate the performance of our system on an unseen subject. In this experiment, we consider two scenarios: the scenario of an unseen subject in a seen room and the scenario of an subject in an unseen room. We first collect data for the aforementioned 8 subjects in two rooms and use these data for training. Then we ask the $9th$ subject (the unseen subject) to perform the activities in one of the above two rooms (the seen room) and the third room (the unseen room), respectively, and use the collected data for testing. We report the classification accuracy in Table 9. The results show that our proposed model outperforms the baseline in both scenarios. For example, in scenario B, our proposed model can achieve an accuracy of 0.812 while the accuracy of the baseline is only 0.699. These experimental results clearly demonstrate that our proposed model is capble of effectively filtering out the domain specific information so that it can generalize well and achieve good performance even on the unseen subjects and rooms.

Table 9. Performance on unseen subject and unseen room.

Room	**The Unseen Subject in a Seen Room**	**The Unseen Subject in an Unseen Room**
VR+HWC	0.751	0.699
DeepMV	0.842	0.812

4.4.7 Experiments to Differentiate the Rooms/Subjects and Their Combinations. Finally, we conduct three experiments to respectively classify the rooms, subjects, and their combinations using our model variant **VR+HWC** (without the domain discriminator). Here we shuffle all the data segments collected in the three rooms, and randomly split them into the training set (66.7% of the data) and the testing set (33.3% of the data). Table 10 reports the classification accuracy for the three experiments.

Table 10. Experiments to differentiate the rooms, the subjects and their combinations.

Classification	**3 Rooms**	**8 Subjects**	**24 Rooms-subjects Combinations**
VR+HWC	0.998	0.596	0.658

From Table 10, we can see that the model VR+HWC can achieve an accuracy of 0.998 when differentiating the three rooms, which implies that the collected data contain substantial domain information related to different rooms and our model is able to classify different rooms. For the experiment to differentiate the 8 subjects, our model can achieve an accuracy of 0.596. The result shows that the collected data also contain some domain information related to different subjects. Please note that in this experiment the subjects are not required to wear the same clothes in different rooms. This may be the reason why the accuracy of classifying different subjects is not as high as room classification. For the classification of 24 different rooms-subjects combinations, our model can still perform well and obtain an accuracy of 0.658, which demonstrates that the collected data contain the domain-specific information related to different room-subjects combinations.

4.4.8 Feasibility to Build a Real Time System. To demonstrate the feasibility of our proposed approach in reality, we empirically study the efficiency of our system. In our system, we use 3 computers (each of them connects to 3 different antennas) and 3 smartphones to collect the WiFi data and the acoustic data, respectively. For every 3 seconds, each computer can send about $378KB$ WiFi data to the server, and each smartphone can send about $517KB$ acoustic data to the server. After receiving the data, the server will concurrently preprocess the received data, and feed the preprocessed data into the well-trained model to obtain the final results. So we consider three aspects when evaluating the efficiency of our proposed system: the data transmitting time, the data preprocessing time, and the deep learning model running time. For the data transmitting time, it takes about $300ms$ to transmit both WiFi and acoustic data ($2.62MB$ in total). For the data preprocessing time, we empirically show that the server needs $583ms$ to preprocess the $378KB$ Wi-Fi data from each computer and $1383ms$ to preprocess the $517KB$ acoustic data from each smartphone. Since data preprocessing can be conducted concurrently on a multi-core CPU server, the latency of this step is equal to the preprocessing time of acoustic data, i.e., $1383ms$. As for the deep learning model running time, it takes about $18ms$ on average. The total time for the three aspects is about $300ms + 1383ms + 18ms = 1701ms$, which is much less than the data collection time i.e., $3000ms$. Thus, our proposed system is much efficient in practice.

5 RELATED WORK

The problem and methodologies presented in this paper are highly related to the following two research areas: device-free human activity recognition and multi-view learning.

5.1 Device-Free Human Activity Recognition

Human activity recognition has been a hot topic for quite a long time. However, the traditional methods such as vision based [16, 56, 85] and special device based [3, 12, 43, 45, 57] methods either have privacy and complexity problems or require subjects to wear extra devices. Currently, more and more researchers begin to utilize wireless signals (e.g., Acoustic, WiFi) to implement device-free human activity recognition process. Below I introduce some representative work in this area.

- **RSS-based methods:** As an indication of power level being received at the receiver, received signal strength (RSS) can be used to measure the distance as well as the channel condition between the transmitter and receiver. Some research work [2, 63] propose to recognize human activities by analyzing the RSS change in a special space. For example, in [2], by analyzing the RSS change caused around mobile device, the authors were able to recognize subjects' in-air hand gestures. Furthermore, the authors in [62] made full use of both the RSS change and the 3D topology of the wireless sensor network to implement the human activity recognition process.
- **CSI-based methods:** As a widely adopted channel property of communication link, CSI can reflect the combined effects of scatting, fading and even the power delay with distance. That is, compared with RSS, CSI can capture the fine-grained changes of wireless channels. Because of the release of Linux 802.11n CSI Tool [32], recently a lot of research work [22, 30, 47, 50, 69, 79, 81, 84, 96] have been conducted to utilize CSI to implement human activity recognition. Most of the existing work focuses their attention on the analysis of the information provided by a single sender-receiver pair. For example, [75] utilize directional antenna to capture the change of CSI caused by speakers' lips to identify speakers' spoken words. By analyzing the CSI change caused by users' typing behavior, [6] can even classify different keystrokes. And [76, 95] attempt to use Fresnel Zone to conduct human respiration detection and human activity recognition.
- **Acoustic-based methods:** Given that the sound frequency generated from commercial speakers can achieve as high as about 20 kHz, considering the Doppler effect caused by the relative movements between human and the speakers, in some research work [14, 21, 25, 31, 61], the authors can recognize human

gestures and activities with the help of analyzing frequency shift over a continuous time interval. Moreover, some researchers used a single smart phone to recognize the keystroke the user types just by calculating the time-difference-of-arrival of the received acoustic signals.

- **Light-based methods:** Considering that each human activity can have different continues shadow maps, some research work [8, 48, 49] can recognize human activities or gestures by analyzing those continue shadow maps.

Different from the above work, which either explore the information from just a single view or combine multi-view data using naive methods, in this paper, we develop a systematic and general multi-view deep learning framework that can benefit a wide spectrum of activity recognition applications.

5.2 Multi-View Learning

Multi-view learning has been widely studied for many years, though still having many challenges to deal with. Traditionally, Canonical Correlation Analysis (CCA) [38] and its kernel extensions [9, 33] are commonly used multi-view learning techniques. Later, they are further improved by incorporating the techniques of topic learning [15], sparse coding [41, 51] and Markov networks[18, 86]. However, these shallow models are suffering when dealing with rapidly increasing data sizes and dimensions. Recently, deep learning [34, 35] models have drawn significant attention, due to its strong ability of feature extraction, model generalization and denoising. The denoising autoencoders [73] and its followed work [74] are proposed to explicitly remove the noise in the data. Recently, CNN based [4, 11] and GAN based [70] imaging denoising methods are proposed to remove noise from noisy image. And there are some existing work using deep learning model to handle heterogeneous data. Based on the classic Restricted Boltzmann Machine (RBM) model, some researchers designed multi-modal Deep Boltzmann Machine (DBM) to fuse different views (e.g., imagery and text) in a high-order feature space [23, 28, 39, 55, 65, 67]. Multi-modal deep autoencoders were also proposed to learn the shared representation among different/heterogeneous modalities (views) [7, 37, 54, 59, 64, 77, 92]. In the field of automatic driving, the objects detection, tracking and imaging are always achieved by adjusting the network structure and fusing heterogeneous inputs (e.g., radar, camera, and lidar) [13, 17, 46]. Specifically in the field of HAR, CNN/DNN based multimodal deep architecture is proposed in [58] to interpret user activity and context captured by multi-sensor systems. In most of the aforementioned multi-view deep models, the joint representations are learned mainly using the parameter sharing architecture. In contrast, we derive the combined representation by explicitly fusing the view representations according to the relative significance of each view, which can not be captured by existing models. In addition, DeepSense [89] is proposed to leverage the representation power of both CNN and RNN to solve classification and regression problems. The architecture of DeepSense includes three layers of local CNN, three layers of global CNN and two layers of GRU. Generally, there are three differences between our model and DeepSense. First, our model is able to handle heterogeneous data due to the flexible parameter settings. Second, our DeepMV model takes the view quality into consideration by learning different view weights. Third, the domain adaptation technique is incorporated in our model which makes it more robust to the changing environment. Some previous work [88, 90, 91] also use the attention mechanism to capture the qualities of different sensors. However, they are sensitive to environment changing since they do not take the domain information into consideration.

5.3 Domain Adversarial Training

Domain adversarial training is attracting significant attentions recent years because it provides a unified architecture to jointly perform feature learning, domain adaptation and classifier learning [27]. [5, 26, 27] are the first to use adversarial networks to tackle the domain adaptation problem. Different from most of the previous domain adaptation approaches which mainly worked with fixed feature representations, the domain adversarial network

achieves the domain adaptation through learning a good nonlinear feature representation which is discriminative on the classification task but invariant to the change of domain. This framework shows its advantage in some practical cross-domain classification tasks [71, 72]. Zhao *et al.* [97] contributes to the area through proposing a conditional adversarial architecture which retains the information related to the classification tasks in the feature representation when removing the domain-specific information from it. However, this architecture is deigned for supervised tasks without taking unlabeled data into consideration. [42] further proposes an unsupervised domain adaptation approach called EI, which imposes constraints on the unlabeled data and the training process to boost the performance. In spite of the remarkable achievement previous work have got, they did not explore the design of the feature extractor when using the domain adversarial network. Usually they use an off-the-shelf network structure as the the feature extractor. In contrast, in this paper we propose a feature extractor with a view representation module and a hierarchically-weighted-combination module which is able to achieve better domain adversarial training performance than existing approaches.

6 CONCLUSIONS

Device-free human activity recognition has become a hot research topic due to its considerable advantages, such as the elimination of the need for users to wear dedicated sensors. Most of the existing work in this area only utilizes the data collected from one pair of transmitter and receiver. However, in reality, the same subject can be observed by multiple different types of wireless devices, each of which can be regarded as a view. To unleash the power of multi-view information, in this paper, we propose a unified deep learning framework, named DeepMV, to extract informative features from heterogeneous device-free data. In order to improve the performance of human activity recognition, the proposed DeepMV model is able to combine the complementary information of multiple views by incorporating the weighted-combination features and extract common representations shared across different environments. For validation, a real-world testbed is built using commercialized WiFi and acoustic devices. Experimental results on the collected activity datasets show that DeepMV can achieve better results than the stat-of-the-art device-free human activity recognition approaches, and hence justify the effectiveness of our proposed DeepMV model for the task of human activity recognition.

ACKNOWLEDGMENTS

This work was supported in part by the US National Science Foundation under Grants IIS-1924928, IIS-1938167, OAC-1934600 and CNS-1652503. And we gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

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

[3] Fadel Adib, Zachary Kabelac, Dina Katabi, and Robert C Miller. 2014. 3D Tracking via Body Radio Reflections.. In *NSDI*, Vol. 14. 317–329.

[4] Byeongyong Ahn and Nam Ik Cho. 2017. Block-matching convolutional neural network for image denoising. *arXiv preprint arXiv:1704.00524* (2017).

[5] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. 2014. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446* (2014).

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

[7] Hadi Amiri, Philip Resnik, Jordan Boyd-Graber, and Hal Daumé III. 2016. Learning text pair similarity with context-sensitive autoencoders. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1882–1892.

[8] Chuankai An, Tianxing Li, Zhao Tian, Andrew T Campbell, and Xia Zhou. 2015. Visible light knows who you are. In *Proceedings of the 2nd International Workshop on Visible Light Communications Systems*. ACM, 39–44.

[9] Francis R Bach and Michael I Jordan. 2002. Kernel independent component analysis. *Journal of machine learning research* 3, Jul (2002), 1–48.

[10] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[11] Steve Bako, Thijs Vogels, Brian McWilliams, Mark Meyer, Jan Novák, Alex Harvill, Pradeep Sen, Tony Derose, and Fabrice Rousselle. 2017. Kernel-predicting convolutional networks for denoising Monte Carlo renderings. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 97.

[12] Sourav Bhattacharya and Nicholas D Lane. 2016. From smart to deep: Robust activity recognition on smartwatches using deep learning. In *Pervasive Computing and Communication Workshops (PerCom Workshops), 2016 IEEE International Conference on*. IEEE, 1–6.

[13] Mario Bijelic, Fahim Mannan, Tobias Gruber, Werner Ritter, Klaus Dietmayer, and Felix Heide. 2019. Seeing Through Fog Without Seeing Fog: Deep Sensor Fusion in the Absence of Labeled Training Data. *arXiv preprint arXiv:1902.08913* (2019).

[14] Nguyen Dang Binh. 2015. Sound Waves Gesture Recognition for Human-Computer Interaction. In *International Conference on Context-Aware Systems and Applications*. Springer, 41–50.

[15] David M Blei and Michael I Jordan. 2003. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 127–134.

[16] Robert Bodor, Bennett Jackson, and Nikolaos Papanikolopoulos. 2003. Vision-based human tracking and activity recognition. In *Proc. of the 11th Mediterranean Conf. on Control and Automation*, Vol. 1. Citeseer.

[17] Simon Chadwick, Will Maddern, and Paul Newman. 2019. Distant vehicle detection using radar and vision. *arXiv preprint arXiv:1901.10951* (2019).

[18] Ning Chen, Jun Zhu, and Eric P Xing. 2010. Predictive subspace learning for multi-view data: a large margin approach. In *Advances in neural information processing systems*. 361–369.

[19] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in neural information processing systems*. 577–585.

[20] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[21] Amit Das, Ivan Tashev, and Shoaib Mohammed. 2017. Ultrasound based gesture recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 406–410.

[22] Shihong Duan, Tianqing Yu, and Jie He. 2018. WiDriver: Driver Activity Recognition System Based on WiFi CSI. *International Journal of Wireless Information Networks* (2018), 1–11.

[23] Max Ehrlich, Timothy J Shields, Timur Almaev, and Mohamed R Amer. 2016. Facial attributes classification using multi-task representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 47–55.

[24] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073* (2016).

[25] Biying Fu, Jakob Karolus, Tobias Grosse-Puppendahl, Jonathan Hermann, and Arjan Kuijper. 2015. Opportunities for activity recognition using ultrasound doppler sensing on unmodified mobile phones. In *Proceedings of the 2nd international Workshop on Sensor-based Activity Recognition and Interaction*. ACM, 8.

[26] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*. 1180–1189.

[27] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.

[28] Liang Ge, Jing Gao, Xiaoyi Li, and Aidong Zhang. 2013. Multi-source deep learning for information trustworthiness estimation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 766–774.

[29] Rohit Girdhar and Deva Ramanan. 2017. Attentional pooling for action recognition. In *Advances in Neural Information Processing Systems*. 33–44.

[30] Xiaonan Guo, Bo Liu, Cong Shi, Hongbo Liu, Yingying Chen, and Mooi Choo Chuah. 2017. WiFi-Enabled Smart Human Dynamics Monitoring. (2017).

[31] Sidhant Gupta, Daniel Morris, Shwetak Patel, and Desney Tan. 2012. Soundwave: using the doppler effect to sense gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1911–1914.

[32] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. 2011. Tool release: Gathering 802.11 n traces with channel state information. *ACM SIGCOMM Computer Communication Review* 41, 1 (2011), 53–53.

[33] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation* 16, 12 (2004), 2639–2664.

[34] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural computation* 18, 7 (2006), 1527–1554.

[35] Geoffrey E Hinton and Ruslan R Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *science* 313, 5786 (2006), 504–507.
[36] Tin Kam Ho. 1995. Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on*, Vol. 1. IEEE, 278–282.
[37] C. Hong, J. Yu, J. Wan, D. Tao, and M. Wang. 2015. Multimodal Deep Autoencoder for Human Pose Recovery. *IEEE Transactions on Image Processing* 24, 12 (Dec 2015), 5659–5670.
[38] Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika* 28, 3/4 (1936), 321–377.
[39] Haifeng Hu, Bingquan Liu, Baoxun Wang, Ming Liu, and Xiaolong Wang. 2013. Multimodal DBN for Predicting High-Quality Answers in cQA portals.. In *ACL (2)*. 843–847.
[40] Che-Wei Huang and Shrikanth Shri Narayanan. 2017. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In *Multimedia and Expo (ICME), 2017 IEEE International Conference on*. IEEE, 583–588.
[41] Yangqing Jia, Mathieu Salzmann, and Trevor Darrell. 2010. Factorized latent spaces with structured sparsity. In *Advances in Neural Information Processing Systems*. 982–990.
[42] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. 2018. Towards Environment Independent Device Free Human Activity Recognition. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 289–304.
[43] Bryce Kellogg, Vamsi Talla, and Shyamnath Gollakota. 2014. Bringing Gesture Recognition to All Devices.. In *NSDI*, Vol. 14. 303–316.
[44] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[45] Nicholas D Lane, Petko Georgiev, and Lorena Qendro. 2015. DeepEar: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 283–294.
[46] Vladimir Lekic and Zdenka Babic. 2019. Automotive radar and camera fusion using Generative Adversarial Networks. *Computer Vision and Image Understanding* 184 (2019), 1–8.
[47] Hong Li, Wei Yang, Jianxin Wang, Yang Xu, and Liusheng Huang. 2016. WiFinger: talk to your smart devices with finger-grained gesture. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 250–261.
[48] Tianxing Li, Qiang Liu, and Xia Zhou. 2016. Practical human sensing in the light. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 71–84.
[49] Tianxing Li, Xi Xiong, Yifei Xie, George Hito, Xing-Dong Yang, and Xia Zhou. 2017. Reconstructing hand poses using visible light. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 71.
[50] Jian Liu, Yan Wang, Yingying Chen, Jie Yang, Xu Chen, and Jerry Cheng. 2015. Tracking vital signs during sleep leveraging off-the-shelf wifi. In *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 267–276.
[51] Weifeng Liu, Dacheng Tao, Jun Cheng, and Yuanyan Tang. 2014. Multiview Hessian discriminative sparse coding for image annotation. *Computer Vision and Image Understanding* 118 (2014), 50–60.
[52] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1903–1911.
[53] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 807–814.
[54] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 689–696.
[55] Lei Pang and Chong-Wah Ngo. 2015. Mutlimodal learning with deep boltzmann machine for emotion prediction in user generated videos. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 619–622.
[56] Ronald Poppe. 2010. A survey on vision-based human action recognition. *Image and vision computing* 28, 6 (2010), 976–990.
[57] Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel. 2013. Whole-home gesture recognition using wireless signals. In *Proceedings of the 19th annual international conference on Mobile computing & networking*. ACM, 27–38.
[58] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. 2018. Multimodal Deep Learning for Activity and Context Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 157.
[59] Sarah Rastegar, Mahdieh Soleymani, Hamid R Rabiee, and Seyed Mohsen Shojaee. 2016. Mdl-cw: A multimodal deep learning framework with cross weights. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2601–2609.
[60] Mengye Ren and Richard S Zemel. 2017. End-to-end instance segmentation with recurrent attention. *arXiv preprint arXiv:1605.09410* (2017).
[61] Wenjie Ruan, Quan Z Sheng, Lei Yang, Tao Gu, Peipei Xu, and Longfei Shangguan. 2016. AudioGest: enabling fine-grained hand gesture detection by decoding echo signal. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 474–485.

[62] Markus Scholz, Till Riedel, Mario Hock, and Michael Beigl. 2013. Device-free and device-bound activity recognition using radio signal strength. In *Proceedings of the 4th Augmented Human International Conference*. ACM, 100–107.
[63] Stephan Sigg, Shuyu Shi, Felix Buesching, Yusheng Ji, and Lars Wolf. 2013. Leveraging RF-channel fluctuation for activity recognition: Active and passive systems, continuous and RSSI-based signal features. In *Proceedings of International Conference on Advances in Mobile Computing & Multimedia*. ACM, 43.
[64] Carina Silberer and Mirella Lapata. 2014. Learning Grounded Meaning Representations with Autoencoders.. In *ACL (1)*. 721–732.
[65] Kihyuk Sohn, Wenling Shang, and Honglak Lee. 2014. Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*. 2141–2149.
[66] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.
[67] Nitish Srivastava and Ruslan Salakhutdinov. 2014. Multimodal Learning with Deep Boltzmann Machines. *Journal of Machine Learning Research* 15 (2014), 2949–2980. http://jmlr.org/papers/v15/srivastava14b.html
[68] Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. LSTM-based deep learning models for non-factoid answer selection. *arXiv preprint arXiv:1511.04108* (2015).
[69] Sheng Tan and Jie Yang. 2016. WiFinger: leveraging commodity WiFi for fine-grained finger gesture recognition. In *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 201–210.
[70] Subarna Tripathi, Zachary C Lipton, and Truong Q Nguyen. 2018. Correction by projection: Denoising images with generative adversarial networks. *arXiv preprint arXiv:1803.04477* (2018).
[71] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. 2015. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*. 4068–4076.
[72] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. 4.
[73] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*. ACM, 1096–1103.
[74] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research* 11, Dec (2010), 3371–3408.
[75] Guanhua Wang, Yongpan Zou, Zimu Zhou, Kaishun Wu, and Lionel M Ni. 2016. We can hear you with wi-fi! *IEEE Transactions on Mobile Computing* 15, 11 (2016), 2907–2920.
[76] Hao Wang, Daqing Zhang, Junyi Ma, Yasha Wang, Yuxiang Wang, Dan Wu, Tao Gu, and Bing Xie. 2016. Human respiration detection with commodity wifi devices: do user location and body orientation matter?. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 25–36.
[77] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. 1083–1092.
[78] Wei Wang, Alex X Liu, and Muhammad Shahzad. 2016. Gait recognition using wifi signals. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 363–373.
[79] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2015. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st annual international conference on mobile computing and networking*. ACM, 65–76.
[80] Wei Wang, Alex X Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, 82–94.
[81] Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu. 2014. 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*. ACM, 617–628.
[82] Yuxi Wang, Kaishun Wu, and Lionel M Ni. 2017. Wifall: Device-free fall detection by wireless networks. *IEEE Transactions on Mobile Computing* 16, 2 (2017), 581–594.
[83] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.
[84] Dan Wu, Daqing Zhang, Chenren Xu, Hao Wang, and Xiang Li. 2017. Device-Free WiFi Human Sensing: From Pattern-Based to Model-Based Approaches. *IEEE Communications Magazine* 55, 10 (2017), 91–97.
[85] Lu Xia, Chia-Chih Chen, and JK Aggarwal. 2012. View invariant human action recognition using histograms of 3d joints. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 20–27.
[86] Eric P Xing, Rong Yan, and Alexander G Hauptmann. 2012. Mining associated text and images with dual-wing harmoniums. *arXiv preprint arXiv:1207.1423* (2012).
[87] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*. 2048–2057.

[88] Hongfei Xue, Wenjun Jiang, Chenglin Miao, Ye Yuan, Fenglong Ma, Xin Ma, Yijiang Wang, Shuochao Yao, Wenyao Xu, Aidong Zhang, et al. 2019. DeepFusion: A Deep Learning Framework for the Fusion of Heterogeneous Sensory Data. In *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*. 151–160.

[89] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. 2017. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 351–360.

[90] Shuochao Yao, Yiran Zhao, Shaohan Hu, and Tarek Abdelzaher. 2018. QualityDeepSense: Quality-Aware Deep Learning Framework for Internet of Things Applications with Sensor-Temporal Attention. In *Proceedings of the 2nd International Workshop on Embedded and Mobile Deep Learning*. ACM, 42–47.

[91] Shuochao Yao, Yiran Zhao, Huajie Shao, Dongxin Liu, Shengzhong Liu, Yifan Hao, Ailing Piao, Shaohan Hu, Su Lu, and Tarek F Abdelzaher. 2019. SADeepSense: Self-Attention Deep Learning Framework for Heterogeneous On-Device Sensors in Internet of Things Applications. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 1243–1251.

[92] Ye Yuan, Guangxu Xun, Kebin Jia, and Aidong Zhang. 2017. A multi-view deep learning method for epileptic seizure detection using short-time fourier transform. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 213–222.

[93] Ye Yuan, Guangxu Xun, Fenglong Ma, Qiuling Suo, Hongfei Xue, Kebin Jia, and Aidong Zhang. 2018. A Novel Channel-aware Attention Framework for Multi-channel EEG Seizure Detection via Multi-view Deep Learning. In *Biomedical & Health Informatics (BHI), 2018 IEEE EMBS International Conference on*. IEEE.

[94] Ye Yuan, Guangxu Xun, Fenglong Ma, Yaqing Wang, Nan Du, Kebin Jia, Lu Su, and Aidong Zhang. 2018. Muvan: A multi-view attention network for multivariate temporal data. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 717–726.

[95] Fusang Zhang, Kai Niu, Jie Xiong, Beihong Jin, Tao Gu, Yuhang Jiang, and Daqing Zhang. 2019. Towards a diffraction-based sensing approach on human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 1–25.

[96] Jin Zhang, Bo Wei, Wen Hu, and Salil S Kanhere. 2016. Wifi-id: Human identification using wifi signal. In *Distributed Computing in Sensor Systems (DCOSS), 2016 International Conference on*. IEEE, 75–82.

[97] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. 2017. Learning sleep stages from radio signals: a conditional adversarial architecture. In *International Conference on Machine Learning*. 4100–4109.

[98] Rui Zhou, Xiang Lu, Pengbiao Zhao, and Jiesong Chen. 2017. Device-Free Presence Detection and Localization With SVM and CSI Fingerprinting. *IEEE Sensors Journal* 17, 23 (2017), 7990–7999.