Adversarial Multi-view Networks for Activity Recognition

LEI BAI, The University of New South Wales, Australia
LINA YAO, The University of New South Wales, Australia
XIANZHI WANG, University of Technology Sydney, Australia
SALIL S. KANHERE, The University of New South Wales, Australia
BIN GUO, Northwestern Polytechnical University, China
ZHIWEN YU, Northwestern Polytechnical University, China

Human activity recognition (HAR) plays an irreplaceable role in various applications and has been a prosperous research topic for years. Recent studies show significant progress in feature extraction (i.e., data representation) using deep learning techniques. However, they face significant challenges in capturing multi-modal spatial-temporal patterns from the sensory data, and they commonly overlook the variants between subjects. We propose a Discriminative Adversarial MUlti-view Network (DAMUN) to address the above issues in sensor-based HAR. We first design a multi-view feature extractor to obtain representations of sensory data streams from temporal, spatial, and spatio-temporal views using convolutional networks. Then, we fuse the multi-view representations into a robust joint representation through a trainable Hadamard fusion module, and finally employ a Siamese adversarial network architecture to decrease the variants between the representations of different subjects. We have conducted extensive experiments under an iterative left-one-subject-out setting on three real-world datasets and demonstrated both the effectiveness and robustness of our approach.

CCS Concepts: • **Human-centered computing** → *Ubiquitous and mobile computing design and evaluation methods*; • **Computing methodologies** → *Artificial intelligence.*

Additional Key Words and Phrases: Activity Recognition, Deep Learning, Multi-view Representation

ACM Reference Format:
Lei Bai, Lina Yao, Xianzhi Wang, Salil S. Kanhere, Bin Guo, and Zhiwen Yu. 2020. Adversarial Multi-view Networks for Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 0, 0, Article 1 (January 2020), 22 pages. https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

1 INTRODUCTION

Human Activity Recognition (HAR) has been a long-standing problem in ubiquitous computing and human-computer interaction. Among the related techniques, sensor-based HAR plays an irreplaceable role in many scenarios such as elderly assisted living, sports monitoring, and surgical operation recording, given its easiness of deployment and better privacy [25]. The main tasks of sensor-based HAR involve partitioning multi-variate data streams from one or more sensors into segments and assigning an appropriate activity label to each segment [43].

Authors' addresses: Lei Bai, The University of New South Wales, Sydney, Australia, baisanshi@gmail.com; Lina Yao, The University of New South Wales, Sydney, Australia; Xianzhi Wang, University of Technology Sydney, Sydney, Australia; Salil S. Kanhere, The University of New South Wales, Sydney, Australia; Bin Guo, Northwestern Polytechnical University, Xi'an, China; Zhiwen Yu, Northwestern Polytechnical University, Xi'an, China.

2474-9567/2020/1-ART1 $15.00
https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

Previous studies in the area represent segments of raw sensory streams using hand-crafted features in statistical (e.g., mean, variance) and frequency (e.g., power spectral density) domains [6, 35] and project feature vectors to activity labels based on traditional machine learning models such as Support Vector Machine (SVM) and Random Forest. This way, the performance of those methods highly depends on the effectiveness of the extracted features, which are heuristic, task-independent, and not specially designed for HAR [49]. Since crafting HAR-specific features is labor-intensive and requires significant domain knowledge [14], recent studies leverage the exceptional data representation ability of deep learning methods to expedite feature extraction. Such studies [14, 39, 40, 47] typically use deep neural networks, e.g., Convolution Neural Networks (CNN) [37], Long-Short Term Memory (LSTM) [23], as the feature extractor to learn representations of the input sensory segments automatically; then, they map the representations to corresponding labels using another neural network (usually a fully-connected layer). The whole process is trained and tested in an end-to-end manner.

Although deep learning models can well capture spatial and temporal correlations, they face several challenges when handling multi-modal sensor streams from diverse subjects (i.e., users). First, unlike image data, which inherently have 2-D or 3-D structures, segments of multiple sensor streams have limited structural information. Current studies either regard a segment as a 2-D matrix [14, 39, 40, 47] or consider each time-step as a 1-D vector [18, 38, 49]. Such practices cannot maximally preserve the structural information in data and thus cannot fully capture the multi-modal spatial-temporal correlations within sensory streams. Second, current studies overlook the discrepancies between subjects in performing activities—they usually map the input data from all subjects indiscriminately to high-level feature representations with neural network based representation module directly. However, users usually perform the same class of activities differently due to their varied personal characteristics, such as habits and physical strength, which makes the corresponding sensory data highly disparate. This requires learning a stable discriminative representation that is robust to such an intra-class variability. Third, current studies either apply subject-dependent settings (where the testing samples may appear in the training set [13, 50]) or select only one target user as the test case. Such evaluation protocols are prone to biased results and inconsistent with the real-world deployment scenarios, where the HAR algorithms will encounter users that have varying personal traits with limited training data. To ensure wide applicability of HAR methods, it is essential to check that they have high robustness and generalization ability to achieve sound accuracy, irrespective of the subjects.

In this work, we propose a novel adversarial multi-view network to address the above challenges. The idea is to design a more robust feature extractor and to reduce the gap between the representation spaces of different subjects explicitly. We first organize data segments into various structures with specific meanings by looking at the input segments from different views and then employ CNN [37] to extract distinct multi-modal spatial-temporal representations under each view; finally, we empirically choose three most influential views and merge their representations into a joint representation by giving them weights adaptively via a trainable Hadamard fusion module. Besides, we design a Siamese adversarial architecture that integrates the generative adversarial training process with activity classification to enforce a unified representation space among subjects. Furthermore, we analyze the shortcomings of existing protocols for evaluating HAR models and propose an Iterative Left-One-Subject-Out (ILOSO) protocol to reduce biases in evaluating the recognition performance of an HAR model. In a nutshell, we make the following contributions in this paper:

- We propose a Discriminative Adversarial MUlti-view Network (DAMUN) for sensor-based HAR. The network can learn more comprehensive spatial-temporal representations from raw sensory data and meanwhile decrease the discrepancies between subjects.
- We represent multivariate sensory streams from multiple views and extract spatial-temporal correlations from each view with a 2-D CNN module. The view-specific representations are then combined via a trainable Hadamard fusion module to obtain a robust joint representation of data segments.

- We design a Siamese adversarial network to minimize the discrepancies between the sensory data representations of different subjects, and thereby ensure all subjects are presented in a consistent representation space. By explicitly decreasing the subject divergence, it improves the generalization ability of HAR methods to new users.
- We take the ILOSO strategy to evaluate both the robustness and generalization ability of HAR models. Our extensive experiments on three datasets demonstrate the effectiveness of our approach. The experiments also show the existence of some “hard-to-distinguish" subjects and “hard-to-distinguish" activities, which impact the recognition performance significantly.

The rest of this paper is organized as follows. Section 2 introduce the related work. Section 3 formulates the multi-modal sensor-based HAR problem. Section 4 gives the details of the proposed DAMUN. Section 5 reports our experimental settings and evaluation results. Finally, we conclude the paper in Section 6.

2 RELATED WORK

In this section, we first briefly introduce the general background of human activity recognition and then elaborate on deep learning methods for sensor-based HAR.

2.1 Human Activity Recognition

In general, human activity recognition is divided into two lines of research. One is vision-based HAR and another is sensor-based HAR. Vision-based solutions have been extensively studied in the past decades. For instance, Khurana et.al [31] present a vision-based system to track exercise of multiple users in the gym environment with a single camera. To detect movements, they extract optical flow trajectories from the video and then convert the trajectories to 27 features for exercise classification. Ahuja et.al [1] develop a comprehensive sensing system that detects the actions and interactions between instructors and students based on visual and audio signals. Considering the complexity associated with labeling the sensory data, Radu and Henne [44] propose to annotate the sensor data with video and computer vision classifiers automatically. They transfer the knowledge between computer vision and mobile sensing data by updating the sensor-based activity recognition model with vision generated labels. Current human activity recognition schemes are severely limited by the variability and complexity of the activities, Heilbron et.al [9] introduce a large scale video benchmark for human activity understanding that contains 203 simple and complex activities to ease the problem. Ke et.al [30] provide a survey about the vision based human activity recognition.

Sensor-based human activity recognition also has a long-standing history in the ubiquitous and wearable research community [18]. Compared with vision-based, sensor-based approaches own some unique merits in terms of pervasiveness, low computational complexity and better privacy preserving [15]. In the meanwhile, it makes a good complement to vision-based solutions. Most early studies in the area aim to recognize human activities from solely accelerometer signals [6, 19, 34]. These methods generally have unreliable performance because the single modal information cannot characterize the variability and diversity between different activities. Recent studies move their attention to recognize human activity with more signals sources, especially, gyroscopes and magnetometer. Extracting and fusing the useful multi-modal information are twos essential steps. These efforts can be divided into early fusion and late fusion, depending on when the fusion is conducted. Early fusion methods combine the raw data from different sources together and process them as an integral entity [28, 39, 40, 43, 47]. On the contrary, late fusion methods extract information from each modality separately at first and then merge the information with an additional ensemble layer [19, 38, 48]. On the other hand, according to how the information is extracted, the previous work can be divided into feature-based methods and deep learning-based methods. Feature-based methods focus on designing hand-crafted features to capture the data distribution of each activity. Besides the most frequently used time-domain features (e.g., mean, variance, and

skewness) and frequency domain features (e.g., power spectral density) [27], a few methods endeavor to design new features containing temporal and structural information. For example, Hammerla et al. [21] propose the Empirical Cumulative Density Function (ECDF) feature to preserve the spatial information of the signal frames. Kwon et al. [35] enhance this work by adding temporal structures to ECDF and get improved results. Different from feature-based methods, deep learning-based methods target on designing neural networks to get the robust representation of sensory data automatically.

2.2 Deep Learning Methods for Sensor-based HAR

2.2.1 Temporal Correlations in HAR. Most deep learning-based HAR methods focus on capturing the temporal correlations. Yang et al. [47] tackle the problem with convolutional neural networks, in which the convolution and pooling filters are designed along the temporal dimensions to process the readings of all sensors. Their work can capture long-term temporal correlation by stacking multiple CNN layer. Ordóñez et al. [40] further extend this model to DeepConvLSTM by integrating LSTM [23] to CNN. The proposed DeepConvLSTM framework contains four CNN layers and two LSTM layers to capture the short-term and long-term temporal correlations, separately. Based on this framework, Peng et al. [42] propose to divide human activities into simple activities and complex activities, which are closely linked. Several consecutive simple activities form one complex activity and share the LSTM network in the DeepConvLSTM model. The model recognizes simple activities and complex activities at the same time and is optimized under the multi-task learning paradigm. Instead of designing new models, Hammerla et al. [20] explore the performance of basic temporal convolution network, LSTM, and Bi-directional LSTM in capturing the temporal correlation. However, the best results vary from dataset to dataset. To advance SOTA in HAR, Guan and Plötz [18] use the ensemble method and combine multiple LSTM networks into a powerful model. They employ the Epoch-wise Bagging scheme in the training procedure and select the LSTM in different training epochs as basic learners. In the testing phase, the outputs of these selected LSTM learners are fused together through score level fusion and obtain improved performance.

The above models have one drawback in that they assume the signals in all time steps are relevant contribute equally to the target activity, which may not always stand as the sensory data may contain activity inconsequential components [49]. Based on this observation, Murahari and Plötz [39] extend the DeepConvLSTM [40] by adding the temporal attention module after the LSTM layer. The attention mechanism [46] aligns the output vector at the last time step with other vectors at earlier steps to learn a relative importance score for each previous time step. The final representation of the segment is the sum of re-weighted representations at each time step. In this way, the model can be trained to focus on more important parts to learn a more accurate representation. Similarly, Zeng et al. [49] use the LSTM to extract temporal correlations and apply the attention mechanism on the last LSTM layer to highlight the important part of time-series signals. Considering that time-series signals are continuous, they claim that nearby signals should have consecutive significance scores and add a continuity regularization to the attention scores.

2.2.2 Spatial-Temporal Correlations in HAR. Besides temporal correlations, there is also spatial correlations among sensors. Some recent studies have started to capture both spatial and temporal correlations for HAR. Jiang et al. [28] first propose to treat the sensory streams segment as an image and process it by multiple CNN layers with 2D kernels for capturing the spatial and temporal correlations. Chen et al. [12] propose the RAAF model to capture the spatial and temporal correlations by cascade CNN and LSTM network. In the RAAF, data from different sensors are combined and organized to activity frame through complex permuting operation to ensure each pair of sensors are adjacent. Then, CNN networks are applied to the frame for extracting spatial correlation followed by a glimpse network and LSTM network to capture the temporal correlation.

Our work differs from the previous work by comprehensively capturing spatial, temporal, and spatial-temporal correlations. Instead of solely representing the raw data as a 1D vector or 2D frame, we fuse and organize the

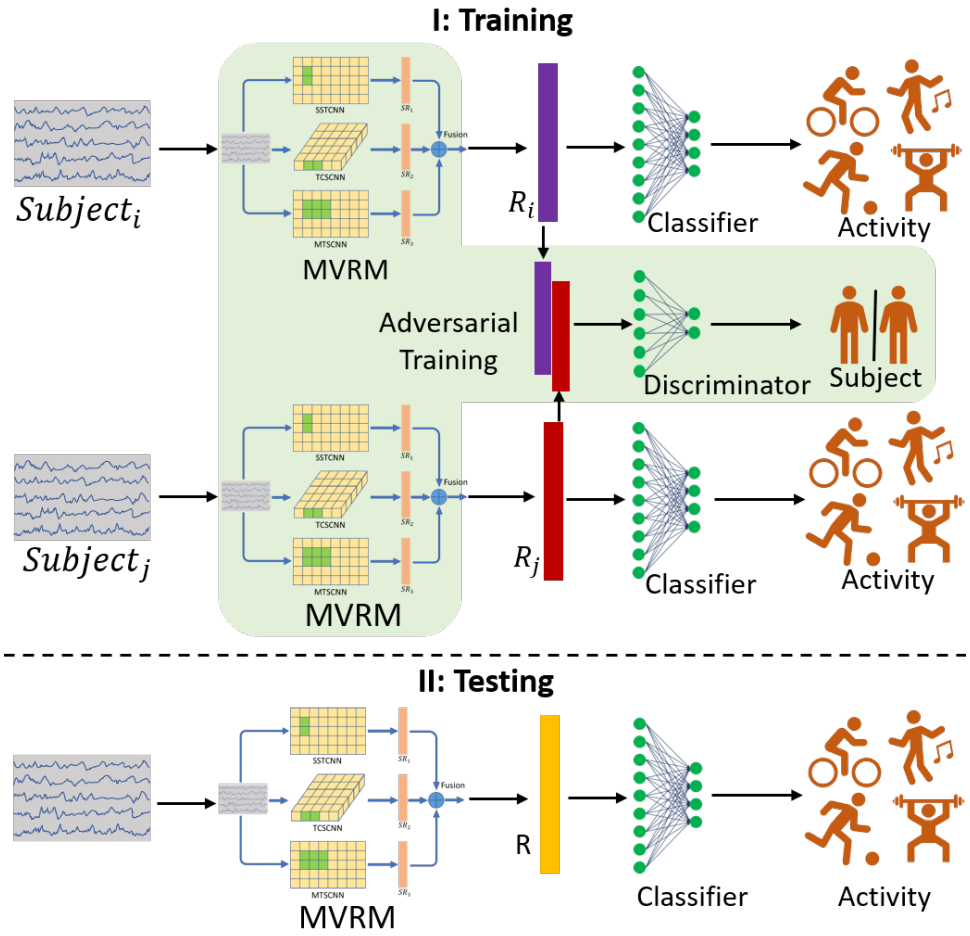
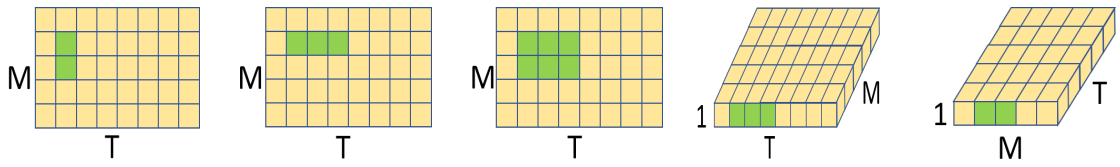


Fig. 1. Discriminative Adversarial Multi-view Network.

multi-modal data into a 2D matrix and 3D matrix from multiple views, followed by employing CNN networks with different kernels to capture spatial-temporal correlations. The learned representations contain both spatial correlations, temporal correlations, and spatial-temporal correlations, respectively. We further design a trainable Hadamard fusion module to merge the representations from different views considering their importance. Besides the robust multi-view data representation module, we further consider and explicitly model the divergence between different subjects' representation space, which is not included in the previous work. The details will be elaborated in the following.

3 PROBLEM DEFINITION

Recognizing human activity with multi-modal data involves multiple devices, e.g., smartphone, inertial measurement unit (IMU) attached to different parts of the human body. Each device carries multiples sensors, e.g., an IMU normally contains 3-axis accelerometer, 3-axis gyroscope, and 3-axis magnetometer. To simplify the



(a) STS (b) SST (c) MST (d) SCT (e) TCS

Fig. 2. Different views of data streams, which reflect different types of correlations. STS, SST, MST, SCT, TCS denote Single-Time-Spatial correlation, Single-Sensor-Temporal correlation, Multiple-Sensor-Temporal correlation, Sensor-as-Channel-Temporal correlation, and Time-as-Channel-Spatial correlation, respectively.

description, we refer to a 3-axis device as three sensors in this paper. For instance, the 3-axis accelerometer contains x-accelerometer, y-accelerometer, and z-accelerometer. Thus, a typical inertial measurement unit contains nine sensors. Let M be the total number of sensors embedded in multiple devices and $s_{i,t}$ be the reading of sensor i ($1 \leq i \leq M$) at time point t, then the readings of sensor i along the time axis is a time series: $\boldsymbol{s}_{i,:} = [s_{i,1}, s_{i,2}, s_{i,3}..., s_{i,t}, ...]$. Similarly, the readings at time point t of all sensors along the spatial axis is a vector: $\boldsymbol{s}_{1:M,t} = [s_{1,t}, ..., s_{i,t}, ..., s_{M,t}]$. The readings of all sensors along time axis form a multi-variate time series.

Following the standard HAR procedure [8], we divide the multi-variate sensory streams into segments with a fixed-size sliding window. Suppose the window size is T, a segment can be both represented as $\boldsymbol{Seg} = [\boldsymbol{s}_{1,1:T}, \boldsymbol{s}_{2,1:T}, ..., \boldsymbol{s}_{M,1:T}]$ and $\boldsymbol{Seg} = [\boldsymbol{s}_{1:M,1}, \boldsymbol{s}_{1:M,2}, ..., \boldsymbol{s}_{1:M,T}]$. Let there be N potential activities to be recognized, $C = \{c_1, c_2, ..., c_N\}$, the purpose of HAR is to learn a function, $\mathcal{F}(\boldsymbol{Seg}, \bullet)$, to infer the correct activity label for the given segment Seg, where $\bullet$ represents all the parameters to be learned during the training process.

4 METHODOLOGY

Our proposed Discriminative Adversarial MUlti-view Network (DAMUN) (Fig. 1) includes three components: (i) a robust multi-view representation module (MVRM), which learns comprehensive multi-modal representations of the input segments from different views; (ii) a classifier, which classifies the input representation and outputs the related activity label; and (iii) a Siamese adversarial framework, which decreases the divergence between subjects and enforces a consistent representation space over all subjects. The Siamese adversarial framework reuses MVRM as the generator and further introduces a discriminator. At the training stage, we train MVRM, the classifier, and the discriminator simultaneously, following the Siamese adversarial training process to obtain robust sensor segment representations module for accurate classification. At the testing stage, we directly use MVRM and the classifier for activity classification Fig. 1.II) without the discriminator.

In the following, we will elaborate on MVRM (Section 4.1), the Siamese Adversarial Framework (Section 4.2), and the adversarial training and optimization process (Section 4.3). We omit details about the classifier as it simply consists of two fully-connected layers.

4.1 Robust Multi-view Representation

We first design a multi-view representation module (a.k.a., feature extractor) for learning robust multi-modal representations from multiple views of data streams. As introduced in Section 3, the input data to the human activity recognition model contains multiple time series. Different from the image data, which inherently have fixed structural information, the segment only contains limited structural information (time series). The limited structural information character provides researchers the feasibility to organize the segment into different structures to extract corresponding correlations with different neural networks. In the literature, Multi-Layer

Perception (MLP), RNN (e.g., LSTM, GRU) and CNN are the three most widely used basic neural networks, based on which more complicated methods are developed. MLP formulates the general interactions within the input and do not specially focus on spatial or temporal correlations [49]. RNN processes the input of each time step sequentially and are good at capturing the temporal correlation in the time-series data with "Memory" mechanism [42]. At the same time, CNN can capture diverse spatial, temporal, or spatial-temporal correlations according to the data manipulated by different kernels. For example, CNN networks with 2-D kernels in the computer vision [41] and urban computing [4] area are considered to extract spatial correlations between different pixels/regions. On the other hand, CNN networks with 2-D kernels in the multi-variate time-series domain are considered to extract correlations between different series and different time points (spatial-temporal correlations) [36]. Based on these observations, we propose to organize the input segment into different 2-D or 3-D matrix and apply CNN networks with different kernels to extract distinct spatial-temporal correlations without losing valuable information. Each pair of segment organization and kernel shape can capture correlations from one view. Specially, we consider extracting correlations from the following five views (shown in Fig. 2, where M denotes the number of sensors, and T denotes the size of the sliding window):

- STS View: This view represents a segment $\boldsymbol{Seg}$ by a 2D matrix ($M \times T$) and applies 1D CNN kernels along the sensors axis. As the kernels operate on the readings of different sensors at the same time point, it can capture the spatial pattern among sensors shared by each time point.
- SST View: This view also represents a segment $\boldsymbol{Seg}$ as a 2D matrix but instead, applies 1D CNN kernels along the time axis. The kernels operate on readings of the same sensor at different time points and thus can model the common individual temporal correlation of one sensor shared by all sensors.
- MST View: This view applies 2D CNN kernels along both sensor and time axes to capture the spatial and temporal correlations simultaneously.
- SCT View: This view represents a segment by a 3D matrix ($1 \times T \times M$) and conducts 2D CNN operations. It focuses on extracting the temporal correlations collectively for all sensors by treating all sensors' readings at one time-point as a unit (channel).
- TCS View: This view represents a segment by a 3D matrix ($1 \times M \times T$) and conducts 2D CNN operations. It focuses on extracting the spatial correlations between different sensors' data streams collectively.

We denote the CNN networks implemented based on the above five views as STSCNN, SSTCNN, MSTCNN, SCTCNN, and TCSCNN, respectively.

4.1.1 CNN Network Architecture. Besides the kernel shapes, these CNN networks used in each view follows the same architecture. As shown in Fig. 3, each network contains three consecutive CNN blocks, which is comprised of a convolutional layer with rectified linear units (ReLu) activation function, a max_pooling layer and a batch normalization layer. The convolutional layer performs the main function of pattern extraction, which employs several kernels of the same shape to filter the input data X and extract meaningful patterns. Formally, we formulate a convolution layer with the ReLu activation function as follows:

$$X_j^{l+1} = \sigma(\sum_{i=1}^{i=F^l} W_{i,j} * X_i^l + b_j^l) \tag{1}$$

where X_i^l is the i_{th} channel of the input for the l_{th} convolutional layer, F^l is the feature map (channel) numbers, $W_{i,j}$ is the j_{th} kernel, b_j^l is the bias and $\sigma(\cdot)$ is the ReLu function defined as: $\sigma(X^{l+1}) = max(0, X^{l+1})$. Then, the max pooling layer is employed as the sampling method to down-sampling the extracted representations while keeping the most prominent patterns.

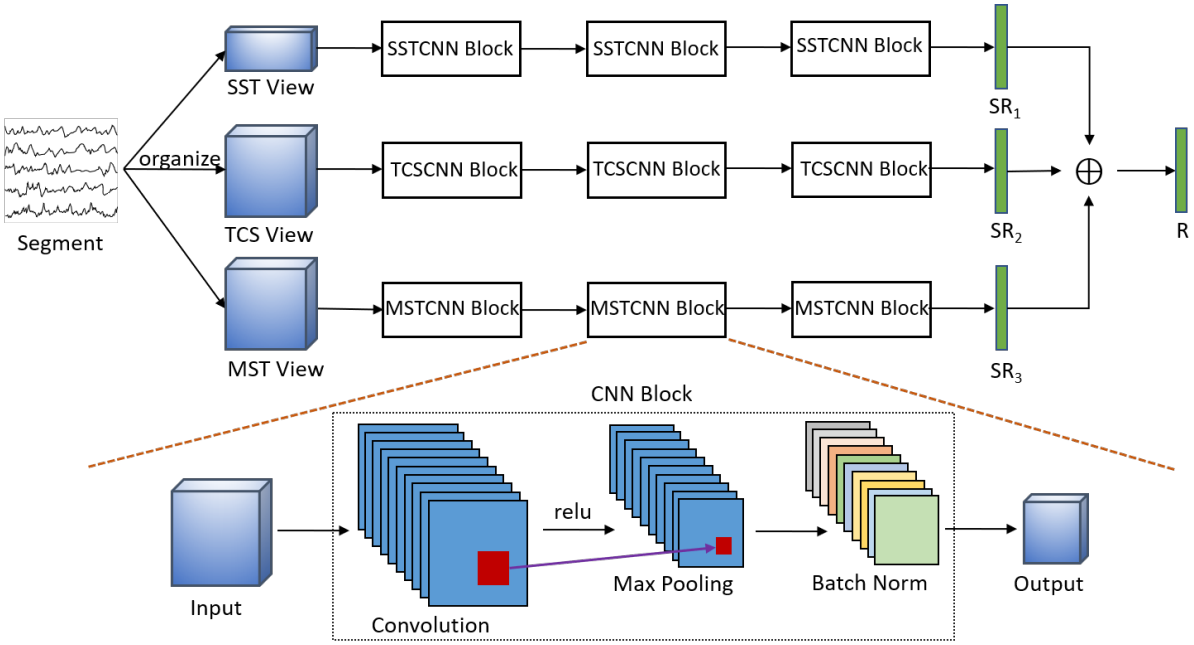


Fig. 3. Multi-View data Representation Module (MVRM)

We further integrate the batch normalization layer [26] to the CNN blocks to achieve faster and more stable training. The batch normalization layer normalizes the layer input with batch mean and batch variance to force the input of every layer to have similar distributions [26]:

$$X = \gamma \frac{X - \mu}{\sqrt{\vartheta^2 + \epsilon}} + \beta \tag{2}$$

where μ and ϑ are the mean and variance of a batch of input for the layer, respectively, γ and β are the trainable parameters in the batch normalization layer.

4.1.2 Trainable Hadamard Fusion Module. We empirically prove that the SST view, MST view, and TCS view can lead to better classification accuracy than the STS view and SCT view (as shown in Section 5.3). Based on the observation, we develop our Multi-View Representation Module (MVRM) to obtain a robust representation of the input segments. As shown in Fig. 3, MVRM consists of three of the above CNN networks: SSTCNN, TCSCNN, and MTSCNN, which process the input segment separately and each generates a sub-representation. We merge the sub-representations of different views using a trainable Hadamard Fusion Module to ensure the final representation contains the most robust and discriminative features in the signals:

$$R = \alpha_1 * SR_1 + \alpha_2 * SR_2 + \alpha_3 * SR_3 \tag{3}$$

where SR_1, SR_2, and SR_3 are three sub-representations generate from the three views shown in Fig. 3; α_1, α_2, and α_3 are trainable importance scores for the three views, respectively. The learned representation R not only contains different types of spatial and temporal correlations but also takes into account the importance of correlations.

4.2 Siamese Adversarial Framework

We then design a novel Siamese adversarial framework forcing MVRM to learn a consistent representation space for all subjects. The framework aims to overcome the varied characteristics of people in performing activities. People tend to perform activities in different ways due to their different characteristics, such as gender, height, weight, and strength. One example is that men usually perform activities at a larger magnitude than women. However, existing learning techniques assume identical distributions of the training and testing data, which does not hold for new users' activities. Neglect of such distribution shifts could degrade a model's generalization ability.

Specifically, we deploy two representation modules with the same parameters in the Siamese architecture, considering the sensory data of two arbitrary subjects (illustrated in Fig. 1), we denote by R_i and R_j the learned representations of the two input subjects. We train a Generative Adversarial Network (GAN) [17] to enforce the identical distribution of R_i and R_j. The generative adversarial training process takes the representation modules as generators and further introduces a discriminator in the latent space: the discriminator conducts binary classification to distinguish between R_i and R_j, while the generator tries to fool the discriminator. Specifically, we employ the Wasserstein GAN (WGAN) [3] rather than the original GAN—WGAN uses the Earth-Mover distance (Wasserstein-1), instead of Jensen-Shannon divergence (used by GAN), to measure the divergence between R_i and R_j, and thus achieves better training stability.

Let $g_\theta(\cdot)$ be the representation module with parameter θ, $\boldsymbol{U_i}$ and $\boldsymbol{U_j}$ be the original sensory data space of $subject_i$ and $subject_j$. Then the representation spaces of $subject_i$ and $subject_j$ are $\mathbb{P}_\theta(\boldsymbol{U_i})$ and $\mathbb{P}_\theta(\boldsymbol{U_j})$, respectively. The Earth-Mover Distance between $\mathbb{P}_\theta(\boldsymbol{U_i})$ and $\mathbb{P}_\theta(\boldsymbol{U_j})$ can be formulated by [3]:

$$W(\mathbb{P}_\theta(\boldsymbol{U_i}), \mathbb{P}_\theta(\boldsymbol{U_j})) = \sup_{||d(\cdot)||_L \leq 1} \mathbb{E}_{r\sim\mathbb{P}_\theta(U_i)}[d(r)] - \mathbb{E}_{r\sim\mathbb{P}_\theta(U_j)}[d(r)] \tag{4}$$

where $||\cdot||_L$ is the Lipschitz constant, and the supremum operation is over all the 1-Lipschitz functions $||d(\cdot)||$. If replacing $||d(\cdot)||_{L\leq 1}$ with $||d(\cdot)||_{L\leq K}$ (consider K-Lipschitz function) and have a family of functions $\{d_w(\cdot)\}_{w\in\mathcal{W}}$ parameterized by w that are all K-Lipschitz for some K [3], we can get:

$$W(\mathbb{P}_\theta(\boldsymbol{U_i}), \mathbb{P}_\theta(\boldsymbol{U_j})) \propto \max_{w\in\mathcal{W}} \mathbb{E}_{x\sim U_i}[d_w(g_\theta(x))] - \mathbb{E}_{z\sim U_j}[d_w(g_\theta(z))] \tag{5}$$

where x and z and sensory data segments sampled from $\boldsymbol{U_i}$ and $\boldsymbol{U_j}$ separately. The generator should minimize the distance between $\mathbb{P}_\theta(\boldsymbol{U_i})$ and $\mathbb{P}_\theta(\boldsymbol{U_j})$ to make the representation of $subject_i$ and $subject_j$ in the same space. To this end, we define the loss function for the generator as follows:

$$\mathcal{L}_g(\theta; x, z) = \mathbb{E}_{x\sim U_i}[d_w(g_\theta(x))] - \mathbb{E}_{z\sim U_j}[d_w(g_\theta(x))] \tag{6}$$

Likewise, we define the loss function of the discriminator:

$$\mathcal{L}_d(w; x, z) = -\mathbb{E}_{x\sim U_i}[d_w(g_\theta(x))] + \mathbb{E}_{z\sim U_j}[d_w(g_\theta(x))] \tag{7}$$

4.3 Training and Optimization

To ensure the learned consistent representation space can lead to accurate activity classification, we jointly minimize the classification error and conduct adversarial training. In each iteration, we sample and feed arbitrary two subjects' sensory data to MVRM (shown in Fig. 1), and then train a classifier, which takes both representations, R_1 and R_2, as the input. We optimize the classification error by minimizing the cross-entropy loss:

$$\mathcal{L}_c(\theta, \phi; x, y) = -\mathbb{E}_{x\sim S_i\cup S_j}[y * log(f_\phi(g_\theta(x)))] \tag{8}$$

where y is the activity label for $x \in (\boldsymbol{S_i} \cup \boldsymbol{S_j})$. We then update the discriminator and generator by minimizing $\mathcal{L}_d(w; x, y)$ and $\mathcal{L}_g(\theta; x, y)$. Through repeatly approximating the representation spaces of arbitrary two subjects,

Algorithm 1 Training and Optimization

Require: the training set $\boldsymbol{L} = \{(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{u})\}$ ($\boldsymbol{u}$ is the subjects set in training), batch size B, maximum training iteration $Iter$, and the number of classifier training iteration $iter_c$.

```
{θ, φ, w} = RandomInitialize()
for iter = 0; iter < Iter do
    Random choose two subjects u_1 and u_2 from u
    for n = 0; n < iter_c do
        Sample a batch {x_i}_{i+1}^B from U_1 ∪ U_2
        Compute L_c(θ, φ; x, y) with Equation 8
        Backpropagate loss and update θ, φ
    end for
    Sample a batch {x_i}_{i+1}^B from U_1
    Sample another batch {z_i}_{i+1}^B from U_2
    Compute L_d(w; x, z) with Equation 7
    Backpropagate loss and update w
    Sample a batch {x_i}_{i+1}^B from U_1
    Sample another batch {z_i}_{i+1}^B from U_2
    Compute L_g(θ; x, z) with Equation 6
    Backpropagate loss and update θ
end for
```

MVRM obtains the ability to map all subject's data to a consistent space. Since the objective of our work is to classify activities, we emphasize the task by repeating updating $\mathcal{L}_c$ for $iter_c$ times in each iteration to force the model reduce the classification loss and thus obtain higher training accuracy. Algorithm 1 describes the detailed training procedure.

5 EXPERIMENTS

This section reports our experiments on three sensor-based human activity datasets. We first discuss the evaluation protocols used in previous work and introduce our updated protocol for achieving a thorough evaluation. Then the experimental datasets and settings are elaborated. Finally, we evaluate the performance of our proposed method in comparison with state-of-the-art and via ablation studies.

5.1 Evaluation Protocol

The evaluation protocol is critical for creating discriminative human activity recognition methods with better generalization ability. Inappropriate or impractical evaluation protocols may lead to biases or misunderstanding conclusions. Considering whether new users are exploited for testing, previous evaluation protocols can be divided into subject-dependent and subject-independent ones. The subject-dependent protocols depend on subjects because they randomly divide the data of all subjects into a training and testing sets. Thus, the training set and testing set contain data from the same subject set. While 10-fold cross-validation can be utilized under such settings to evaluate the models multiple times independently, the subject-dependent setting does not conform to the real deployment, where the HAR models normally serve new users. The work [10, 34, 50] utilizes subject-dependent protocols and can easily achieve high recognition accuracy because of the information leakage problem (samples in the training set and testing set may come from neighboring windows and contain extremely similar information).

Most recent studies [19, 35, 38–40, 43, 49] deploy subject-independent settings in the evaluation to avoid the information leakage problem and evaluate the models' generalization ability on new subjects. These studies use the leave-one-subject-out evaluation protocol, which selects one subject's data as the testing set and all the other subjects' data as the training set. Under the LOSO setting, the data in the testing set and training set belong to different subjects, which makes it more difficult to achieve high recognition accuracy due to the inconsistency between the training and testing sets. While the LOSO evaluation protocol can check both the model's performance in capturing spatial-temporal correlations for recognition and generalization ability for new users, current works don't evaluate the models completely. Specially, current works employ an impaired LOSO setting which only test the model in one selected subject (for example, [38, 39, 43, 49] use subject six for testing in the PAMAP2 dataset [45]). This impaired setting is biased because a model's performance may vary significantly from one subject to another subject [29].

Considering the shortcomings in previous evaluations, we propose to assess the model iteratively with the LOSO protocol on each subject separately. We name the upgraded evaluation protocol by Iterative-LOSO (ILOSO). In each iteration, we train the model from scratch and test the model with one subject's data. Finally, we will get *subject_number* results for each model. We believe ILOSO is more suitable for analyzing and comparing the performance of HAR models because the averaged results of all subjects are less biased but more comprehensive than the sole result of one selected subject.

5.2 Dataset Description

While several datasets are publicly available for HAR, many of them only contain a few subjects (e.g. the Ubicomp 08 dataset [24] only has one subject) or human activities (e.g. the UCI dataset [2] only contains six activities). To demonstrate the superior ability of our method in classifying activities and dealing with subject divergence, we select three datasets with relatively more activities and subjects:

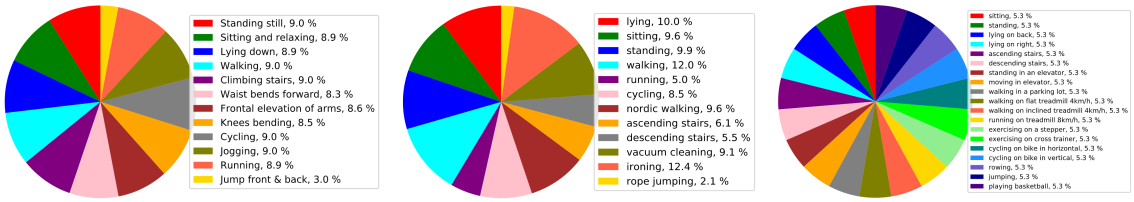
- MHEALTH Dataset[1] This dataset [5] contains body motion and vital signs recordings for ten volunteers of diverse profiles. Each subject performed 12 activities in an out-of-lab environment with no constraints. Three inertial measurement units (IMUs) were placed on the subject's chest, right wrist, and left ankle to measure a 3-axis acceleration (m/s^2), a 3-axis gyroscope (deg/s), and a 3-axis magnetic field (local) of the motion, respectively. Besides, the IMU positioned on the chest also provided 2-lead ECG measurements (mV). All sensing modalities were recorded at a frequency of 50Hz.
- PAMAP2 Dataset[2] The original PAMAP2 dataset [45] was designed to benchmark daily physical activities. It contains data collected from nine subjects related to 18 daily activities such as vacuum cleaning, ironing, and rope jumping. Similar to the MHEALTH dataset, the data was collected with three IMUs placed on the subject's chest, dominant wrist, and dominant ankle, respectively, under the sampling frequency of 100Hz.
- UCIDSADS Dataset[3] The UCIDSADS Dataset [7] was specially devised for daily and sports activities. It comprises the motion sensor data of 19 daily and sports activities such as walking on a treadmill, exercising on a stepper, and rowing. Each activity was performed by eight subjects for five minutes in their own style without constraints. Five units on the torso and the four limbs are calibrated to acquire data at the sampling frequency of 25Hz. Each unit contains nine sensors: 3-axis accelerometers, 3-axis gyroscopes, and 3-axis magnetometers.

Table 1 provides some statistics of these three datasets. For the MHEALTH and UCIDSADS datasets, we use all subjects' data in our experiments. For the PAMAP2 dataset, we remove six activities (watching TV, computer work, car driving, folding laundry, house cleaning, and playing soccer) because they are executed by only one

[1] http://archive.ics.uci.edu/ml/datasets/mhealth+dataset
[2] https://archive.ics.uci.edu/ml/datasets/pamap2+physical+activity+monitoring
[3] https://archive.ics.uci.edu/ml/datasets/daily+and+sports+activities

Table 1. Statistics of datasets (# denotes the "number").

Dataset	Subject#	Activity#	Frequency	Window size	Devices#	Sensors#	Sample#
MHEALTH	10	12	50 Hz	20	3	23	34 097
PAMAP2	8	12	100 Hz	20	3	36	191 309
UCIDSADS	8	19	25 Hz	20	5	45	113 848



(a) mhealth (b) pamap2 (c) ucidsads

Fig. 4. Activity distribution of the three datasets (best viewed in color)

subject; therefore, 12 activities from eight subjects are retained for our experiments. We divide the raw sensory data streams into small segments with fixed-sized sliding windows and an overlap of 50% for all the three dataset. Instead of using a large window size (5 seconds) for PAMAP2 (as in [18, 20]), we set each window to contain 20 time points for the three datasets, resulting in window lengths of 0.4 second, 0.2 second, and 0.8 second for MHEALTH, PAMAP2, and UCIDSADS, respectively. A smaller window size implies collecting and processing less data during each recognition, which helps reduces the latency of the recognition system and higher-level applications. Our experimental results will show that the window size is sufficient for achieving competitive performance. Fig. 4 shows that the class distributions of the MHEALTH and PAMAP2 datasets are both imbalanced—the minority class “Jump front & back" has a much lower count than other classes in the MHEALTH dataset; and the ratio of “walking" and “ironing" outnumber “rope jumping" significantly in the PAMAP2 dataset.

5.3 Experimental Settings

We normalize the datasets using standard normalization method after dividing the streams into segments. The normalized segments are fed into the network directly without sophisticated operations like hand-crafted feature extraction or Fourier transformation [38]. We initialize the network parameters with Xavier Normal initialization [16] and optimize them by Adam optimizer [32] at a learning rate of 0.0001 for all the three datasets. We set the Batch_size to 320 to reduce the training time and the length of learned representations to 128 for each segment. We implement the model based on Pytorch running on an NVIDIA TITAN X Pascal GPU. We use weighted Accuracy (Acc), Precision (Pre), and F-measure (F_w) as the performance metrics for the evaluation and report the mean result, worst result, and best result of all subjects as $mean[worst, best]$, which can reflect both the overall performance and the generalization ability of HAR models. While the widely used weight metrics can better reflect the overall performance, they may not be effective enough on highly imbalanced datasets such as PAMAP2 dataset (shown in Fig. 4). In this regard, we additionally report the classification confusion matrix to investigate the performance of each class and thus gain deeper insights into the model (Section 5.8).

Table 2. Evaluation on different views

MHEALTH	View	STS	SST	MST	SCT
	Acc	93.67 [84.72, 98.34]	94.01 [87.17, 99.35]	95.45 [90.40, 99.24]	92.89 [82.08, 98.94]
	View	TCS	SST+MST+TCS	SST+MST+TCS+SCT	ALL
	Acc	94.20 [87.14, 98.87]	**96.07 [92.29, 99.50]**	95.73 [88.53, 99.67]	**96.59 [92.29, 99.73]**
PAMAP2	View	STS	SST	MST	SCT
	Acc	72.30 [47.87, 87.61]	81.76 [58.70, 94.14]	77.58 [52.77, 92.83]	74.17 [40.16, 91.70]
	View	TCS	SST+MST+TCS	SST+MST+TCS+SCT	ALL
	Acc	75.77 [52.09, 92.77]	**83.21 [62.64, 93.96]**	81.79 [58.10, 94.23]	80.32 [53.11, 93.57]
UCIDSADS	View	STS	SST	MST	SCT
	Acc	85.03 [81.63, 90.66]	90.68 [85.97, 94.82]	88.07 [84.34, 94.53]	85.49 [77.98, 89.29]
	View	TCS	SST+MST+TCS	SST+MST+TCS+SCT	ALL
	Acc	85.60 [78.09, 89.63]	**92.14 [87.40, 96.21]**	90.44 [84.31, 95.27]	88.61 [75.01, 94.22]

5.4 Evaluation on Different Views

We first conduct an experiment to explore the performance of each view and some of their combinations with the Siamese adversarial training process. The results (Table 2) show the SST view consistently achieves a high classification accuracy on all the three datasets, showing the importance of the single-sensor temporal correlation for human activity recognition. This is reasonable as each sensor's data is a time-series. However, the performance of the SCT view is poor, which means the collectively temporal correlation is fragile when regrading all sensors as a unit. On the other hand, the performance of the STS view is among the worst, showing the weakness of the spatial correlations among different sensors at each time point. However, the MST and TCS view performs moderately, inditing the existence of spatial correlations among different sensors. The results of the three views (SCT view, MST view, and TCS view) also imply that a better way to organize all the sensors together (e.g., graph structure) is necessary for extracting the spatial correlation and collectively temporal correlation.

Based on these observations, we form MVRM by the SST view, MST view, and TCS view and find that MVRM outperforms single-view models by a large margin on all three datasets. Besides, due to the weakness of the STC view and the STS view, adding them to MVRM additionally will decrease the performance.

5.5 Comparison with State-of-the-art

To verify the overall performance of the proposed DAMUN model, we compare our model with the following baseline and state-of-the-arts:

- SVM [22]: the traditional support vector machine (SVM) with the radial basis function (RBF) kernel.
- MC-CNN [47]: a state-of-the-art CNN-based model formed of 3 convolutional layers with kernels along the time axis to capture temporal correlations.
- b-LSTM-S [20]: a bi-directional LSTM variant to capture both forward and backward information.
- DeepConvLSTM [40]: a hybrid model with four convolutional layers and 2 LSTM layers to capture both the spatial and temporal correlations.
- Ensemble-LSTM [18]: an LSTM-based method that combines multiple individual LSTM learners with epoch-wise bagging.
- AttConvLSTM [39]: an SOTA hybrid model integrating an attention-based output module to DeepConvLSTM for capturing temporal dynamics.

Table 3. Performance of compared methods. Each cell consists of the mean score of a method in one evaluation metric, followed by the corresponding minimum and maximum scores in brackets. The best performance values are in bold.

MHEALTH		SVM	MC-CNN	Bi-LSTM-S	DeepConvLSTM
	Acc	78.49 [60.93, 93.45]	92.72 [87.96, 97.97]	89.94 [83.39, 96.28]	91.34 [86.83, 99.21]
	Pre	79.53 [65.29, 94.47]	93.51 [84.11, 98.18]	87.16 [76.51, 95.41]	89.37 [81.48, 99.21]
	F_w	76.73 [59.33, 92.80]	92.17 [85.34, 97.98]	87.90 [79.01, 94.85]	89.89 [81.49, 99.22]
		Ensemble_LSTM	AttDeepConvLSTM	Multi-Agent	DAMUN
	Acc	86.03 [78.34, 98.56]	92.19 [84.02, 98.18]	91.57 [84.04, 98.01]	**96.07 [92.29, 99.44]**
	Pre	84.81 [74.57, 98.59]	89.96 [78.30, 98.21]	91.87 [80.51, 98.06]	**96.52 [92.38, 99.52]**
	F_w	84.64 [70.32, 98.55]	90.75 [80.36, 98.17]	91.20 [81.12, 98.01]	**96.07 [92.23, 99.50]**
PAMAP2		SVM	MC-CNN	Bi-LSTM-S	DeepConvLSTM
	Acc	69.47 [38.72, 86.41]	80.27 [58.29, 93.40]	71.51 [36.06, 92.02]	75.69 [50.55, 92.67]
	Pre	70.77 [41.69, 88.76]	80.64 [57.65, 93.82]	71.12 [29.01, 92.21]	73.04 [36.42, 92.95]
	F_w	68.11 [36.72, 86.68]	78.05 [52.09, 93.37]	68.65 [32.34, 91.94]	72.36 [41.67, 92.65]
		Ensemble_LSTM	AttDeepConvLSTM	Multi-Agent	DAMUN
	Acc	73.36 [47.79, 89.66]	74.13 [46.58, 88.62]	72.48 [35.16, 88.02]	**83.21 [62.64, 93.96]**
	Pre	73.90 [36.88, 90.93]	73.92 [50.40, 85.02]	73.35 [36.22, 89.88]	**83.57 [55.69, 94.63]**
	F_w	71.98 [42.09, 88.84]	71.83 [44.79, 86.58]	71.39 [31.70, 87.14]	**82.13 [57.31, 93.91]**
UCIDSADS		SVM	MC-CNN	Bi-LSTM-S	DeepConvLSTM
	Acc	71.95 [65.21, 80.66]	87.93 [72.69, 94.62]	88.98 [79.67,93.71]	89.66 [80.54, 94.17]
	Pre	70.60 [63.19, 78.84]	87.18 [64.01, 95.42]	89.72 [74.29, 95.25]	89.58 [79.88, 95.27]
	F_w	67.74 [60.25, 78.33]	85.52 [66.57, 94.53]	87.73 [75.36, 93.28]	88.42 [77.95, 94.08]
		Ensemble_LSTM	AttDeepConvLSTM	Multi-Agent	DAMUN
	Acc	83.13 [76.13, 90.92]	88.73 [79.03, 94.22]	85.98 [76.48, 91.21]	**92.14 [87.40, 96.21]**
	Pre	84.06 [72.65, 93.51]	88.24 [74.57, 94.78]	87.45 [79.48, 92.91]	**92.99 [86.84, 96.18]**
	F_w	81.09 [71.48, 90.19]	86.75 [74.64, 94.22]	84.26 [73.03, 90.70]	**91.59 [86.79, 95.36]**

- Multi-Agent [14]: a spatial-temporal attention method based on multi-agent collaboration to select informative modalities and their active periods.

For the state-of-the-art methods, we replicated each method with the same settings as introduced in the original papers, except the data pre-processing steps, where we use the same window size and overlap as ours. We evaluate them with the ILOSO evaluation protocol to achieve a fair and thorough comparison. Table 3 shows the experimental results.

The results show that all the methods achieve higher performance on the MHEALTH and UCIDSADS datasets than on the PAMAP2 dataset. The reason could be the data quality or the hard-recognizing activities in the PAMAP2 dataset. Besides, the performance variance of the PAMAP2 dataset is also the largest, reflecting the diversity between different subjects in PAMAP2.

All the deep learning methods perform better than the baseline model, showing the promising ability of deep learning in capture the nonlinear spatial-temporal correlations. Moreover, the MC-CNN model fits pretty good in the MHEALTH and PAMAP2 datasets but poor in the UCIDSADS dataset. This could own to the fact that the MC-CNN model only considers the single-sensor temporal correlation, which ignores spatial correlations and limits the performance when data contain more features. In addition, we notice that the complex reinforcement learning-based Multi-agent model does not work very well as reported in [14], indicating the difficulty of selecting important modalities for numerous and more complex activities compared to their background activity recognition

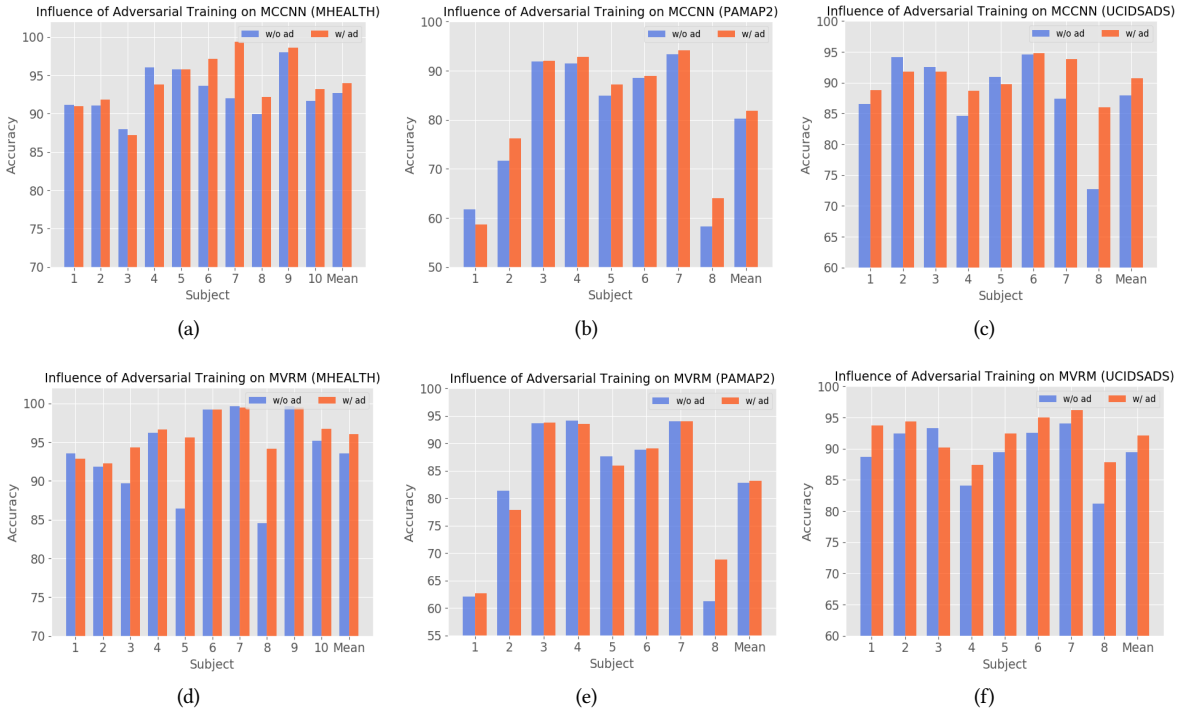


Fig. 5. Effect of the Siamese adversarial training. (a)-(c): Performance of MC-CNN with (w/ ad) or without (w/o ad) adversarial training; (d)-(f): Performance of MVRM with (w/ ad) or without (w/o ad) adversarial training (best viewed in color).

setting (only category six activities [19]). Our method consistently outperforms all the state-of-the-art methods on the three datasets. It's mean recognition accuracy achieves 3.35%, 2.94%, and 2.48% absolute improvements over the best SOTA in the MHEALTH, PAMAP2 and UCIDSADS datasets, respectively. The comparison demonstrates the effectiveness of our proposed DAMUN model.

5.6 Effect of the Siamese Adversarial Training

To further evaluate the effect of our Siamese adversarial training framework, we conduct an ablation study to check the performance of feature extractors with or without the adversarial training process. We analyze two feature extractors: MVRM and an SOTA MC-CNN model. Fig. 5 reports the classification accuracy for each subject and the mean accuracy on the three datasets. We have three observations. First, the performance of all models varies significantly among subjects—A model that works positively for one subject may not generalize well to other subjects and achieve consistent performance. This observation demonstrates that testing HAR models only on one selected subjected (as previous work does) is insufficient and could lead to wrong conclusions. Take the UCIDSADS dataset as an example: if we only select subject 3 as the testing subject, researchers may conclude that the Siamese adversarial training process will decrease the recognition performance. However, more thorough evaluations with ILOSO protocol show that the Siamese adversarial training framework can indeed improve the recognition accuracy (i.e., from 87.93% to 90.69% with MC-CNN model and from 89.44% to 92.14% with our MVRM). Second, although the average performances among all subject are pretty good for the

Table 4. Evaluation results of different fusion methods

Method	Average	Concatenate	Self-Attention	Hadamard
MHEALTH	94.86 [91.93, 99.44]	94.81 [84.52, 99.82]	94.12 [87.17, 99.41]	**96.07 [92.29, 99.44]**
PAMAP2	81.86 [62.17, 93.73]	80.98 [59.09, 93.86]	80.75 [58.95, 94.01]	**83.21 [62.64, 93.96]**
UCIDSADS	92.05 [84.42, 95.37]	91.00 [86.27, 96.50]	90.09 [80.18, 94.20]	**92.14 [87.40, 96.21]**

three datasets, there exist some hard-to-distinguish subjects (e.g., subject 8 in PAMAP2 dataset and UCIDSADS dataset). The main reason could be the subject-divergence—these hard-to-distinguish subjects conduct activities in a different manner from other subjects in the training set; as a result, the models trained on other subjects cannot be generalized to these "hard-to-distinguish" subjects. Third, while the performance of a few subjects declines, more subjects see improvements with the Siamese adversarial training framework. Overall, MVRM and MC-CNN with the Siamese adversarial training process consistently boost the performance on all datasets. These results demonstrate the effectiveness and necessity of our adversarial training framework in improving human activity recognition performance and reducing subject divergence. To interpret the results, we consider the adversarial training process as a regularizer that enforces the feature extractor to focus on subject-irrelevant common patterns when learning segment representations and thus learn a consistent representation space for all subjects. In contrast, traditional deep learning models (e.g. MC-CNN without adversarial training) can easily suffer overfitting, which leads to poor generalization ability and low recognition accuracy on the "hard-to-distinguish" testing subjects.

5.7 Effect of the Fusion Method

Finally, we also evaluate the influence of the fusion method by comparing our trainable Hadamard fusion method with other three fusion strategies: 1) average-based fusion, which sums and averages the representation of each view directly to get the final joint representation of the input segment; 2) concatenate-based fusion, which gets the final representation by concatenating representations of all views; 3) self-attention based fusion, which uses the self-attention mechanism [46] to calculate the importance of each view and merges the re-weighted sub-representations by summing. Our experimental results (Table 4) show simple average-based fusion performs better than concatenate-based fusion and self-attention based fusion. Besides, our Hadamard fusion achieves the best performance, confirming the importance of considering the significance of each view. In contrast, considering more complex interactions between representations with self-attention mechanism is not helpful.

5.8 Confusion Matrix Analysis

To better understand the activity recognition performance for each class, we further conduct a deeper analysis based on the confusion matrix. As described in the evaluation protocol part (Section 5.1), we iteratively evaluate the performance of our method on each subject separately, resulting in *subject_number* confusion matrices for each dataset. Due to the space limitation, we only report the best classified subjects (the subject that achieves the highest classification accuracy within the dataset) and the worst classified subjects of the three datasets in Fig. 6. The class labels in confusion matrices are corresponding to the activity name in Fig. 4 in the same order. We can observe that: 1) our method can work perfectly in the MHEALTH and UCIDSADS datasets for most activities with an accuracy of nearly 100%, which shows the exceptional classification performance of the proposed model; 2) the classification accuracy for the minority classes are still accurate, as observed from Fig. 6.(a) and Fig. 6.(d) for "Jump front & back" activity (class 12) in the MHEALTH dataset and from Fig. 6.(b) for "rope jumping" activity (class 12) in the PAMAP2 dataset; 3) the classification performance varies significantly between different classes. Take the worst classified subject in the UCIDSADS dataset (subject 7, Fig. 6.(f)) as an example, the classification

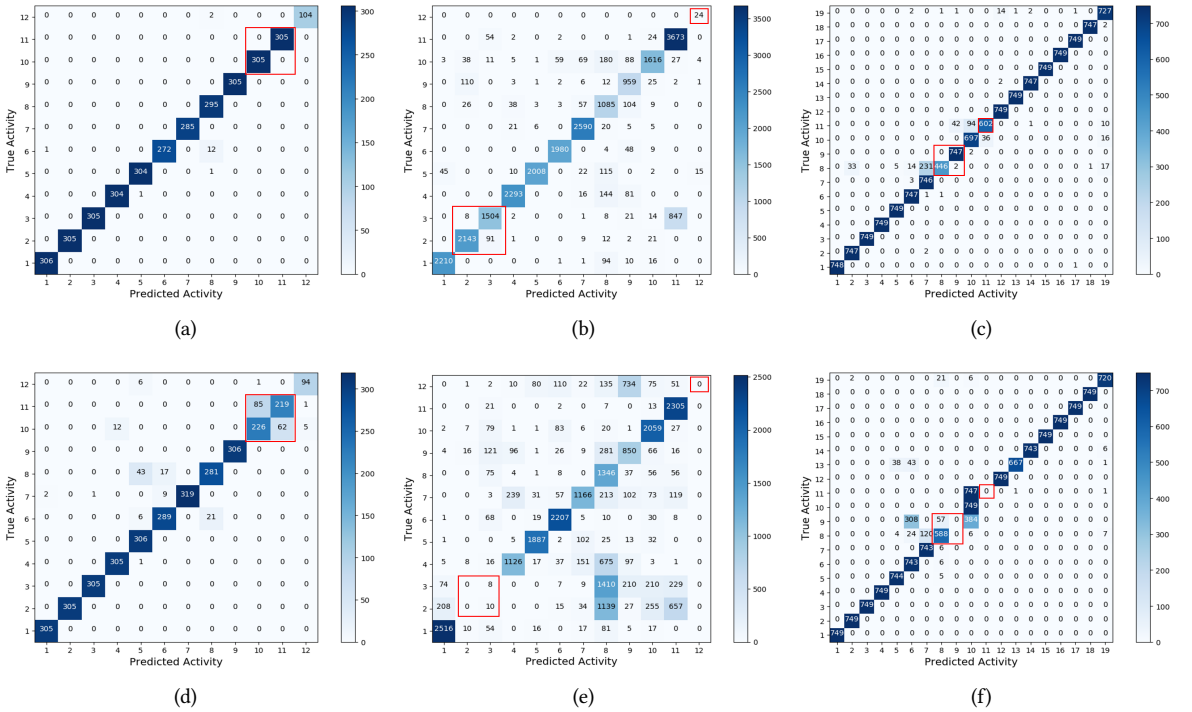


Fig. 6. (a)-(c): Confusion Matrices of the best-classified subject in the (a) MHEALTH, (b) PAMAP2, and (c) UCIDSADS datasets; (d)-(f): confusion matrix of the worst-classified subject in the (d) MHEALTH, (e) PAMAP2, and (f) UCIDSADS datasets (best viewed in color).

accuracy for most classes is very high except for class 9 and class 11. Besides, the classification performance also varies dramatically between different subjects for the same activity, as can be observed from the red rectangle area in Fig. 6); 4) Last but not the least, the poor performance on the worst classified subjects (hard-to-distinguish subjects) is mainly caused by a few “hard-to-distinguish" activities (e.g., class 10 and 11 in the MHEALTH dataset, class 2 and 3 in the PAMAP2 dataset). This suggest that we should pay more attention to “hard-to-distinguish" activities for the “hard-to-distinguish" subjects to achieve accurate and widely applicable HAR.

5.9 Case Study

To verify the overall performance of the proposed DAMUN out of the existing public datasets, we further conduct a case study in our laboratory environment. Specifically, we collect an activity dataset containing five activities (sitting, standing, walking, ascending stairs, and descending stairs) from eight participants (six male and two female). Three Phidget Spatial IMU sensors working at the frequency of 70Hz are attached to the dominant wrist, the waist, and the dominant side’s ankle to record the acceleration, angular velocity, and magnetism. The same evaluation protocol (ILOSO) and experimental settings as introduced in section 5.3 are employed for the evaluation. As shown in Fig. 7, our proposed DAMUN can still achieve positive recognition performance, and it consistently outperforms the SOTA MC-CNN model.

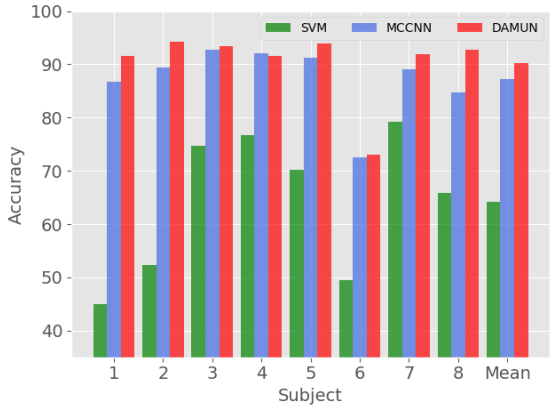


Fig. 7. Classification accuracy of SVM, MC-CNN and the proposed DAMUN in the self-collected dataset

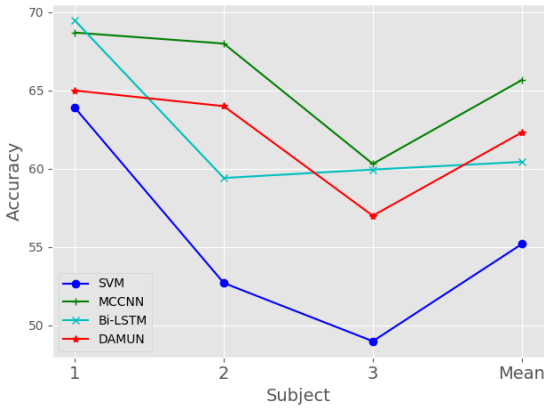


Fig. 8. Classification results in the Opportunity dataset

5.10 Discussion

In this section, we provide a brief discussion on potential of DAMUN in the recognition of sporadic activities, and shed light on the relationship between the proposed model with Siamese networks. Some future works in the area are also discussed.

5.10.1 Performance on the Sporadic Activities. We have evaluated the performance of our proposed DAMUN thoroughly on both public datasets and local deployment. The results demonstrate the superior performance of DAMUN. While these datasets cover a variety of activities in different application scenarios, they mainly are repetitive or recurring activities. To examine the efficacy of DAMUN on sporadic activities, we further carry out the experiments with the Opportunity dataset [11] containing three subjects. Fig. 8 presents the performance of DAMUN against the other three comparison methods following the introduced ILOSO evaluation protocol and experimental settings. It can be observed that the performance of all methods degrade at different levels, among which MC-CNN achieves the best performance. The performance of Bi-LSTM is similar to MC-CNN for subject 3 but worse than MC-CNN and DAMUN on average under the subject-independent ILOSO evaluation protocol. This finding reveals that CNN-based models still work better in recognizing non-repetitive activities with designated architectures, and that the previous work may be biased due to the compromised evaluation protocols. DAMUN achieves reasonable performance even though it is a bit lower than MC-CNN, which might be attributed to the small number of subjects in the Opportunity dataset. When there are only two subjects available in the training dataset, the effect of the adversarial training component is compromised as the model would only capture the patterns shared by the two training subjects instead of more general common patterns. This limitation could be improved in real-world deployment, where more training subjects are available.

5.10.2 Relation to the Siamese Network. In this section, we compare our Siamese adversarial framework with the Siamese network [33], which is a metric learning-based method in the computer vision area. Siamese network takes two samples as the inputs: one is the query; the other is the baseline. The two samples are operated by the same network to get their representations. Then, the distance between the two representations is compared by a predefined metric function to determine whether the query input belongs to the same class with the baseline. While our Siamese adversarial framework also organizes two samples, it is not a metric learning-method. Our method organizes the two samples from different subjects and uses a discriminator to impose the learned

representations only contain subject-irrelevant patterns. The framework is designed for cross-subject training during the training process to decrease the subject divergence and does not influence the classification and testing phase (as shown in Fig. 1). In summary, our Siamese adversarial framework only shares a similar appearance with the Siamese network but works differently and for a different purpose. The performance gain of our Siamese adversarial comes from the adversarial training instead of the Siamese inputs.

5.10.3 Future Works. In the previous sections, we have analysed our proposed model from different perspectives thoroughly and demonstrated the robustness and effectiveness of our design in capturing multi-modal spatial-temporal correlations and reducing subject divergence, which can facilitate the deployment of HAR algorithms into real-world applications broadly, In order to better understand the spatial-temporal correlations among the multi-modal sensors and the subject divergence, there are also some future works we would to like to highlight and explore in the next stage:

- By iteratively testing the model on each subjects without their data, our ILOSO evaluation protocol can reflect the model's generalization ability to new users and is closer to the real deployment scenario. However, it would still be overfitted as we use data from many more users for training than in the real scenario. Evaluating the performance of HAR models with different number of training subjects thoroughly is better for assessing the model's generalization performance.
- Accurate recognition of human activities and reducing subject divergence require large scale dataset from more subjects. However, collecting and labeling sensor-based human activity dataset is still a challenge. On the other hand, we observed that there are some overlaps in different sensory datasets. For instance, both the MHEALTH dataset and the UCIDSADS dataset place the IMU devices in subjects' arm and leg. How to transfer the information across datasets to lessen the data collection is a promising topic.
- Among the datasets we have tested, we found they employed different number of sensors in distinct parts of body. While deep learning methods can capture the complex spatial-correlations among these sensors automatically, the importances/benefits of each sensors are not clear. Thus, exploring the minimum number and best locations of the sensors would be an interesting and practical direction.
- Our experiments have demonstrate that there exist "hard-to-distinguish" subjects and "hard-to-distinguish" activities in all the three datasets (shown in Fig. 5 and Fig. 6). Our current approach can boost the recognition performance on some of them (e.g., subject 8 in the UCIDSADS dataset) but cannot fully solve the problem. Developing more robust and discriminative HAR models to improve the performance on "hard-to-distinguish" subjects and "hard-to-distinguish" activities is essential.

6 CONCLUSION

The main focus of this work is developing discriminative human activity recognition method that is powerful and robust to different sensor streams and generalizes well to new subjects, both of which are prerequisites for wider adoption of HAR models/algorithms in real-world applications. As such, we proposed a discriminative adversarial multi-view network for sensor-based human activity recognition. We first design a novel multi-view representation module (MVRM) to capture multi-modal spatial-temporal correlations and to obtain high-level representations of the raw sensory data. Compared with previous representation modules, MVRM can capture both the spatial, temporal and spatial-temporal correlations with multi-modal sensory streams and thus is more powerful and robust. Then, we present a Siamese adversarial training framework to ensure that MVRM can map all subjects' data to the same space, thus reducing subject discrepancies. To evaluate the HAR models more thoroughly, we introduced ILOSO evaluation protocol based on the traditional LOSO protocol and conducted extensive experiments on three public datasets and a new dataset collected in our laboratory environment. The results validate the superior performance and generalization ability of our approach to state-of-the-art. We

also discuss the performance of our model on the non-repetitive activities and give some future directions for wide-scale adoption of sensor-based HAR models in pervasive systems and applications.

REFERENCES

[1] Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. 2019. EduSense: Practical classroom sensing at Scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–26.

[2] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. 2013. A public domain dataset for human activity recognition using smartphones.. In *Esann*.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*. 214–223.

[4] Lei Bai, Lina Yao, Salil S. Kanhere, Xianzhi Wang, and Quan Z. Sheng. 2019. STG2Seq: Spatial-Temporal Graph to Sequence Model for Multi-step Passenger Demand Forecasting. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 1981–1987.

[5] Oresti Banos, Rafael Garcia, Juan A Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. 2014. mHealthDroid: a novel framework for agile development of mobile health applications. In *International workshop on ambient assisted living*. Springer, 91–98.

[6] Ling Bao and Stephen S Intille. 2004. Activity recognition from user-annotated acceleration data. In *International conference on pervasive computing*. Springer, 1–17.

[7] Billur Barshan and Murat Cihan Yüksek. 2014. Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. *Comput. J.* 57, 11 (2014), 1649–1667.

[8] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)* 46, 3 (2014), 33.

[9] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*. 961–970.

[10] Cagatay Catal, Selin Tufekci, Elif Pirmit, and Guner Kocabag. 2015. On the use of ensemble of classifiers for accelerometer-based activity recognition. *Applied Soft Computing* 37 (2015), 1018–1022.

[11] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R Millán, and Daniel Roggen. 2013. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters* 34, 15 (2013), 2033–2042.

[12] Kaixuan Chen, Lina Yao, Xianzhi Wang, Dalin Zhang, Tao Gu, Zhiwen Yu, and Zheng Yang. 2018. Interpretable parallel recurrent neural networks with convolutional attentions for multi-modality activity modeling. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[13] Kaixuan Chen, Lina Yao, Dalin Zhang, Xiaojun Chang, Guodong Long, and Sen Wang. 2019. Distributionally robust semi-supervised learning for people-centric sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3321–3328.

[14] Kaixuan Chen, Lina Yao, Dalin Zhang, Bin Guo, and Zhiwen Yu. 2019. Multi-agent attention activity recognition. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*.

[15] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. 2020. Deep Learning for Sensor-based Human Activity Recognition: Overview, Challenges and Opportunities. *arXiv preprint arXiv:2001.07416* (2020).

[16] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 249–256.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[18] Yu Guan and Thomas Plötz. 2017. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 11.

[19] Haodong Guo, Ling Chen, Liangying Peng, and Gencai Chen. 2016. Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1112–1123.

[20] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 1533–1540.

[21] Nils Y Hammerla, Reuben Kirkham, Peter Andras, and Thomas Ploetz. 2013. On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution. In *Proceedings of the 2013 International Symposium on Wearable Computers*. ACM, 65–68.

[22] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications* 13, 4 (1998), 18–28.

[23] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[24] Tâm Huynh, Mario Fritz, and Bernt Schiele. 2008. Discovery of activity patterns using topic models.. In *UbiComp*, Vol. 8. 10–19.

[25] Sozo Inoue, Naonori Ueda, Yasunobu Nohara, and Naoki Nakashima. 2015. Mobile activity recognition for a whole day: recognizing real nursing activities with big dataset. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1269–1280.

[26] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*. 448–456.

[27] Majid Janidarmian, Atena Roshan Fekr, Katarzyna Radecka, and Zeljko Zilic. 2017. A comprehensive analysis on wearable acceleration sensors in human activity recognition. *Sensors* 17, 3 (2017), 529.

[28] Wenchao Jiang and Zhaozheng Yin. 2015. Human activity recognition using wearable sensors by deep convolutional neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia*. Acm, 1307–1310.

[29] Artur Jordao, Antonio C Nazare Jr, Jessica Sena, and William Robson Schwartz. 2018. Human activity recognition based on wearable sensor data: A standardization of the state-of-the-art. *arXiv preprint arXiv:1806.05226* (2018).

[30] Shian-Ru Ke, Hoang Le Uyen 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 (2013), 88–131.

[31] 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.

[32] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[33] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, Vol. 2. Lille.

[34] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. 2011. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter* 12, 2 (2011), 74–82.

[35] Hyeokhyen Kwon, Gregory D Abowd, and Thomas Plötz. 2018. Adding structural characteristics to distribution-based accelerometer representations for activity recognition using wearables. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*. ACM, 72–75.

[36] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. 2018. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 95–104.

[37] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436.

[38] Haojie Ma, Wenzhong Li, Xiao Zhang, Songcheng Gao, and Sanglu Lu. 2019. AttnSense: multi-level attention mechanism for multimodal human activity recognition. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 3109–3115.

[39] Vishvak S Murahari and Thomas Plötz. 2018. On attention models for human activity recognition. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*. ACM, 100–103.

[40] Francisco Ordóñez and Daniel Roggen. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16, 1 (2016), 115.

[41] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2018. Spatial as deep: Spatial cnn for traffic scene understanding. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[42] Liangying Peng, Ling Chen, Zhenan Ye, and Yi Zhang. 2018. AROMA: A Deep Multi-Task Learning Based Simple and Complex Human Activity Recognition Method Using Wearable Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 74.

[43] Hangwei Qian, Sinno Jialin Pan, Bingshui Da, and Chunyan Miao. 2019. A novel distribution-embedded neural network for sensor-based activity recognition. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 5614–5620.

[44] 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.

[45] Attila Reiss and Didier Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th International Symposium on Wearable Computers*. IEEE, 108–109.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[47] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. 2015. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[48] 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.

[49] Ming Zeng, Haoxiang Gao, Tong Yu, Ole J Mengshoel, Helge Langseth, Ian Lane, and Xiaobing Liu. 2018. Understanding and improving recurrent networks for human activity recognition by continuous attention. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*. ACM, 56–63.

[50] Xiang Zhang, Lina Yao, Chaoran Huang, Sen Wang, Mingkui Tan, Guodong Long, and Can Wang. 2018. Multi-modality sensor data classification with selective attention. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 3111–3117.