Attend and Discriminate: Beyond the State-of-the-Art for Human Activity Recognition Using Wearable Sensors

ALIREZA ABEDIN, The University of Adelaide, Australia
MAHSA EHSANPOUR, The University of Adelaide, Australia
QINFENG SHI, The University of Adelaide, Australia
HAMID REZATOFIGHI, Monash University, Australia
DAMITH C. RANASINGHE, The University of Adelaide, Australia

Wearables are fundamental to improving our understanding of human activities, especially for an increasing number of healthcare applications from rehabilitation to fine-grained gait analysis. Although our collective know-how to solve Human Activity Recognition (HAR) problems with wearables has progressed immensely with end-to-end deep learning paradigms, several fundamental opportunities remain overlooked. We rigorously explore these new opportunities to learn enriched and highly discriminating activity representations. We propose: i) learning to exploit the *latent* relationships between multi-channel sensor modalities and specific activities; ii) investigating the effectiveness of *data-agnostic augmentation* for multi-modal sensor data streams to regularize deep HAR models; and iii) incorporating a classification loss criterion to encourage minimal intra-class representation differences whilst maximising inter-class differences to achieve more discriminative features. Our contributions achieves *new* state-of-the-art performance on *four* diverse activity recognition problem benchmarks with large margins—with up to 6% relative margin improvement. We extensively validate the contributions from our design concepts through extensive experiments, including *activity misalignment* measures, *ablation* studies and insights shared through both quantitative and qualitative studies. The *code* base and trained network parameters are open-sourced on GitHub https://github.com/AdelaideAuto-IDLab/Attend-And-Discriminate to support further research.

CCS Concepts: • **Human-centered computing → Ubiquitous computing**; • **Computing methodologies → Neural networks**; **Supervised learning by classification**.

Additional Key Words and Phrases: activity recognition; deep learning; attention; cross-channel interaction encoder, center-loss, data augmentation, wearable sensors; time-series data

ACM Reference Format:
Alireza Abedin, Mahsa Ehsanpour, Qinfeng Shi, Hamid Rezatofighi, and Damith C. Ranasinghe. 2021. Attend and Discriminate: Beyond the State-of-the-Art for Human Activity Recognition Using Wearable Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 1 (March 2021), 22 pages. https://doi.org/10.1145/3448083

1 INTRODUCTION

Wearable sensors provide an infrastructure-less multi-modal sensing method. Current trends point to a pervasive integration into our lives with wearables providing the basis for wellness and healthcare applications from

Authors' addresses: Alireza Abedin, alireza.abedinvaramin@adelaide.edu.au, The University of Adelaide, Adelaide SA 5005, Australia; Mahsa Ehsanpour, mahsa.ehsanpour@adelaide.edu.au, The University of Adelaide, Adelaide SA 5005, Australia; Qinfeng Shi, javen.shi@adelaide.edu.au, The University of Adelaide, Adelaide SA 5005, Australia; Hamid Rezatofighi, Hamid.Rezatofighi@monash.edu, Monash University, Australia; Damith C. Ranasinghe, damith.ranasinghe@adelaide.edu.au, The University of Adelaide, Adelaide SA 5005, Australia.

2474-9567/2021/3-ART1 $15.00
https://doi.org/10.1145/3448083

rehabilitation, caring for a growing older population to improving human performance [1, 3, 9, 10, 12, 13, 15, 23, 26, 33, 40]. Fundamental to these applications is our ability to automatically and accurately recognize human activities from often tiny sensors embedded in wearables.

Wearables capture individuals' activity dynamics by continuously recording measurements through different sensor channels over time and generate *multi-channel time-series* data streams. Consequently, the problem of human activity recognition (HAR) with wearables involves temporal localization and classification of actions embedded in the generated sequences. Adoption of deep neural networks for HAR has created pipelines for end-to-end learning of activity recognition models yielding state-of-the-art (SoA) performance [16, 28, 30, 50].

1.1 Problem and Motivations

Despite progress towards end-to-end deep learning architectures for achieving state-of-the-art performance on HAR problems, we uncover key under explored dimensions with significant potential for improving the performance of state-of-the-art frameworks:

- HAR data acquisition often involves recording of motion measurements over number of sensors and channels. Therefore, we can expect the capability of different sensor modalities and channels to capture and encode some activities better than others whilst having complex interactions between sensors, channels and activities. Thus, we hypothesize that learning to exploit the relationships between multi-channel sensor modalities and specific activities can contribute to learning enriched activity representations—*this insight remains unstudied.*
- Human actions, for example walking and walking up-stairs, exhibit significant intra-class variability and inter-class similarities. This suggests imposing optimization objectives for training that not only ensure class separability but also encourage compactness in the established feature space. However, *the commonly adopted cross-entropy loss function does not jointly accommodate both objectives*.
- Due to the laborious process of collecting annotated sequences with wearables, sensor HAR datasets are often small in size. While expanding the training data with virtual samples has proved beneficial in achieving better *generalization* performance for general machine learning problems, exploration of data augmentation for HAR has been largely limited to hand-crafted techniques that alter the data sequences with the assumption of being able to preserve the activity label semantics. However, achieving label-preserving transformations for wearable HAR sensor data is not obvious and intuitively recognizable [41]. Thus, the augmented data may not necessarily preserve salient characteristics embedded within the original data, leading to alteration of the activity labels and potentially deluding the supervised training process. For instance, in the image domain, a flipped image of a person is still a meaningful illustration of the person concept whilst applying the same method and flipping sensor channels of an inertial sensor leads to a completely different signal.

1.2 Our Contributions

Motivated by the opportunities to further HAR research, our *key contribution* is to propose a *new HAR framework* built upon multiple architectural elements and demonstrate its capability to realize new state-of-the-art performance that generalizes across multiple diverse wearable sensor datasets. We illustrate our framework in Fig. 1 and summarize our key contributions below:

(1) We *propose and design* a *cross-channel interaction encoder* to incorporate a self-attention mechanism to learn to exploit the different capabilities of sensor modalities and latent interactions between multiple sensor channels capturing and encoding activities. The encoder module captures latent correlations between multi-sensor channels to generate self-attention feature maps and enrich the convolutional feature representations (**Section 3.2** and **Fig. 9**).

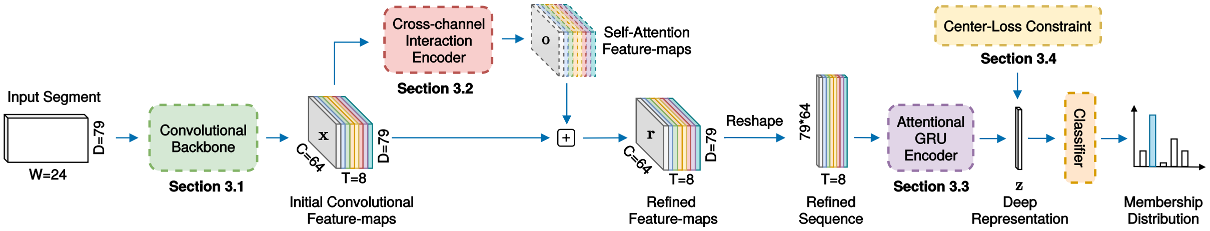


Fig. 1. An overview of our proposed HAR framework

(2) Temporal attention layers were recently shown in [28] to improve performance by capturing temporal context in a network constructed using LSTM (long short-term memory) layers capable of learning dependencies in sequences. Therefore, we *design* an *attentional GRU (gated recurrent unit) encoder* to enhance the sequence of self-attention enriched features by further capturing the relevant temporal context (**Section 3.3**). Compared with LSTMs, GRUs are easier to train and leverage fewer parameters [56].

(3) In recognizing the intra-class variations of HAR activities, we propose adopting the *center-loss criterion* to encourage minimal intra-class representation differences whilst maximising inter-class differences to achieve more discriminative features and demonstrate the effectiveness of center-loss penalisation for learning highly discriminative activity representations for wearable HAR problems (**Section 3.4).**

(4) In recognizing the difficulty of ensuring label-preserving augmentation with hand-crafted approaches in wearable HAR problems, we propose adopting *mixup* method to take into account both data and label information for multi-modal sensor data augmentation, investigate the effectiveness of the method and demonstrate the seamless integration of mixup for wearable HAR problems (**Section 3.5**). Importantly unlike existing augmentation approaches that are dataset dependent and thus require domain expert knowledge for effective adoption, mixup is domain independent and simple to apply; an important consideration for wearable HAR problems based on multiple different sensor modalities and sensor specific semantic and signal characteristics.

(5) Under a unified evaluation protocol, our proposed framework achieves significant improvements against the state-of-the-art on four diverse HAR benchmarks and, thus, highlights the effectiveness and generalizability of our framework (**Section 4.4**). Further, we share our insights through extensive quantitative and qualitative results as well as an ablation study to comprehensively demonstrate the contributions made by the architectural elements in our new HAR framework (**Section 4.5**).

2 RELATED WORK

Traditionally, the standard activity recognition pipeline for time-series sensory data involved sliding window segmentation, manual hand-crafted feature design, and subsequent activity classification with classical machine learning algorithms [6]. Studies along these line have extensively explored statistical [4, 35], basis transform [19], multi-level [55], and bio-mechanical features [48] coupled with the employment of shallow classifiers including support vector machines [7], decision trees [4], joint boosting [24], graphical models [38], and multi-layer perceptrons [34]. While this manually tuned procedure has successfully acquired satisfying results for relatively simple recognition tasks, its generalization performance is limited by heavy reliance on domain expert knowledge to engineer effective features.

Over the past years, the emerging paradigm of deep learning has presented unparalleled performance in various research areas including computer vision, natural language processing and speech recognition [25]. When applied to sensor-based HAR, deep learning allows for automated end-to-end feature extraction and thus, largely alleviates the need for laborious feature engineering procedures. Motivated by these, sensor-based human activity recognition has witnessed extensive and successful adoption of deep learning paradigms in diverse HAR application settings [18, 21, 29].

Pioneering studies in the field have explored Restricted Boltzmann Machines (RBMs) for automatic representation learning [2, 15, 32, 54]. Recently, deep architectures based on convolutional neural networks (CNNs) have been predominantly leveraged to automate feature extraction from sensor data streams while mutually enhancing activity classification performance [5, 37, 49, 51]. These studies typically employ a cascaded hierarchy of 1D convolution filters along the temporal dimensions to capture salient activity features at progressively more abstract resolutions. The acquired latent features are ultimately unified and mapped into activity class scores using a fully connected network. Taking a different approach, [50] presents an efficient dense labeling architecture based on fully convolutional networks that allows making activity predictions for every sample of a sliding window segment.

Another popular architecture design for HAR adopts deep recurrent neural networks (RNNs) that leverage memory cells to directly model temporal dependencies between subsequent sensor samples. In particular, [16] investigates forward and bi-directional long short-term memory (LSTM) networks, and [14] explores ensemble of diverse LSTM learners to exploit the sequential nature of sensor data. Combining these concepts, [30] proposes DeepConvLSTM by pairing convolutional and recurrent networks in order to model the temporal correlations at a more abstract representation level. In [28], the recurrent network of DeepConvLSTM is expanded with attention layers to model the relevant temporal context of sensor data. We refer readers to [44] for a curated list of recent HAR studies with deep neural networks.

Despite the great progress in the field, we can see that the unique opportunities we discussed in Section 1 for learning from multi-channel time-series data generated by body-worn sensors remain. Conventionally, the feature-maps generated by convolutional layers are trivially vectorized and fed to fully connected layers or recurrent networks to ultimately produce classification outcomes. However, such manipulation of the convolutional feature-maps fails to explicitly capture and encode the inter-channel interactions that can aid accurate recognition of activities. Moreover, regardless of the architectural designs, cross-entropy loss constitutes the common choice for supervised training of deep HAR models. Yet, this optimization objective alone does not cater for the need to achieve minimal intra-class compactness of feature representations [47] necessary to counter the significant intra-class variability of human activities.

In addition, while data augmentation has shown great potential for regularizing deep neural networks for general machine learning applications, only limited research efforts in HAR have focused on investigating systematic data augmentation techniques for wearable sensor data. In particular, [41] investigates hand-crafted augmentation approaches including jittering, scaling, cropping, permutation and axis rotations for monitoring of Parkinson's disease using wearable sensors with convolutional neural networks. In [27], data augmentation is applied to sensor data in order to specifically counter sampling-jitters resulting from software and hardware heterogeneity in diverse sensing devices. In another study, [11] explores a series of sequentially applied transformations—rotation, time-warp, scaling and jittering—in a semi-supervised transfer learning framework for complex human activity recognition. While the use of data augmentation in these studies consistently demonstrates improved generalisation to unseen data, the incorporated strategies are dataset-dependent and rely on the use of domain expert knowledge for effective and meaningful adoption; *e.g.* it is not straightforward and clear what degree of sensor data scaling is considered reasonable to apply without altering the semantic activity labels of the original data. This becomes even more problematic as wearable data are often captured over multitude of sensor channels with diverse magnitudes and innate properties; thus, complicating manual design of label-preserving sensor augmentations. This motivates *further investigation of data-agnostic augmentation approaches* for multi-channel times-series data in HAR that can be applied to effectively expand the training data captured by diverse sensing modalities without reliance on domain expert knowledge.

3 OUR PROPOSED HAR FRAMEWORK

The goal is to develop an end-to-end deep HAR model that directly consumes raw sensory data captured by wearables and seamlessly outputs precise activity classification decisions. In our framework, a network composed of 1D convolutional layers serves as the backbone feature extractor in order to automatically extract an initial feature representation for each sensory segment. Subsequently, we propose a two-staged refinement process to enrich the initial feature representations prior to classification that allows the model to *i*) effectively uncover and encode the underlying sensor channel interactions at each time-step, and *ii*) learn the relevant temporal context within the sequence of refined representations. Moreover, we encourage intra-class compactness of representations with center-loss while regularizing the network with mixup data augmentation during training. In what follows, we elaborate on the components of our framework[1], illustrated in Fig. 1.

3.1 1D Convolutional Backbone

Following the sliding window segmentation, the input to the network is a slice of the captured time-series data $x \in \mathbb{R}^{\mathrm{D}\times\mathrm{W}}$, where D denotes the number of sensor channels used for data acquisition and W represents the choice for the window duration. For automatic feature extraction, the input is then processed by a convolutional backbone operating along the temporal dimension. Given the 1D structure of the adopted filters, progressively more abstract temporal representations are learned from nearby samples without fusing features in-between different sensor channels. Ultimately, the backbone yields a feature representation $\mathbf{x} \in \mathbb{R}^{\mathrm{C}\times\mathrm{D}\times\mathrm{T}}$, where in each of the C feature maps, the sensor channel dimension D is preserved while the temporal resolution is down-sampled to T. Without loss of generality, in this paper we employ the convolutional layers of a state-of-the-art HAR model [30] as the backbone feature extractor; the input segment is successively processed by four layers, each utilizing 64 one-dimensional filters of size 5 along the temporal axis with ReLU non-linearities.

3.2 Cross-Channel Interaction Encoder (CIE)

Accurate realization of fine-grained human actions using wearables is often associated with utilizing multitude of on-body sensing devices that capture activity data across multiple channels. Measurements captured by different sensor channels provide different views of the same undergoing activity and are thus, inherently binded together in an unobservable latent space. Accordingly, we seek to design an end-to-end trainable module that takes as input the initial convolutional feature-maps at each time-step, learns the interactions between any two sensor channels within the feature-maps, and leverages this overlooked source of information to enrich the sensory feature representations for HAR.

Motivated by the emerging successful applications of self-attention [43, 45, 53] in capturing global dependencies by computing relations at any two positions of the input, here we design a *Cross-Channel Interaction Encoder (CIE)* that adopts self-attention mechanism to effectively process the initial feature representations and uncover the latent channel interactions. To this end, we first compute the normalized correlations across all pairs of sensor channel features $\mathbf{x}_t^d$ and $\mathbf{x}_t^{d'}$ using the embedded Gaussian function at each time-step t,

$$\mathbf{a}_t^{d,d'} = \frac{\exp\left(f(\mathbf{x}_t^d)^\intercal g(\mathbf{x}_t^{d'})\right)}{\sum_{d'=1}^{\mathrm{D}} \exp\left(f(\mathbf{x}_t^d)^\intercal g(\mathbf{x}_t^{d'})\right)}, \tag{1}$$

where $\mathbf{a}_t^{d,d'}$ indicates the attendance of our model to the features of sensor channel d' when refining representations for sensor channel d. Subsequently, the extracted correlations are leveraged in order to compute the response for

[1] Notably, in our title, *Attend* refers to the self-attention module and the attentional GRU encoder modules; *Discriminate* refers to the impact of the center-loss criterion and the mixup method to learn discriminative activity representations that generalize well to unseen data in our proposed framework.

the d^{th} sensor channel features $\mathbf{x}_t^d \in \mathbb{R}^{\text{C}}$ and generate the corresponding self-attention feature-maps $\mathbf{o}_t^d$ at each time-step

$$\mathbf{o}_t^d = v\Big(\sum_{d'=1}^{\text{D}} \mathbf{a}_t^{d,d'} h(\mathbf{x}_t^{d'})\Big). \tag{2}$$

Technically, the self-attention in the CIE module functions as a non-local operation which computes the response for sensor channel d at each time-step by attending to all present sensor channels' representations in the feature-maps at the same time-step. In the above, f, g, h, and v all represent linear embeddings with learnable weight matrices ($\in \mathbb{R}^{\text{C}\times\text{C}}$) that project feature representations into new embedding spaces where computations are carried out. Having obtained the self-attention feature-maps, the initial feature-maps are then added back via a residual link (indicated by $\bigoplus$ in Fig. 1) to encode the interactions and generate the refined feature representations $\mathbf{r}_t$,

$$\mathbf{r}_t^d = \mathbf{o}_t^d + \mathbf{x}_t^d. \tag{3}$$

With the residual connection in place, the model can flexibly decide to use or discard the correlation information. During training, the HAR model leverages the CIE module to capture the interactions between different sensor channels. The discovered correlations are encoded inside the self-attention weights and leveraged at inference time to help support the model's predictions.

3.3 Attentional GRU Encoder (AGE)

As a result of employing the CIE module, the feature-maps generated at each time-step are now contextualized with the underlying cross-channel interactions. As shown in Fig. 1, we vectorize these representations at each time-step to obtain a sequence of refined feature vectors $(\mathbf{r}_t \in \mathbb{R}^{\text{CD}})_{t=1}^{\text{T}}$ ready for sequence modeling. Given that not all time-steps equally contribute in recognition of the undergoing activities, it is crucial to learn the relevance of each feature vector in the sequence when representing activity categories. In this regard, applying attention layers to model the relevant temporal context of activities has proved beneficial in recent HAR studies [28]. Adopting a similar approach, we utilize a 2-layer *attentional GRU Encoder (AGE)* to process the sequence of refined representations and learn soft attention weights for the generated hidden states $(\mathbf{h}_t)_{t=1}^{\text{T}}$. In the absence of attention mechanism in the temporal domain, classification decision would only be based on the last hidden state achieved after observing the entire sequence. By contrast, empowering the GRU encoder with attention alleviates the burden on the last hidden state and instead, allows learning a holistic summary $\mathbf{z}$ that takes into account the relative importance of the time-steps

$$\mathbf{z} = \sum_t \beta_t \mathbf{h}_t, \tag{4}$$

where β_t denotes the computed attention weight for time-step t. Technically, attention values are obtained by first mapping each hidden state into a single score with a linear layer and then normalizing these scores across the time-steps with a softmax function.

3.4 Center Loss Augmented Objective

Intra-class variability and inter-class similarity are two fundamental challenges of HAR with wearables. The former phenomena occurs since different individuals may execute the same activity differently while the latter challenge arises when different classes of activities reflect very similar sensor patterns. To counter these challenges, the training objective should encourage the model to learn discriminative activity representations; *i.e.*, representations that exhibit large inter-class differences as well as minimized intra-class variations.

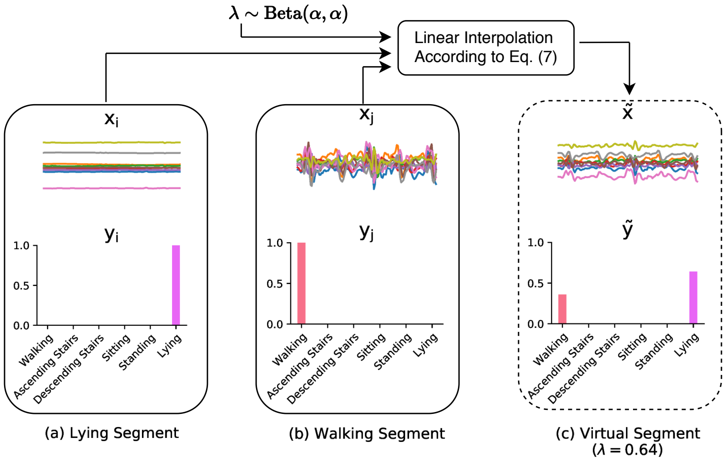


Fig. 2. We leverage mixup data augmentation technique to generate virtual sequences during training. We interpolate in-between samples. Here, we visualize (a) a sequence of sensor data from the training split corresponding to the `lying` activity and its one-hot encoded label representation, (b) a training sensor data segment corresponding to the `walking` activity, and (c) a *virtual* or generated sequence and its target label according to Eq. 7 with a drawn λ value of 0.64 (sampled from a Beta distribution). The visualized data corresponds to a subset of sensor channels in the PAMAP2 dataset [36].

Existing HAR architectures solely rely on the supervision signal provided by the cross-entropy loss during their training phase. While optimizing for this criteria directs the training process towards yielding inter-class separable activity features, it does not explicitly encourage learning intra-class compact representations. To boost the discriminative power of the deep activity features within the learned latent space, we propose to incorporate center-loss [47] for training our HAR model. The auxiliary supervision signal provided by center-loss penalizes the distances between activity representations and their corresponding class centers and thus, reduces intra-class feature variations. Formally, center-loss is defined as

$$\mathcal{L}_c = \frac{1}{2}\sum_{i=1} \|\mathbf{z}_i - \mathbf{c}_{y_i}\|_2^2, \tag{5}$$

where $\mathbf{z}_i \in \mathbb{R}^z$ denotes the deep representation for sensory segment x_i, and $\mathbf{c}_{y_i} \in \mathbb{R}^z$ denotes the y_ith activity class center computed by averaging the features of the corresponding class. We enforce this criteria on the activity representations obtained from the penultimate layer of our network to effectively pull the deep features towards their class centers.

In each iteration of the training process, we leverage the joint supervision of cross-entropy loss together with center-loss to simultaneously update the network parameters and the class centers $\mathbf{c}_y$ in an end-to-end manner. Hence, the aggregated optimization objective is formulated as

$$\Theta^* = \arg\min_{\Theta} \mathcal{L} + \gamma \mathcal{L}_{\mathrm{c}}, \tag{6}$$

where $\mathcal{L}$ represents the cross-entropy loss, γ is the balancing coefficient between the two loss functions, and Θ denotes the collection of all trainable parameters.

3.5 Mixup Data Augmentation for HAR

Due to the laborious task of collecting annotated datasets from wearables, current HAR benchmarks are characterized by their limited sizes. Therefore, introducing new modules and increasing the network parameters without employing effective regularization techniques, makes the model prone to overfitting and endangers its generalization. In this regard, while extending the training data with augmented samples achieved by *e.g.* slight rotations, scaling, and cropping has consistently led to improved generalization performance for computer vision applications, these methods are not directly applicable to multi-channel time-series data captured by wearables.

In this paper, we explore the effectiveness of a recently proposed data-agnostic augmentation strategy, namely *mixup* [52], for time-series data in order to regularize our deep HAR model. This approach has demonstrated its potential in significantly improving the generalization of deep neural networks by encouraging simple linear behavior in-between training data. In addition, unlike existing augmentation approaches that are dataset dependent and thus require domain expert knowledge for effective adoption, mixup strategy is domain independent and simple to apply. In essence, mixup yields augmented virtual example $(\tilde{x}, \tilde{y})$ through linear interpolation of training example pairs (x_i, y_i) and (x_j, y_j),

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1 - \lambda) x_j \\ \tilde{y} &= \lambda y_i + (1 - \lambda) y_j, \end{aligned} \tag{7}$$

where λ sampled from a Beta(α, α) distribution is the mixing-ratio and α is the mixup hyper-parameter controlling the strength of the interpolation. Notably, mixup augmentation allows efficient generation of virtual examples on-the-fly by randomly picking pairs from the same minibatch in each iteration. In this work, we adopt mixup strategy to augment the time-series segments in each mini-batch and train the model end-to-end with the generated samples. We visually explain the augmentation process with an example in Fig. 2, where a pair of randomly drawn training data sequences are linearly interpolated to yield a novel virtual sequence.

4 EXPERIMENTS AND RESULTS

4.1 Benchmark Datasets

To validate our framework and provide empirical evidence of its generalizability, we employ four HAR benchmarks exhibiting great diversity in terms of the sensing modalities used and the activities to be recognized. We provide a brief description of the datasets in what follows.

Opportunity Dataset [8]. This dataset is captured by multiple body-worn sensors. Four participants wearing the sensors were instructed to carry out naturalistic kitchen routines. The data is recorded at a frequency of 30 Hz and is annotated with 17 sporadic gestures as well as a `Null` class. Following [16], the 79 sensor channels not indicating packet-loss are used. For hold-out evaluation, we use runs 4 and 5 from subjects 2 and 3 as the holdout test-set, run 2 from participant 1 as the validation-set, and the remaining data as the training-set.

PAMAP2 Dataset [36]. This dataset is aimed at recognition 12 diverse activities of daily life. Data was recorded over 52 channels with annotations covering prolonged household and sportive actions. Replicating [16] for hold-out evaluation, we use runs 1 and 2 from subject 6 as the holdout test-set, runs 1 and 2 from subject 5 as the validation-set, and the remaining data for training.

Skoda Dataset [39]. The dataset covers the problem of recognizing 10 manipulating gestures of assembly-line workers in a manufacturing scenario. Following [14] for hold-out evaluation, we use the data recorded over 60 sensor channels collected from the right arm, utilize the first 80% of each class for the training-set, the following 10% for validation and the remainder as the test-set.

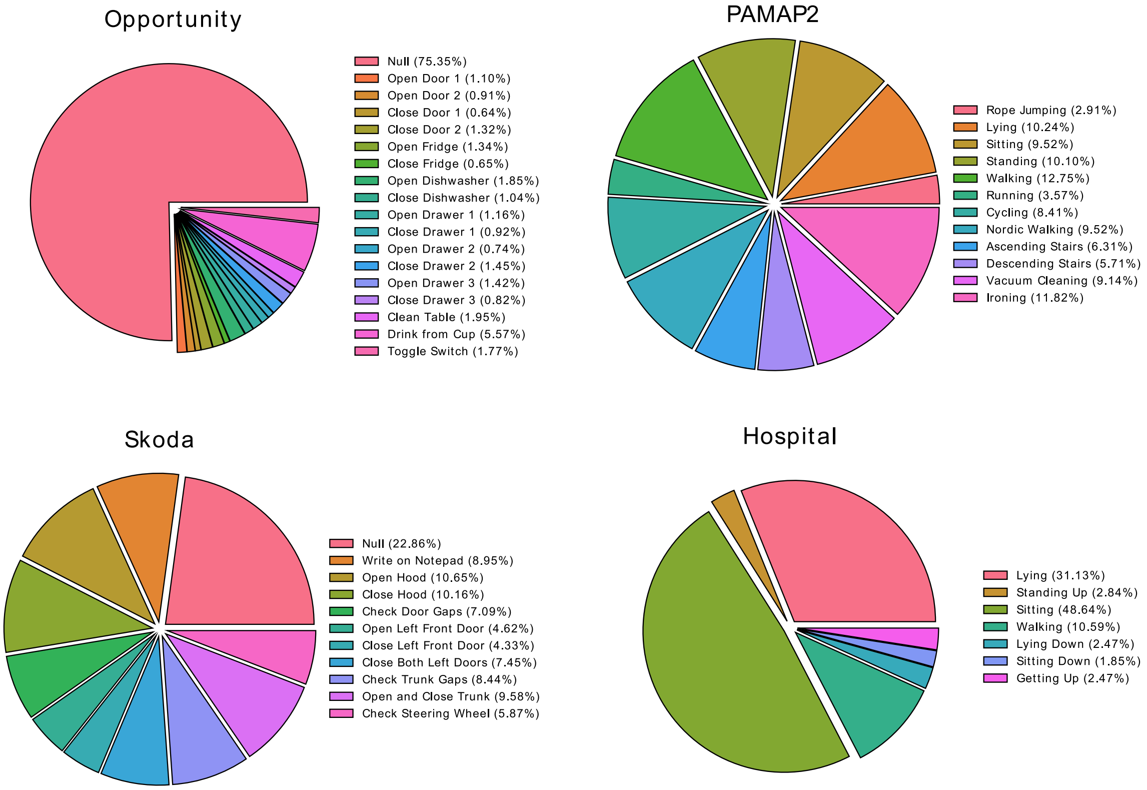


Fig. 3. Benchmark HAR datasets investigated in this paper. We illustrate the activity categories covered and their corresponding distributions within each dataset.

Hospital Dataset [50]. This dataset is collected from 12 hospitalized older patients wearing an inertial sensor over their garment while performing 7 categories of activities. All the data is recorded at 10 Hz. Following [50] for hold-out evaluation, data from the first eight subjects are used for training, the following three for testing, and the remaining for the validation set.

We summarize the list of covered activity categories and their corresponding distributions within each dataset in Fig. 3. As illustrated, while the prevalence of each activity category for PAMAP2 and Skoda datasets is quite balanced, we observe a significant distribution imbalance for Opportunity and Hospital datasets.

4.2 Unified Evaluation Protocol

To ensure a fair comparison, we directly adopt the hold-out evaluation protocol and metrics with standard training and testing splits used in the recent literature [14, 16, 17, 28]. Where possible, sensor data are down-sampled to 33 Hz to achieve a consistent temporal resolution with the Opportunity dataset. Each sensor channel is normalized to zero mean and unit variance using the training data statistics. Following [14, 28], the training data is partitioned into segments using a sliding window of 24 samples (*i.e.*, W=24) with 50% overlap between adjacent windows. For a realistic setup, sample-wise evaluation is adopted to compare the performance on the hold-out test-set; thus, a

Table 1. A summary of hyper-parameter values selected per dataset. All other hyper-parameters were kept constant across all datasets.

Hyper-parameter	Opportunity	PAMAP2	Skoda	Hospital
Dropout ratio p_{feat}	0.5	0.9	0.5	0.5
Dropout ratio p_{cls}	0.5	0.5	0.0	0.5
Weighting coefficient γ	3×10^{-4}	3×10^{-3}	3×10^{-1}	3×10^{-1}

prediction is made for every sample of the test sequence as opposed to every segment. Given the imbalanced class distributions in the datasets (see Figure 3), as in [14, 16, 28], the class-average F-score

$$\text{F-score}_m = \frac{2}{C} \sum_{c=1}^{C} \frac{\text{prec}_c \times \text{recall}_c}{\text{prec}_c + \text{recall}_c} \tag{8}$$

is used as the evaluation metric to reflect the ability of the HAR model to recognise every activity category regardless of its prevalence in the collected data. Here, C denotes the number of activity classes, and prec_c and recall_c respectively represent the precision and recall terms computed for activity class c.

4.3 Implementation Details

We implement our experiments using PyTorch [31]. Our network is trained end-to-end for 300 epochs by back-propagating the gradients of the loss function based on mini-batches of size 256 and in accordance with the Adam [22] update rule. The learning rate is set to 10^{-3} and decayed every 10 epochs by a factor of 0.9. For *mixup* augmentation, we fix α=0.8. All these hyper-parameters are kept constant across all datasets. For each dataset, we choose a dropout probability $p \in \{0, 0.25, 0.5, 0.75, 0.9\}$ for the refined feature-maps (p_{feat}) and the feature vectors fed to the classifier (p_{cls}), and select the center-loss weighting coefficient $\gamma \in 3\times\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, as summarized in Table 1.

4.4 Comparisons with the State-of-the-Art

4.4.1 Classification Measure. We compare our proposed framework against state-of-the-art HAR models on four standard benchmarks in Table 2. As elucidated in Section 4.2, every baseline generates sample-wise predictions on the entire holdout test sequence and the performance is judged based on the acquired class-averaged f1-score

Table 2. A comparison of sample-wise activity recognition performance based on class-averaged f1-scores on the holdout test sequences. The baseline results are quoted from [14, 28], except for (*) where the published code is used with our evaluation protocol.

HAR Model	Opportunity	PAMAP2	Skoda	Hospital*
LSTM Learner Baseline [14]	65.9	75.6	90.4	62.7
DeepConvLSTM [30]	67.2	74.8	91.2	62.8
b-LSTM-S [16]	68.4	83.8	92.1	63.6
Dense Labeling [50]*	62.4	85.4	91.6	62.9
Att. Model [28]	70.7	87.5	91.3	64.1
Ours	**74.6**	**90.8**	**92.8**	**66.6**
(Improvement over Runner-up)	**(5.52%)**	**(3.77%)**	**(0.76%)**	**(3.9%)**

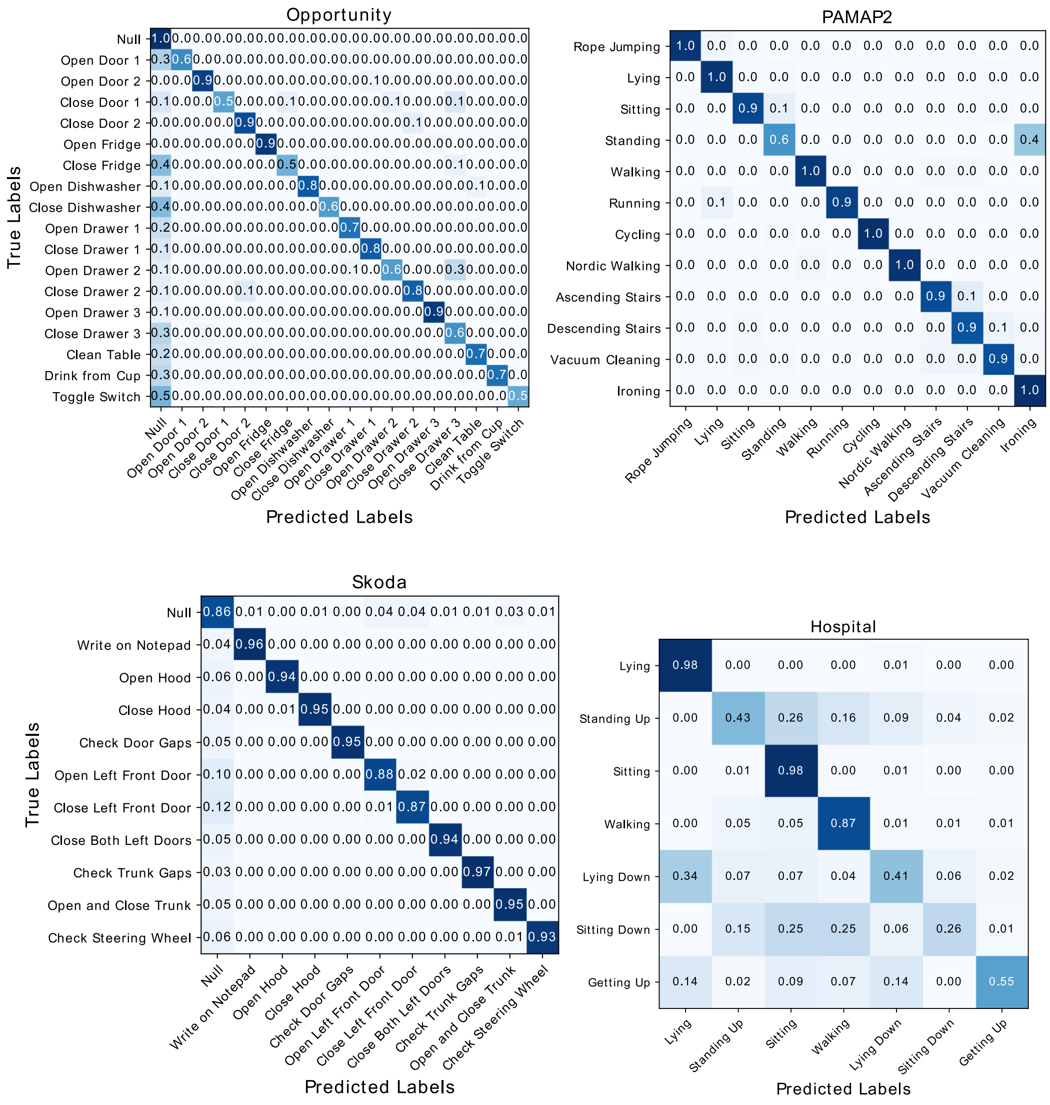


Fig. 4. The confusion matrices highlighting the class-specific recognition performance for the testing splits of Opportunity, PAMAP2, Skoda, and Hospital HAR datasets. The vertical axis represents the ground-truth activity categories and the horizontal axis denotes the predicted activities.

(F_m). The baseline results are directly quoted from [14, 28], except where indicated by (*), where the published code is used with the standard evaluation protocol.

In Table 2, we can see that the elements we introduced into our framework consistently yield significant recognition improvements over the state-of-the-art models. Interestingly, we observe the highest performance gain of 5.52% on the Opportunity dataset characterized by *i*) the largest number of incorporated sensor channels;

Table 3. A comparison of segment-wise activity recognition performance based on class-averaged f1-scores with cross-fold evaluation.

HAR Model	Opportunity	Opportunity (*w/o* Null)	PAMAP2	Skoda	Skoda (*w/o* Null)	Hospital
LSTM Learner Baseline [14]	75.6 ± 0.7	70.2 ± 0.7	97.8 ± 0.1	90.9 ± 0.6	82.6 ± 0.6	71.5 ± 2.1
DeepConvLSTM [30]	73.0 ± 0.8	67.7 ± 0.8	97.9 ± 0.1	90.8 ± 0.2	83.2 ± 0.2	72.1 ± 2.4
b-LSTM-S [16]	77.2 ± 1.1	71.8 ± 1.1	97.9 ± 0.1	90.9 ± 0.2	83.5 ± 0.2	72.4 ± 1.4
Dense Labeling [50]	78.5 ± 0.4	73.1 ± 0.4	98.4 ± 0.1	92.1 ± 0.3	84.1 ± 0.3	70.3 ± 0.7
Att. Model [28]	78.1 ± 0.2	72.3 ± 0.6	98.4 ± 0.1	90.4 ± 0.5	82.8 ± 0.4	72.5 ± 1.7
Ours	**81.1 ± 0.2**	**75.7 ± 0.1**	**98.7 ± 0.1**	**93.2 ± 0.4**	**85.3 ± 0.3**	**73.1 ± 1.9**

ii) the greatest diversity of the actions to recognize; and *iii*) the highest ratio of class imbalance. Our results highlight the significant contribution made by the integrated components in dealing with challenging activity recognition tasks. Notably, our framework still achieves a moderate performance improvement on the *performance saturated* [28] Skoda dataset.

For further insights, we summarize the class-specific recognition results from our model by presenting confusion matrices for the four recognition tasks in Fig. 4. We can see that for Opportunity and Skoda datasets with the inclusion of a Null class in the annotations, most of the confusions occur in distinguishing between the ambiguous Null class and the activities of interest. This can be understood since the Null class represents an infinite number of irrelevant activity data for the HAR problem in hand; thus, explicitly modeling this unknown space is a difficult problem.

For completeness, we perform additional extensive cross-fold evaluations across all benchmark datasets in Table 3 to complement our hold-out evaluations presented in Table 2. Following [20], we employ fully non-overlapping windows to generate sensor segments with no temporal overlaps. This is to guarantee that the segment contents do not simultaneously appear both in training and testing splits and prevent data leakage from the training set to the test sets. Subsequently, 3-fold stratified cross-validation is adopted on the datasets to produce the training and testing splits while preserving activity class distributions across all folds. Each constructed fold is in turn utilized once for testing while the remaining folds constitute the training data. The resulting class-average F-score is reported in Table 3 for the benchmark datasets and the corresponding HAR models. In the case of the Opportunity and Skoda datasets, we report the recognition performance both including and ignoring the Null class (*w/o Null* columns in Table 3) during inference. Inclusion of the Null class may result in an overestimation of the recognition performance due to its large prevalence and thus, providing both results gives better insights into the nature of the errors made by the models [30].

Consistent with the observations of hold-out evaluations, our proposed framework presents superior performance in identifying human activity classes from raw sensor data across all benchmarks as compared with the baseline HAR models in Table 3. Interestingly, a comparison of results between the two evaluation methods—*i.e.,* hold-out evaluation in Table 2 and cross-fold evaluation in Table 3—*indicates significantly higher recognition performance for the latter.* This is mainly due to the fact that with cross-fold evaluation, the activity data captured by a subject may appear both in the training and testing folds (despite not having any temporal overlap of the sensor data), thus leading to better generalization performance of the trained models on the testing sets.

In addition, whilst we employed the window size determined in previous studies aiming to improve benchmark performance—see Section 4.2, we also compare our framework with the baselines using a larger window size in Appendix A, for completeness. The results demonstrate that the proposed framework still achieves better results compared to the baseline models.

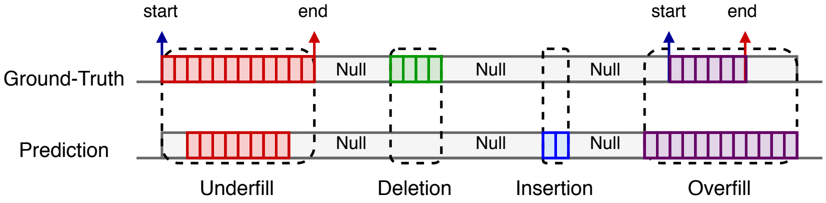


Fig. 5. We illustrate different categories of misalignment measures investigated in this paper. Here, presence of different activity classes is represented with distinct colors and the Null class is denoted with gray. The sequence of continuous sample predictions is compared against the ground-truth sample labels to compute the underfill, overfill, insertion and deletion misalignment measures.

Table 4. Misalignment measures comparison. (*) denotes the best performing state-of-the-art recognition model [28] according to Table 2.

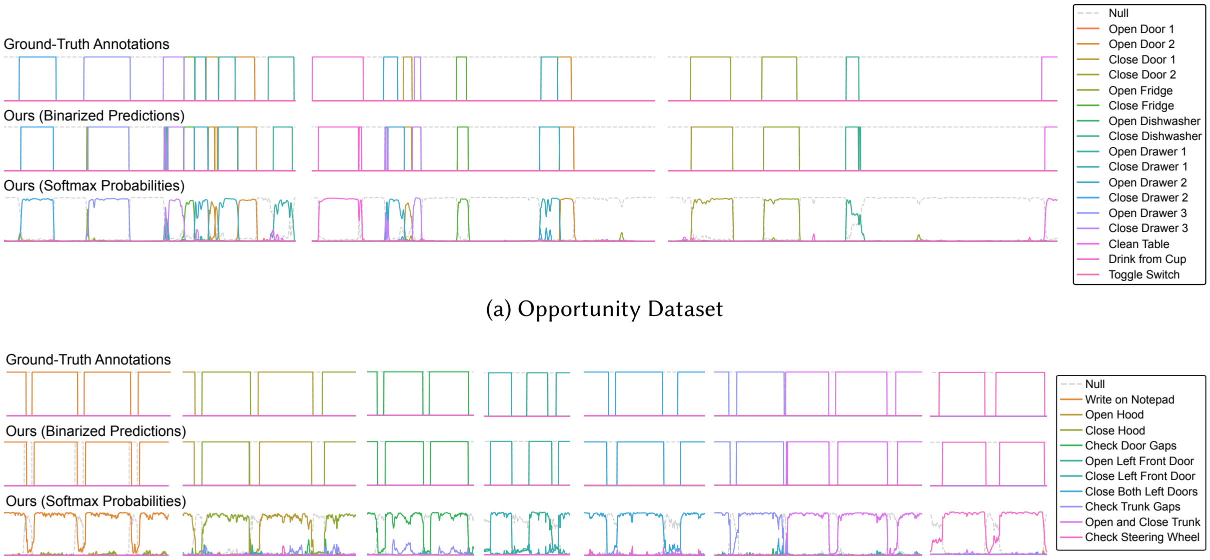
	Opportunity		Skoda	
Alignment Measures	Ours	SoA*	Ours	SoA*
Deletion (↓)	**0.62**	0.69	**0.04**	**0.04**
Insertion (↓)	**2.87**	3.34	**2.01**	3.34
Underfill/Overfill (↓)	**3.71**	4.15	5.33	**5.17**
True Positives (↑)	**92.8**	91.82	**92.62**	91.45

4.4.2 Misalignment Measure. In addition to the reported classification metrics, we further report on the explicitly designed misalignment measures of *overfill/underfill, insertion,* and *deletion* [46] and provide comparisons with the state-of-the-art HAR model [28] in Table 4. These metrics characterize *continuous* activity recognition performance and provide finer details on temporal prediction misalignment with respect to ground truth as illustrated in Fig. 5. Specifically:

- *Overfill* and *Underfill* indicate errors when the predicted start or end time of an activity are earlier or later than the ground-truth timings.
- *Insertion* errors refer to incorrectly predicting an activity when there is `Null` activity.
- *Deletion* represents wrongly predicting `Null` class when an activity exists.

Since some measures require the existence of `Null` class by definition, we report results on Opportunity and Skoda datasets. The quantitative results in Table 4 indicate the improved capability of our model to predict a continuous sequence of activity labels that more accurately aligns with ground-truth timings and better recognizes existence or absence of activities of interest.

Further, we visualize fragments of sensor recordings from these datasets in Fig. 6 for qualitative assessment. The Skoda dataset includes repetitive execution of quality check gestures while the Opportunity dataset is characterized by short duration and sporadic activities. We present the ground-truth annotations (top rows), our model's softmax output probabilities (last rows) and the binarized sequence of predictions (middle rows) obtained after applying argmax operation on the soft scores for each time-step. At every time-step, we color-code and plot the output class probabilities for each activity category, where we observe a strong correspondence between the ground-truth annotations, activity duration and the predicted activity scores.



(a) Opportunity Dataset

(b) Skoda Dataset

Fig. 6. A visualization of our networks' predictions on the holdout test fragments. Our proposed HAR model accurately localizes and classifies short duration gestures embodied in sequences of sensor signals captured by wearables. We visualize the model's predictions against the ground-truth annotations for sequence fragments of Opportunity and Skoda datasets which include a Null class label representing activities of non-interest.

4.4.3 Efficiency Analysis. In Fig. 7, we present a computational complexity comparison among activity recognition models explored in this study. In particular, we illustrate the number of trainable network parameters associated with the HAR baselines for each activity recognition benchmark dataset in Fig. 7-a. Clearly, FCN [50] demonstrates significantly lower number of parameters as compared with the other models due to its fully convolutional structure and abandoning fully connected layers entirely. This is beneficial for realizing activity recognition on edge devices where storage constraints may be a concern. As for the baseline LSTM learner [14], we observe a similar number of parameters across all benchmark datasets. This is due to the fact that the number of parameters within the LSTM networks are heavily influenced by the number of adopted hidden units. This also holds for the b-LSTM-S [16] architecture, however employing roughly twice the number of learnable parameters due to the bi-directional connections. The remaining HAR frameworks—DeepConvLSTM [30], Att. Model [28], and *Ours*—employ identical backbone feature extractors and mainly differ in terms of the recurrent networks and the attentional components. *Notably, by replacing the LSTM recurrent network with the GRU modules, our proposed HAR framework reduces the number of trainable parameters within the architecture.*

Additionally, in Fig. 7-b, we analyze the inference time required by each HAR model to process a single sensory window of 24 samples for all benchmark datasets. To simulate a real-time deployment scenario, the holdout test sets of the benchmark datasets are segmented and sequentially processed—*i.e.*, with a batch size of one—by the HAR baselines and the corresponding total processing time is divided by the total of number of processed segments to generate the results in Fig. 7-b. While all frameworks are suitable for real-time predictions—*i.e.*, consuming a processing time of approximately 0.8-1.9 milliseconds—the b-LSTM-S network demonstrates the highest inference time.

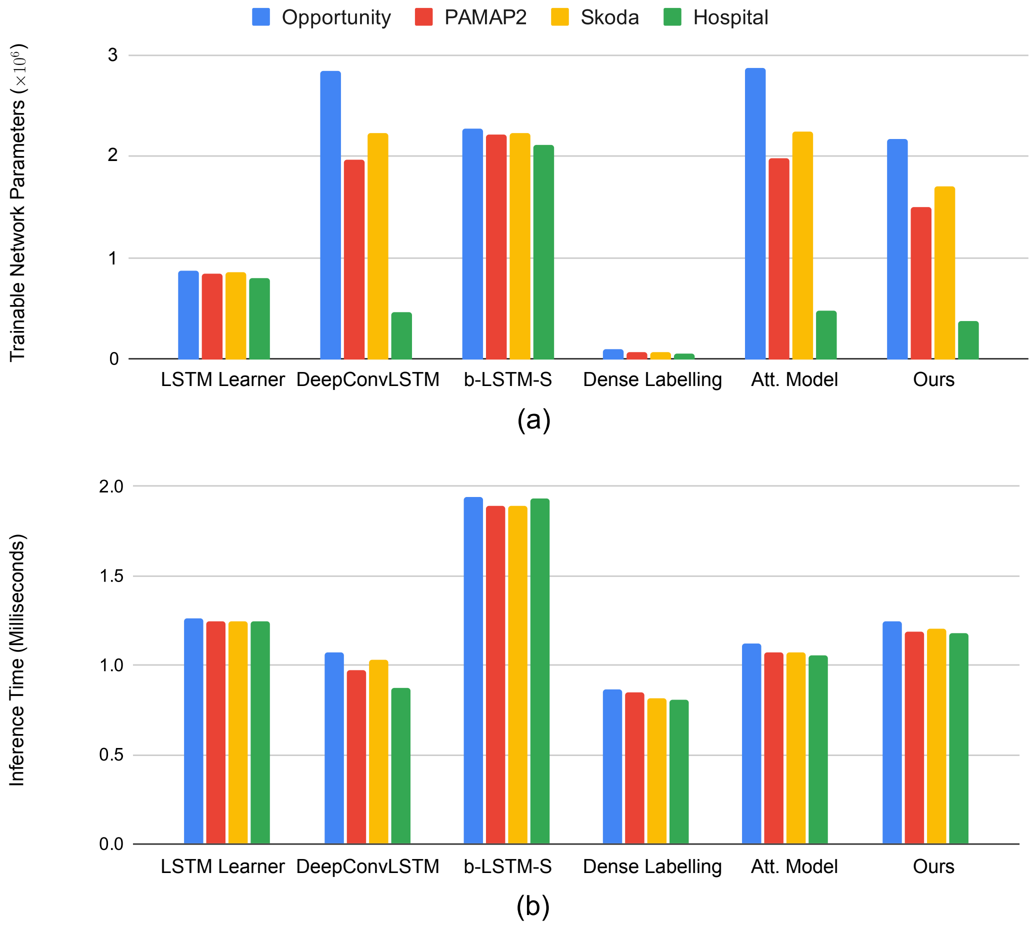


Fig. 7. A computational complexity comparison among activity recognition models explored in this study: (a) the number of trainable network parameters, and (b) the inference time required for processing a single sensory segment associated with the HAR baselines for each activity recognition benchmark dataset.

4.5 Ablation Studies and Insights

Given that our proposed HAR model integrates several key ideas into a single framework, we conduct an ablation study on the Opportunity dataset to understand the contribution made by the various components for the human activity recognition task in Table 5. For each ablated experiment, we remove specific modules of our framework and as a reference we include DeepConvLSTM—the backbone of our network as illustrated in Fig. 1. Unsurprisingly, removing any component handicaps the HAR model and reduces performance (to 67.2%—see DeepConvLSTM baseline performance) while incorporating all components together yields the highest performing HAR model (74.6%—see mixup+CenterLoss+CIE+AGE).

Notably, the effectiveness of mixup augmentation in regularizing models learnt from time-series wearable HAR data is demonstrated by the significant relative improvement of 5.2% over the DeepConvLSTM Baseline compared to employing mixup alone (an improvement from 67.2% to 70.7%). The virtual multi-channel time-series data attained through in-between sample linear interpolations expand the training data and effectively

Table 5. We investigate the contribution of integrated modules by conducting an ablation study on the Opportunity dataset.

HAR Model	F_m
DeepConvLSTM Baseline	67.2
Ours (CenterLoss + CIE + AGE)	70.2
Ours (mixup)	70.7
Ours (mixup + CenterLoss)	72.2
Ours (mixup + AGE)	71.7
Ours (mixup + CIE)	73.0
Ours (mixup + CenterLoss + AGE)	72.3
Ours (mixup + CenterLoss + CIE)	73.2
Ours (mixup + CIE + AGE)	74.0
Ours (mixup + CenterLoss + CIE + AGE)	**74.6**

Table 6. A comparison of activity recognition performance on Opportunity dataset based on the class-averaged f1-scores achieved while employing different data augmentation strategies.

	No Augmentation	Jittering	Scaling	Magnitude Warping	Mixup
F_m	70.2	69.4	70.4	70.3	74.6

improve the generalization of learned activity features to unseen test sequences. This is particularly important when incorporating HAR architecture designs with increased number of trainable parameters to counter the phenomenon of overfitting.

We also investigated conventional hand-crafted augmentation strategies adopted for wearable HAR. We adopt the recent methods studied in [41] including jittering, scaling and magnitude warping using the officially provided

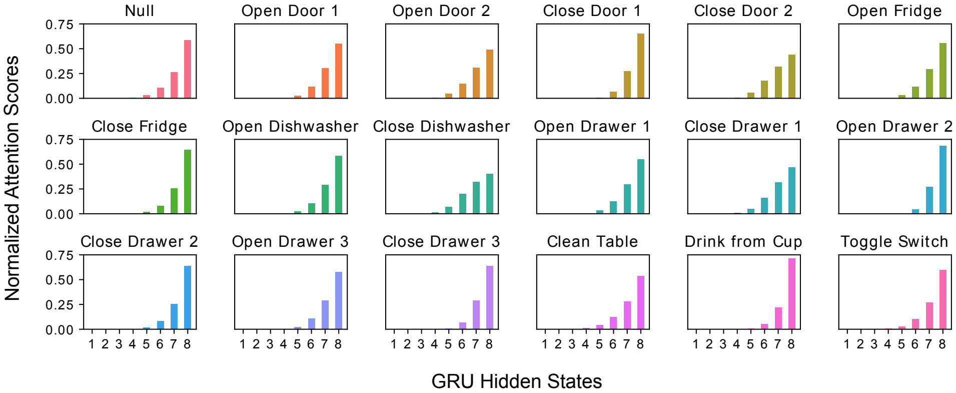


Fig. 8. A visualization of discovered temporal attentions by our AGE module. The vertical axis represents the normalized attention scores and the horizontal axis denotes hidden states $(\mathbf{h}_t)_{t=1}^{T=8}$ of our GRU encoder. The GRU becomes progressively more informed about the activity and thus, places higher attention on the few last hidden states with the last state dominating the attention weights.

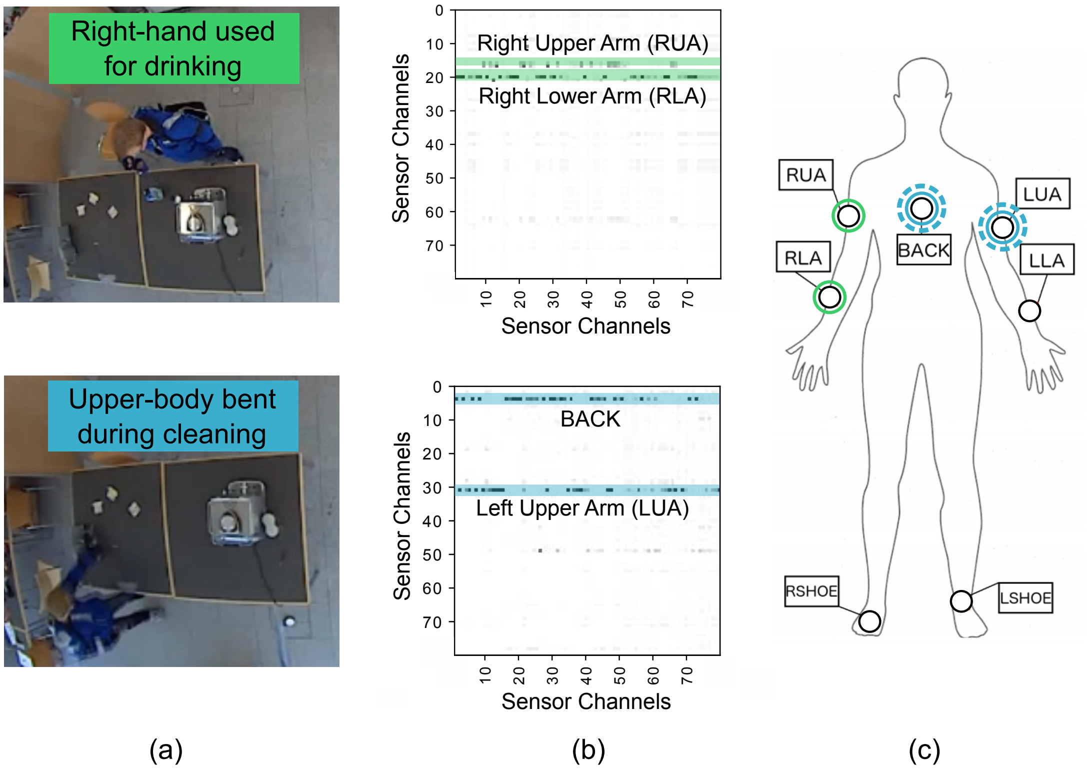


Fig. 9. A visualization of learned self-attention correlations by our CIE module. (a) Subject engaged in two activities; (b) discovered cross-channel correlations by our model selected for *Right-hand used for* `drinking from cup` (dark shaded marks along the rows highlighted in green) and *Upper-body bent during* `cleaning table` (dark shaded marks along the rows highlighted in blue) as shown in video snapshots recorded during the data collection process and shown in (a); and (c) highly attended sensor locations for each activity—color-coded to match green and blue highligts in (b)—in the Opportunity dataset.

implementations in Table 6. Here, jittering simulates additive sensor noise, scaling changes the magnitude of segment data by multiplying by a random scalar, and magnitude warping convolves the segment with a random sinusoidal curve using arbitrary amplitude, frequency and phase.

According to the results, mixup clearly outperforms existing augmentation methods by a large margin. Moreover, in line with the observations made in [41], we see that data augmentation techniques may adversely affect recognition performance if not carefully tuned and applied, as is the case here for the jittering approach. We argue that depending on the target task—*i.e.,* the activities to be recognized, sensor channel characteristics, intra-class variations and inter-class similarities—hand-crafted augmentation methods demand domain expert knowledge for effective adoption. In particular, for the Opportunity dataset with 79 sensor channels and 18 fine-grained activity classes, it is not trivial and straight-forward to design channel specific augmentations.

Most importantly, as opposed to the conventional hand-crafted augmentation strategies, mixup augmentation takes into account both *data* and *label* information (see Eq. 7) when generating novel samples. This approach largely alleviates the concerns regarding the label-preservation of transformations, and allows simple adoption for diverse activity recognition problem scenarios. This is substantiated in Table 6 comparisons of recognition performance on the Opportunity dataset, where mixup provides improved results over no augmentation whilst hand-crafted augmentation strategies such as jittering negatively impacts performance and other methods provide approximately similar results to those achieved with no augmentation.

As hypothesized, encouraging minimal intra-class variability of representations with *center-loss* consistently improves the recognition performance for activities (mixup+CenterLoss). In addition, while both *CIE* and *AGE* modules allow learning better representations of activities reflected by the enhanced metrics for mixup+CIE+AGE compared to mixup (4.7% relative improvement), we observe a larger performance gain when incorporating *CIE* module as compared with *AGE*; the former encodes the cross-channel sensor interactions with self-attention while the latter learns the relevance of time-steps with temporal attention. Presumably this is due to the fact that within the Opportunity challenge setup, the sequence of representations fed to the GRU is quite short in length (*i.e.*, T=8) and therefore, the last hidden state alone captures most of the information relevant to the activity. In order to verify this, we extract the learned attention scores β_t corresponding to the hidden states $(\mathbf{h}_t)_{t=1}^{T=8}$ of our GRU encoder and present an illustration for every activity category of Opportunity dataset in Fig. 8. In line with the observations made in [28], the recurrent neural network progressively becomes more informed about the activity and thus, proportionally places higher attention on the few last hidden states with the last state dominating the attendance.

On the other hand, we observed exploiting latent channel interactions to significantly improve activity representations as highlighted in ablation results in Table 5. To visually explain the learned self-attention correlations from our proposed cross-channel encoder, we graph two segments associated with activity classes of `drinking from cup` and `cleaning table`. The CIE module consumes an input sequence and generates a normalized score matrix of size D×D, corresponding to the attention between each pair of D=79 channels. In Fig. 9, we present the normalized self-attention scores, $\mathbf{a} \in \mathbb{R}^{79\times79}$ (attained from softmax operation) in Eq. 1, where each column in the attention matrix indicates the extent that a particular sensor channel attends to available sensor channels.

We observe a clear and meaningful focus on a subset of channels vital to the recognition of activities indicated by dark rows in the matrices. For example, we notice high attendance: *i*) to the inertial measurement units (IMUs) on the right arm when *right hand is being used for* `drinking from cup`; and *ii*) to the IMUs placed on the back and left-upper arm when *upper-body is bent during* `cleaning table`. Thus, the explicit modeling of sensor channel interactions not only leads to improved recognition performance as substantiated by our ablation study in Table 5, but also facilitates visual explanation through interpretable scores.

5 CONCLUSIONS

Over the past years, human activity recognition (HAR) using wearables has created increasingly new opportunities for healthcare applications. In this paper, we present a new HAR framework and demonstrate its effectiveness through significant performance improvements achieved over state-of-the-art and its generalizability by evaluations across four diverse benchmarks. In particular, we: *i*) enriched activity representations by exploiting latent correlations between sensor channels; *ii*) incorporated center-loss to alleviate dealing with intra-class variations of activities; and *iii*) augmented multi-channel time-series data with mixup for better generalization. We believe that our work will provide new opportunities to further research in HAR using wearables.

REFERENCES

[1] Alireza Abedin, S. Hamid Rezatofighi, Qinfeng Shi, and Damith C. Ranasinghe. 2019. SparseSense: Human Activity Recognition from Highly Sparse Sensor Data-streams Using Set-based Neural Networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 5780–5786. https://doi.org/10.24963/ijcai.2019/801

[2] Mohammad Abu Alsheikh, Ahmed Selim, Dusit Niyato, Linda Doyle, Shaowei Lin, and Hwee-Pink Tan. 2016. Deep activity recognition models with triaxial accelerometers. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

[3] Akin Avci, Stephan Bosch, Mihai Marin-Perianu, Raluca Marin-Perianu, and Paul Havinga. 2010. Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In *23th International conference on architecture of computing systems 2010*. VDE, 1–10.

[4] Ling Bao and Stephen S. Intille. 2004. Activity recognition from user-annotated acceleration data. In *Proceedings of the 2nd International Conference on Pervasive Computing*, Alois Ferscha and Friedemann Mattern (Eds.). 1–17.

[5] Sourav Bhattacharya and Nicholas D Lane. 2016. Sparsification and separation of deep learning layers for constrained resource inference on wearables. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*. 176–189.

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

[7] Andreas Bulling, Jamie A. Ward, and Hans Gellersen. 2012. Multimodal recognition of reading activity in transit using body-worn sensors. *ACM Transactions on Applied Perception* 9, 1 (2012), 2:1–2:21.

[8] 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.

[9] Federico Cruciani, Ian Cleland, Kåre Synnes, and Josef Hallberg. 2018. Personalized Online Training for Physical Activity monitoring using weak labels. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 567–572.

[10] Jordana Dahmen, Alyssa La Fleur, Gina Sprint, Diane Cook, and Douglas L Weeks. 2017. Using wrist-worn sensors to measure and compare physical activity changes for patients undergoing rehabilitation. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 667–672.

[11] Abu Zaher Md Faridee, Md Abdullah Al Hafiz Khan, Nilavra Pathak, and Nirmalya Roy. 2019. AugToAct: Scaling Complex Human Activity Recognition with Few Labels. In *Proceedings of the 16th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*.

[12] Jordan Frank, Shie Mannor, and Doina Precup. 2010. Activity and gait recognition with time-delay embeddings. In *AAAI Conference on Artificial Intelligence*.

[13] Mehmet Gövercin, Y Költzsch, M Meis, S Wegel, M Gietzelt, J Spehr, S Winkelbach, M Marschollek, and E Steinhagen-Thiessen. 2010. Defining the user requirements for wearable and optical fall prediction and fall detection devices for home use. *Informatics for Health and Social Care* 35, 3-4 (2010), 177–187.

[14] Yu Guan and Thomas Plötz. 2017. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–28.

[15] Nils Yannick Hammerla, James Fisher, Peter Andras, Lynn Rochester, Richard Walker, and Thomas Plötz. 2015. PD disease state assessment in naturalistic environments using deep learning. In *AAAI conference on artificial intelligence*.

[16] Nils Y. Hammerla, Shane Halloran, and Thomas Plötz. 2016. Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables. In *Proceedings of International Joint Conference on Artificial Intelligence*. 1533–1540.

[17] Harish Haresamudram, David Anderson, and Thomas Plötz. 2019. On the Role of Features in Human Activity Recognition. In *Proceedings of International Symposium on Wearable Computers*. 78–88.

[18] HM Sajjad Hossain, MD Abdullah Al Haiz Khan, and Nirmalya Roy. 2018. DeActive: scaling activity recognition with active deep learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–23.

[19] Tâm Huynh and Bernt Schiele. 2005. Analyzing features for activity recognition. In *Proceedings Conference on Smart Objects and Ambient Intelligence: Innovative Context-aware Services: Usages and Technologies*. 159–163.

[20] 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).

[21] Md Abdullah Al Hafiz Khan, Nirmalya Roy, and Archan Misra. 2018. Scaling human activity recognition via deep learning-based domain adaptation. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 1–9.

[22] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *The International Conference on Learning Representations*.

[23] Matthias Kranz, Andreas MöLler, Nils Hammerla, Stefan Diewald, Thomas PlöTz, Patrick Olivier, and Luis Roalter. 2013. The mobile fitness coach: Towards individualized skill assessment using personalized mobile devices. *Pervasive and Mobile Computing* 9, 2 (2013), 203–215.

[24] Oscar D Lara, Alfredo J Pérez, Miguel A Labrador, and José D Posada. 2012. Centinela: A human activity recognition system based on acceleration and vital sign data. *Pervasive and Mobile Computing* 8, 5 (2012), 717–729.

[25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

[26] Andrea Mannini, Mary Rosenberger, William L Haskell, Angelo M Sabatini, and Stephen S Intille. 2017. Activity recognition in youth using single accelerometer placed at wrist or ankle. *Medicine and science in sports and exercise* 49, 4 (2017), 801.

[27] Akhil Mathur, Tianlin Zhang, Sourav Bhattacharya, Petar Veličković, Leonid Joffe, Nicholas D. Lane, Fahim Kawsar, and Pietro Lió. 2018. Using Deep Data Augmentation Training to Address Software and Hardware Heterogeneities in Wearable and Smartphone Sensing Devices. In *Proceedings of the 17th ACM/IEEE International Conference on Information Processing in Sensor Networks*.

[28] Vishvak S. Murahari and Thomas Plötz. 2018. On Attention Models for Human Activity Recognition. In *Proceedings of International Symposium on Wearable Computers*. 100–103.

[29] Dzung Tri Nguyen, Eli Cohen, Mohammad Pourhomayoun, and Nabil Alshurafa. 2017. SwallowNet: Recurrent neural network detects and characterizes eating patterns. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 401–406.

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

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in Pytorch. In *NIPS Autodiff Workshop*.

[32] Thomas Plötz, Nils Y Hammerla, and Patrick L Olivier. 2011. Feature learning for activity recognition in ubiquitous computing. In *Twenty-second international joint conference on artificial intelligence*.

[33] Thomas Plötz, Nils Y Hammerla, Agata Rozga, Andrea Reavis, Nathan Call, and Gregory D Abowd. 2012. Automatic assessment of problem behavior in individuals with developmental disabilities. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 391–400.

[34] C. Randell and H. Muller. 2000. Context awareness by analysing accelerometer data. In *Proceedings of the 4th International Symposium on Wearable Computers*. 175–176.

[35] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L. Littman. 2005. Activity recognition from accelerometer data. In *Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence*. 1541–1546.

[36] Attila Reiss and Didier Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *Proceedings of International Symposium on Wearable Computers*. 108–109.

[37] Charissa Ann Ronao and Sung-Bae Cho. 2015. Deep convolutional neural networks for human activity recognition with smartphone sensors. In *International Conference on Neural Information Processing*. 46–53.

[38] Roberto Luis Shinmoto Torres, Qinfeng Shi, Anton van den Hengel, and Damith C. Ranasinghe. 2017. A hierarchical model for recognizing alarming states in a batteryless sensor alarm intervention for preventing falls in older people. *Pervasive Mob. Comput.* 40, C (2017), 1–16.

[39] Thomas Stiefmeier, Daniel Roggen, Georg Ogris, Paul Lukowicz, and Gerhard Tröster. 2008. Wearable activity tracking in car manufacturing. *IEEE Pervasive Computing* 2 (2008), 42–50.

[40] Roberto L Shinmoto Torres, Renuka Visvanathan, Derek Abbott, Keith D Hill, and Damith C Ranasinghe. 2017. A battery-less and wireless wearable sensor system for identifying bed and chair exits in a pilot trial in hospitalized older people. *PloS one* 12, 10 (2017), 1–25.

[41] Terry T. Um, Franz M. J. Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. 2017. Data Augmentation of Wearable Sensor Data for Parkinson's Disease Monitoring Using Convolutional Neural Networks. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. Association for Computing Machinery.

[42] Alireza Abedin Varamin, Ehsan Abbasnejad, Qinfeng Shi, Damith C Ranasinghe, and Hamid Rezatofighi. 2018. Deep Auto-Set: A Deep Auto-Encoder-Set Network for Activity Recognition Using Wearables. In *the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. 246–253.

[43] 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.

[44] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters* 119 (2019), 3–11.

[45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 7794–7803.

[46] Jamie A Ward, Paul Lukowicz, and Hans W Gellersen. 2011. Performance metrics for activity recognition. *ACM Transactions on Intelligent Systems and Technology* 2, 1 (2011), 1–23.

[47] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*. 499–515.

[48] Asanga Wickramasinghe, Damith C Ranasinghe, Christophe Fumeaux, Keith D Hill, and Renuka Visvanathan. 2017. Sequence learning with passive RFID sensors for real-time bed-egress recognition in older people. *IEEE Journal of Biomedical and Health Informatics* 21, 4 (2017), 917–929.

[49] Jian Bo 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 *Proceedings of the 24th International Conference on Artificial Intelligence* (Buenos Aires, Argentina). 3995–4001. http://dl.acm.org/citation.cfm?id=2832747.2832806

[50] Rui Yao, Guosheng Lin, Qinfeng Shi, and Damith C. Ranasinghe. 2018. Efficient dense labelling of human activity sequences from wearables using fully convolutional networks. *Pattern Recognition* 78 (2018), 252 – 266. https://doi.org/10.1016/j.patcog.2017.12.024

[51] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang. 2014. Convolutional Neural Networks for human activity recognition using mobile sensors. In *6th International Conference on Mobile Computing, Applications and Services*. 197–205. https://doi.org/10.4108/icst.mobicase.2014.257786

[52] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

[53] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. Self-Attention Generative Adversarial Networks. In *International Conference on Machine Learning*. 7354–7363.

[54] Licheng Zhang, Xihong Wu, and Dingsheng Luo. 2015. Human activity recognition with HMM-DNN model. In *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*. IEEE, 192–197.

[55] Mi Zhang and Alexander A. Sawchuk. 2012. Motion primitive-based human activity recognition using a bag-of-features approach. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. 631–640.

[56] Nenggan Zheng, Jun Wen, Risheng Liu, Liangqu Long, Jianhua Dai, and Zhefeng Gong. 2018. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *Thirty-Second AAAI conference on artificial intelligence*.

A WINDOW SIZE ANALYSIS

The window sizes we have adopted are based on other benchmarking studies—see Section 4.2—aiming to achieve improved performance. However, the window size selected can impact activity recognition performance [50] and is therefore an important parameter to consider in a given HAR task. Hence, for completeness, we extend the activity recognition experiments in Section 4.4 by analysing the effect of sampling window size on the performance of HAR baselines for the Opportunity benchmark dataset with respect to our proposed framework.

We increased the sampling window size adopted in Section 4.4 and incorporated 48 samples in each data segment (*i.e.*, W=48). As in Section 4.4, fully non-overlapping windows are employed to generate sensor segments and 3-fold stratified cross-validation is adopted to produce the training and testing datasets. Subsequently, each fold is in turn used once for evaluation while the remaining folds are utilized as the training data. In Table 7, we summarize the resulting class-average F-score for each baseline and present a comparison against the activity recognition performance achieved with sampling window size of 24.

Interestingly, we can see that all baselines demonstrate a lower activity recognition performance with the increased sampling window size. However, the performance degradation is more pronounced for the baselines that incorporate the window based labelling scheme as compared with the fully convolutional baseline [50] that adopts a dense labelling scheme for training and inference. This can be understood since the sensor sequences in

Table 7. An investigation of the impact of sampling window size on segment-wise activity recognition performance based on class-averaged f1-scores with cross-fold evaluation.

HAR Model	Sampling Size	
	W=24	W=48
LSTM Learner Baseline [14]	75.6 ± 0.7	72.8 ± 1.2
DeepConvLSTM [30]	73.0 ± 0.8	70.2 ± 1.1
b-LSTM-S [16]	77.2 ± 1.1	74.1 ± 1.5
Dense Labeling [50]	78.5 ± 0.4	77.1 ± 0.7
Att. Model [28]	78.1 ± 0.2	75.6 ± 0.7
Ours	**81.1 ± 0.2**	**77.8 ± 0.6**

the Opportunity benchmark dataset include short duration gestures encompassing minority classes. Accordingly, adopting a sampling window size that is too large results in the *multi-class windows problem* [42, 50]; *i.e.,* all samples in a sliding window may not share the same activity label. Here, the window based labelling scheme approximates the ground-truth window label with either the last or most frequent activity label observed in the window. This inevitably leads to loss of original information required for training an effective activity recognition model and, in particular, negatively impacts the performance on minority classes. Nevertheless, our proposed HAR framework still performs better than the baselines.