Triple Cross-Domain Attention on Human Activity Recognition Using Wearable Sensors

Yin Tang, Lei Zhang, Qi Teng, Fuhong Min, and Aiguo Song, *Senior Member, IEEE*

***Abstract*—Efficiently identifying activities of daily living (ADL) provides very important contextual information that is able to improve the effectiveness of various sports tracking and healthcare applications. Recently, attention mechanism that selectively focuses on time series signals has been widely adopted in sensor based human activity recognition (HAR), which can enhance interesting target activity and ignore irrelevant background activity. Several attention mechanisms have been investigated, which achieve remarkable performance in HAR scenario. Despite their success, prior these attention methods ignore the cross-interaction between different dimensions. In the paper, in order to avoid above shortcoming, we present a triplet cross-dimension attention for sensor-based activity recognition task, where three attention branches are built to capture the cross-interaction between sensor dimension, temporal dimension and channel dimension. The effectiveness of triplet attention method is validated through extensive experiments on four public HAR dataset namely UCI-HAR, PAMAP2, WISDM and UNIMIB-SHAR as well as the weakly labeled HAR dataset. Extensive experiments show consistent improvements in classification performance with various backbone models such as plain CNN and ResNet, demonstrating a good generality ability of the triplet attention. Visualization analysis is provided to support our conclusion, and actual implementation is evaluated on a Raspberry Pi platform.**

***Index Terms*—Activity recognition, attention, weakly supervised learning, wearable sensors, convolutional neural networks.**

I. Introduction

DURING recent years, human activity recognition (HAR) using various motion sensors embedded in smartphones or other wearable devices has become a new research hotspot in ubiquitous and mobile computing due to the rapid growth of application demands in domains such as health care, life assistance and exercise monitoring. Sensor-based HAR task [1]–[3] can be regarded as a multi-channel time series classification problem, in which a fixed length sliding window is utilized to split time series signal into equal segments. Various traditional machine learning approaches such as Logistic Regression, Decision Trees, Random Forest and native Bayesian methods have been widely adopted in HAR areas [4], [5], which have achieved remarkable performance. However, these shallow learning methods often require feature extraction from the data, which heavily depends on expert knowledge from specific domain [6]. The handcrafted feature engineering inevitably restricts the practicability of the HAR model when the task is transferred from one domain to the other.

Lately, deep learning techniques [7]–[9] have broken the limit to shallow learning methods, which enables richer feature representations to be learned automatically with no need of domain-specific knowledge. In particular, compared with these shallow learning methods with handcrafted features that only can recognize low-level or simple activities, convolutional neural networks (CNNs) [7] are more suitable for recognizing more complex activities because of its advantages of local dependencies and scale invariance. CNNs have significantly pushed state-of-the-art performance in HAR scenario given its rich representation ability. Despite its effectiveness, deep HAR still faces many key challenges, one of which is ground truth annotation [10]. In a supervised learning setting, the use of deep CNNs relies heavily on strictly labeled activity sensor data for training. Nevertheless, compared with HAR that uses video data (e.g. GoPro motion camera), the high dimensional time series data from motion sensors such as accelerometer is harder to interpret and annotate, which has brought cumbersome and arduous difficulties to HAR.

Such challenges can be tackled by utilizing attention mechanism [11], [12], which shows great potential in a large variety of computer vision or natural language processing tasks. The learning of attention weights can aid the model to focus on the target object, thereby improving the recognition accuracy. On the other hand, for an annotator who is in charge of recording sensor data, it is much simpler to identify whether a target activity occurs in a long sensor sequence. If a specific activity can be recognized according to coarse or weakly labels, it will significantly ease the burden of manual labeling. Intuitively, the attention mechanism is capable of aiding to tell where or what to focus via enhancing selectively the interesting target activity while weakening redundant or even other irrelevant information. Therefore, it deserves further research whether the attention mechanism can promote the state-ot-the-art performance of HAR via consciously improving output feature maps of convolutional network.

Recently, hard attention [13] and soft attention [14] have been proposed respectively in weakly supervised learning scenario,

Manuscript received July 15, 2021; revised November 2, 2021; accepted November 22, 2021. This work was supported in part by the National Science Foundation of China under Grant 61971228 and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20191371. *(Corresponding author: Lei Zhang.)*

Yin Tang, Lei Zhang, Qi Teng, and Fuhong Min are with the School of Electrical and Automation Engineering, Nanjing Normal University, Nanjing 210023, China (e-mail: yinntag@gmail.com; leizhang@njnu.edu.cn; teqi159@gmail.com; minfuhong@njnu.edu.cn).

Aiguo Song is with the School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: a.g.song@seu.edu.cn).

Digital Object Identifier 10.1109/TETCI.2021.3136642

in which sensor data does not need to be strictly labeled. One only needs to know which kind of activity has occurred in a long sensor sequence without the specific location of the target activity. The learned attention weights can help to focus on the target activity from a long background sequence. However, the two attention mechanisms can only tell us where to focus, ignoring channel information, which plays an important role in deciding what to focus on. The dual attention network [15] in weakly supervised HAR applications has demonstrated the advantages of computing multi-attention. Although dual attention mechanism provides significant performance improvements in HAR scenario, it does not account for the importance of capturing cross-dimension interaction, which have successfully shown a favorable impact in computer vision task.

In the paper, we firstly propose a novel triplet attention network in HAR scenario, which mainly blends three attention branches. Given a standard convolutional layer, let us consider its input tensor with shape $C \times T \times S$, in which C, T and S are the channel, temporal and sensor modality respectively. Each branch is responsible for capturing cross-dimension interaction between the spatial dimensions $(T \times S)$ and channel dimension (C) of sensor input. We conduct extensive experiments to evaluate the triplet attention network on several public benchmark HAR datasets consisting of UCI-HAR dataset [16], PAMAP2 dataset [17], WISDM dataset [18] and UNIMIB-SHAR [19] dataset, as well as the weakly labeled HAR dataset. The experimental results manifest that triplet attention perform better than one or two attention respectively. The main contributions of this method are summarized as follows:

- Firstly, we propose a new architecture relying on triple attention mechanism for HAR task, which could aid to extract richer activity feature representations via building three attention branches to capture cross-interaction between sensor dimension, temporal dimension, and channel dimension.
- Second, the triple attention tends to strength the importance of cross-dimension interaction, which is superior to its corresponding predecessors, i.e., one or two attention respectively.
- Finally, extensive experiments are conducted on several public HAR datasets, and several key hyperparameters are analyzed in details. We also examine actual implementation on a Raspberry Pi platform with ARM-based computing core. The experimental results manifest that triplet attention method could provide competitive results at a negligible computational cost.

The rest of the paper is organized as follows. Section II introduces related works on attention based HAR methods. Section III presents an overall architecture of the proposed triplet attention. In Section IV and Section V, we detail experimental results obtained on four public HAR datasets and the weakly labeled HAR dataset, which are compared with the existing SOTAs. Moreover, several ablation studies about the triplet attention method are provided. Section VI summarizes our conclusion.

II. Related Works

Attention in human perception is everywhere, which selectively focus on interesting parts while suppressing the other irrelevant or even misleading information. During the past few years, the attention mechanism has been widely incorporated into various deep CNN architectures, which can significantly improve performance on large scale computer vision tasks. Several related attention mechanisms to our work are introduced as follows. Hu *et al.* for the first time proposed the Squeeze-and-Excitation Networks (SENet) [20], which successfully utilizes global average-pooled features to compute channel attention in an efficient way. This was followed by the introduction of Convolutional Block Attention Module (CBAM) [21], in which the combination of channel attention and spatial attention leads to significant performance improvement. Global-Context Networks (GC-Net) [22] proposed a novel NL-block, which takes into account global context modeling and lightweight modular design. More recently, Landskape *et al.* [23] adopted triplet attention mechanism for a variety of computer vision tasks, which concentrates on cross dimension interaction. However, attention mechanism has been rarely explored in sensor based HAR scenario.

Due to the popularity of attention mechanism in deep learning, a surge of research hotspot has been emerging to utilize attention for handling HAR tasks. Recently, Ma *et al.* [24] proposed a novel AttnSense for HAR, which has incorporated the attention mechanism into a Gated Recurrent Units (GRU) subnet for capturing the dependencies of sensor signals in both spatial and temporal domains. Zeng *et al.* [25] highlighted the important part of different time series and sensor modalities by designing temporal attention and sensor attention with Long Short Term Memory (LSTM). When compared to recurrent neural networks, CNN has better ability of feature extraction. In recent works, two mainstream attention mechanisms, hard attention [13] and soft attention [14], have been incorporated into convolutional architecture to perform the weakly supervised HAR tasks, which ignores the importance of sensor channels. Gao *et al.* [15] proposed a novel dual attention method for HAR that blends channel attention and spatial attention, demonstrating obvious superiority in handling multimodal HAR task. In order to capture cross-domain interaction of sensor signals, we for the first time propose a new triple attention network for HAR task, which is able to extract meaningful cross-dimensional features via building three main attention branches.

III. Model

Actually, the channel attention [20] often needs to compute a singular weight, i.e., a scalar for each channel of input sensor tensor, which can be used to scale these feature maps for generating attention effect. Although the lightweight channel attention is very effective, there is an obvious shortcoming in its computing process. Usually, in order to produce these singular weights for each channel, one has to use global average pooling to spatially subsample the input sensor tensor along each channel, which inevitably leads to a significant loss in

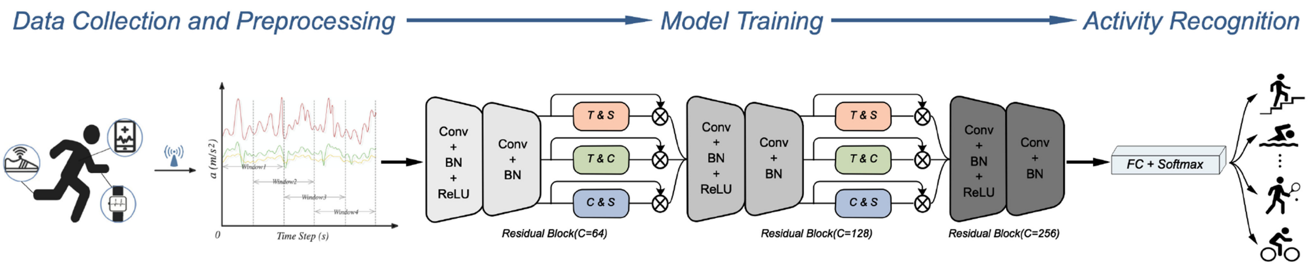


Fig. 1. The overview of our proposed triplet attention (TA) module for HAR system. It simply describes the three pipelines: *Data collection and preprocessing, model training as well as activity recognition*. $T\&S$, $T\&C$ and $C\&S$ represent temporal and sensor interaction, temporal and channel interaction, as well as channel and sensor interaction, respectively.

spatial information. That is to say, the cross-dependence between channel dimension and spatial dimension is lost due to the subsampling by global average pooling. To avoid above drawback, the dual attention [21] computes the spatial attention, which is used as a complement to the channel attention. Simply speaking, the channel attention tells *"what channel"* to focus on, and meanwhile the spatial attention tells *"where in the channel"* to focus. However, its shortcoming lies in that the channel attention and spatial attention are computed independently. As a result, the cross-interaction [23] between the two is ignored. To address above shortcoming, we present the use of cross-dimension interaction for HAR task, which builds three attention branches to capture the interaction between the spatial dimensions (i.e., temporal and sensor modality) and channel dimension of input sensor tensor. The time series sensor data is firstly preprocessed with the sliding window technique, which is then fed into a standard convolution layer. For a given convolutional layer, let us consider an input tensor χ with shape $C \times T \times S$, in which C, T and S are the channel, temporal and sensor modality respectively. The cross-dimension interaction is introduced via three parallel attention branches, each of which is responsible for capturing dependencies between the (*C, T*), (*C, S*) and (*T, S*) dimensions of sensor input tensor respectively. Finally, the weights of three cross-domain attentions are learned in the triplet attention. Fig. 1 shows the framework based triplet attention in HAR system.

A. Rethinking Channel Attention

Let us revisiting channel attention [20], [21] via considering a convolutional layer and its corresponding input tensor $\chi \in R^{C\times H\times W}$. For each independent channel, the SE block compute the channel attention via utilizing global average-pooling technique to squeeze the $H \times W$ dimension. The channel weights are generated by two FC layers that is followed by sigmoid non-linearity function. The research in CBAM shows that the max-pooling operation is also a good choice for aggregating discriminative features. Referring to SE and CBAM block, the weight of the channel attention combining average-pooling and max-pooling can be expressed as:

$$W_c = \sigma\left(f_{\{w_1,w_2\}}(\mathrm{AP}(\chi)) + f_{\{w_1,w_2\}}(\mathrm{MP}(\chi))\right), \qquad (1)$$

in which $\mathrm{AP}(\chi) = \frac{1}{WH}\sum_{i=1,j=1}^{W,H}\chi_{ij}$ and $\mathrm{MP}(\chi) = \max_{i=1,j=1}^{W,H}\chi_{ij}$ is global average-pooling and max-pooling operation respectively. σ is Sigmoid function. The Eq. (1) can further be expressed as:

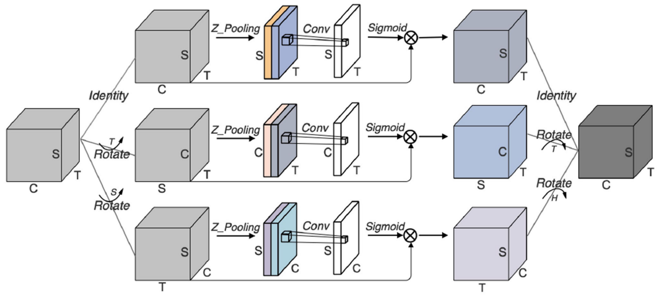


Fig. 2. Description of the triplet attention with three branches.

$$W_c = \sigma\left(w_2\,\mathrm{ReLU}\left(w_1\mathrm{AP}(\chi)\right) + w_2\,\mathrm{ReLU}\left(w_1\mathrm{MP}(\chi)\right)\right). \qquad (2)$$

Note that two FC layers are used as indicated above, where the size of w_1 and w_2 is set by adjusting a scaling factor. On the whole, the Eq. (2) uses two linear projections to assign corresponding weights to each channel.

B. Triplet Attention Using Cross-Domain Interaction

As mentioned above, the triplet attention has three attention branches, which is built via using cross-domain attention module. The given input sensor tensor $\chi \in R^{C\times T\times S}$ will be sent to an attention branch respectively. In fact, there are several aggregation for attention weights, such as addition, multiplication and concatenation. In order to make computation more lightweight, the Z_Pooling technique [23] is used, which can preserve richer feature representations, and meanwhile compressing depth. Unlike the CBAM, two pooled features are concatenated to aggregate information, which can be formulated as:

$$\mathrm{Z\_Pooling}(\chi) = \left[\mathrm{MaxPool}_{0\mathrm{d}}(\chi),\ \mathrm{AvgPool}_{\ 0\mathrm{d}}(\chi)\right], \qquad (3)$$

where 0d is the 0th-dimension across which the max and average pooling operations occur. That is to say, the Z_Pooling can reduce the zeroth dimension given input tensor to two by aggregating the two pooled features. For example, a sensor tensor of shape $(C \times T \times S)$ will be transformed into an output tensor of shape $(2 \times T \times S)$ through Z_Pooling. In fact, every branch is implemented by the following three steps, which can generate a refined tensor. The triplet across-domain attention is shown in Fig. 2.

The first branch is in charge of calculating the cross-interaction between temporal dimension and channel dimension. Firstly, the tensor χ with input shape $(C \times T \times S)$ is rotated 90° counter-clockwise along the *T* axis to generate a new tensor $\widehat{\chi}_1$ with the shape $(S \times T \times C)$; $\widehat{\chi}_1$ is then fed into Z_Pooling, which can generate a tensor $\widehat{\chi}_1^*$ with the shape $(2 \times T \times C)$; As a third stage, $\widehat{\chi}_1^*$ is passed through a standard convolution with $k \times 1$ kernel size (e.g., 3×1, 5×1), followed by a batch normalization, which results in an intermediate output (shape is $1 \times T \times C$); After passed through a sigmoid activation, the intermediate output is turned into attention weights ω_1, which are applied to $\widehat{\chi}_1$, then rotated 90° clockwise along the *T* axis to keep the shape of input χ.

In the second branch, the cross interaction between sensor dimension and channel dimension can be computed in a similar way. The tensor χ with input shape $(C \times T \times S)$ is rotated 90° counter-clockwise along the *S* axis, which provides a new tensor $\widehat{\chi}_2$ with the shape $(T \times C \times S)$; $\widehat{\chi}_2$ is then passed through Z_Pooling layer, which can generate a tensor $\widehat{\chi}_2^*$ with the shape $(2 \times C \times S)$; At the third stage, $\widehat{\chi}_2^*$ is passed through a standard convolution with $k \times 1$ kernel size (e.g., 3×1, 5×1), followed by a batch normalization, which results in an intermediate output $(1 \times C \times S)$; After passed through a sigmoid activation, the intermediate output is turned into attention weights ω_2, which are applied to $\widehat{\chi}_2$, then rotated 90° clockwise along the *S* axis to maintain the shape of input χ.

For the third branch, the channels of input tensor χ are reduced to two via using Z_Pooling operation, which provides the tensor $\widehat{\chi}_3$ with the shape $(2 \times T \times S)$; $\widehat{\chi}_3$ is then fed into a standard convolution with $k \times 1$ kernel size (e.g., 3×1, 5×1), followed by a batch normalization, which results in an intermediate output; The output is then fed into a sigmoid activation, which generates the attention weights ω_3 with shape $(1 \times T \times S)$; The attention weights ω_3 are then applied to the input χ.

Finally, the three refined tensors from three branches are aggregated via learning three weight parameters. For simplicity, it can be represented as:

$$\mathrm{Y} = \frac{1}{3}\left(\mathrm{R}\left(\omega_1 \widehat{\chi}_1\right)\right) + \frac{1}{3}\left(\mathrm{R}\left(\omega_2 \widehat{\chi}_2\right)\right) + \frac{1}{3}\left(\omega_3 \widehat{\chi}_3\right), \qquad (4)$$

where ω_1, ω_2 and ω_3 are the three cross-dimensional attention weights. The $\widehat{\chi}_1$, $\widehat{\chi}_2$ and $\widehat{\chi}_3$ represent the refined tensor, which can be obtained via rotating the input tensor χ 90° counter-clockwise along *T* axis and *S* axis respectively. R means 90° clockwise rotation. Compared with above simple averaging, the model performance can be further improved by introducing a combination of three learnable weight parameters, which can be formulated as:

$$\mathrm{Y} = \alpha_1\left(\mathrm{R}\left(\omega_1 \widehat{\chi}_1\right)\right) + \alpha_2\left(\mathrm{R}\left(\omega_2 \widehat{\chi}_2\right)\right) + \alpha_3\left(\omega_3 \widehat{\chi}_3\right), \qquad (5)$$

which will be detailed in Section V. B.

IV. Experiment

In the following, we will describe the experimental setup and main results in detail. All the experiments are divided into three parts. Firstly, to demonstrate the superiority of the proposed triplet attention method, we compare classification results on four publicly available HAR datasets including UCI-HAR, WISDM, PAMAP2 and UNIMIB-SHAR. All datasets have been recorded by various sensors such as accelerometers and gyroscope, which can reflect human activities in different scenarios. Secondly, detailed ablation experiments are provided to analyze the impact of several hyperparatmers. Finally, we evaluate the performance of triplet attention in the weakly supervised activity recognition task, which uses the weakly labeled dataset collected by He *et al.* [26]. The impact of different cross dimension attention for HAR is explored.

A. Training Details

Our model is trained by minimizing cross-entropy (CE) loss using mini-batch gradient descent, where the batch size is set to 200. An Adam optimizer with dynamic learning rate is used. The initial learning rate is set to 0.001, which will be reduced by a factor of 0.1 after every 100 epochs. All the experiments are implemented in Python using PyTorch framework backend on a server with an Intel i7-6850 K CPU, 64 GB RAM and NVIDIA RTX 3090 GPU. Since there is highly imbalanced class in various naturalistic activity datasets, different class weights need to be reconsidered according their sample proportion. Thus, the mean F_1 score [27] is used as metric to evaluate final performance.

B. Datasets

A comprehensive evaluation of the proposed method is conducted using four popular HAR datasets that include both high-dimensional and low-dimensional sensor modalities. The sensor data is segmented using sliding window technique with different window size and step length, which has an important influence on recognition system's practical performance. We select the same window size and step length adopted in previous successful cases [15], [27] to ensure fair comparison.

• *UCI-HAR [16]:* This dataset was collected by recruiting 30 volunteers. Everyone is required to wear a Samsung Galaxy S II smartphone around their waist to perform six simple daily activities consisting of *"Walking," "Going upstairs," "Going downstairs," "Sitting," "Standing," "Laying"*. Three-axis accelerometer and gyroscope sensor signals are recorded at a fixed frequency of 50 Hz. The raw data is firstly preprocessed by the noise filter, which is then segmented by the sliding window with a fixed length of 128 and 50% overlap. Finally, the whole dataset has been randomly split into two parts, where 70% for training and 30% for test.

• *PAMAP2 [17]:* The Physical Activity Monitoring for Aging People 2 dataset is collected from 9 participants to perform 12 daily activities *("Walking", "Lying down", "Standing", etc.)* and excises *("Watching TV," "Computer work," "Car driving," etc.)* The three inertial measurement units (IMUs) were placed on the hand, chest, and ankle of each subject to collect raw sensor data from accelerometer, gyroscope, magnetometer, and heart rate. At a 100 Hz sampling rate, the collection process lasts around 10 hours. To perform fair comparisons with previous works [27], the sensor signal is down-sampled into 33.3 Hz and with a 5.12 s sliding window and 78% overlap. Generally, this

TABLE I
BRIEF DESCRIPTION FOR EACH BACKBONE

	# Layer	**# Number of Kernels**	**Batch Norm**	**Activation Function**
Standard CNN	Conv2d_1	64	✓	ReLU
	Conv2d_2	128	✓	ReLU
	Conv2d_3	256	✓	ReLU
	FC_1	-	-	Softmax
Equally-sized ResNet	Conv2d_11	64	✓	ReLU
	Conv2d_12	64	✓	-
	Conv2d_21	128	✓	ReLU
	Conv2d_22	128	✓	-
	Conv2d_31	256	✓	ReLU
	Conv2d_32	256	✓	-
	FC_1	-	-	Softmax

dataset are randomly divided into two parts, in which 80% is used for training and 20% for test.

- *WISDM [18]:* The WISDM samples belong to 29 volunteer subjects who performed 6 discriminative human activities *("Walking", "Jogging," "Sitting," "Standing," "Going downstairs" and "Going upstairs")* by placing their mobile phones with Android operating system in front leg pocket. It contains 1,098,213 samples sampled at a rate of 20 Hz from a triaxial accelerometer sensor. Accordingly, the accelerometer sensor data will be preprocessed by a sliding window of 10 seconds and 95% overlap (200 readings/window). This dataset will be split into two parts, in which 80% for training and 20% for test.
- *UNIMIB-SHAR [19]:* This dataset includes 11,771 samples from 30 test subjects for the use of human pose estimation and fall detection. During data collection, a Samsung Galaxy Nexus I9250 smartphone is embedded with a Bosh BMA220 3D accelerometer, which measured sensor signals at a frequency of 50 Hz. The dataset consists of 17 fine-grained categories, which is further split into 9 classes of activities of daily living and 8 classes of falls. Accordingly, the sliding windows of data are produced by a size T = 151 (151 readings/window). Our experiment requires dividing up this dataset into two parts, where 70% for training and the rest for test.

C. Comparison Algorithms

The triplet attention mechanism can be used to update the existing network architectures at a negligible cost. Extensive experiments are conducted to evaluate the performance gain brought by the triplet attention part. To demonstrate generalization ability of the triplet attention and analyze how it influence the classification results, we use standard CNN, equally-sized ResNet [28] as our backbones, which is introduced as follows. Table I presents their detailed architectures.

- *Standard CNN:* The baseline CNN consists of three standard convolution layers. Batch normalization and ReLU activation are applied after each convolutional layer.
- *Equally-sized ResNet:* To demonstrate the effectiveness of triplet attention module, we also incorporate it into the residual networks proposed in the previous work [28]. The residual network consisting of three residual blocks with the same architecture is used as our baseline, in which the contribution of triplet attention mechanism is further evaluated.

V. DISCUSSION

The proposed method is compared with both baselines on four public HAR datasets. We have three major observations from Table II. Firstly, it can be seen that the ResNet outperforms all original CNN by a large margin due to its strong feature extraction ability. For instance, the ResNet outperforms standard CNN by 0.21% in terms of accuracy on UCI-HAR dataset. Secondly, the results indicate that our triplet attention can further improve performance by clear gains compared to these baselines. Results from Table II, it can be easily seen that the proposed method achieves 1.35% and 0.62% performance gains on PAMAP2 dataset when using CNN and ResNet as backbones respectively. Similar results are also reflected on WISDM dataset. Meanwhile, the triplet attention with almost the same complexity is superior to the original CNN and equally-sized ResNet by 0.96% and 1.47% in terms of accuracy on UNIMIB-SHAR dataset, respectively. This comparison consistently verifies the effectiveness of our model on different baselines. That is to say, it can boost the accuracy of baselines significantly, demonstrating that it can generalize well on various models on HAR dataset. Lastly, we note that there is no extra parameter caused by the triplet attention compared to their plain counterparts, which motivates us to update new light-weight network by applying our proposed module.

In addition, the triplet attention method is compared with the other state-of-the-art algorithms [15], [31], [32], [35] accordingly. Table II summarizes main experimental results. Compared with recent state-of-the-art methods, it obtains better or competitive results without increasing model complexity. As shown in Table II, we observe that the integration of the triplet attention with ResNet is superior to Xiao *et al.*'s [31] result by 0.44% that uses a federated learning method on UCI-HAR dataset. Compared with Teng *et al.*'s [27] result using local loss method, the triplet attention achieves 0.23% performance gain in terms of accuracy on PAMAP2 dataset. On WISDM dataset, our method is also able to beat Janarthanan *et al.*'s [35] result by 1.11%. Finally, the triplet attention also achieves very competitive accuracy on UNIMIB-SHAR dataset, which outperforms all previous results [15], [19], [27], [36]. In particular, as mentioned above, it indicates that the triplet attention can be used to update the existing network architecture.

A. Visualization Analysis

To evaluate whether the cross-dimensional interaction provided by triplet attention can capture richer internal representations of sensor signals, we provide sample visualization to better understand the cross-dimensional interaction between sensor dimension, temporal dimension and channel dimension on PAMAP2 dataset. The results show that our triplet attention

TABLE II
THE CLASSIFICATION PERFORMANCE ON FOUR HAR DATASETS

Method		**UCI-HAR**		**PAMAP2**		**WISDM**		**UNIMIB-SHAR**	
		F_1 score(%)	Para.(M)	F_1 score(%)	Para.(M)	F_1 score(%)	Para.(M)	F_1 score(%)	Para.(M)
Standard CNN	Baseline	96.12	0.34	91.13	0.86	96.73	0.42	74.42	0.39
	+TA	**96.60**	0.34	**92.48**	0.86	**97.34**	0.42	**75.38**	0.39
Equally-sized ResNet	Baseline	96.33	0.85	92.58	1.37	98.10	1.01	77.08	0.90
	+TA	**96.77**	0.85	**93.20**	1.37	**98.61**	1.01	**78.55**	0.90
Related Research		Anguita *et al* [16] 95.18		Ma *et al* [24] 89.30		Ignatov *et al* [30] 93.32		Gao *et al* [15] 77.12	
		Khan *et al* [29] 95.37		Zeng *et al* [25] 89.96		Wasle *et al* [33] 98.09		Micucci *et al* [19] 74.66	
		Ignatov *et al* [30] 96.63		Teng *et al* [27] 92.97		Ravi *et al* [34] 98.20		Teng *et al* [27] 78.07	
		Xiao *et al* [31] 96.33		Wan *et al* [32] 91.16		Janarthanan *et al* [35] 97.50		Liu *et al* [36] 76.14	

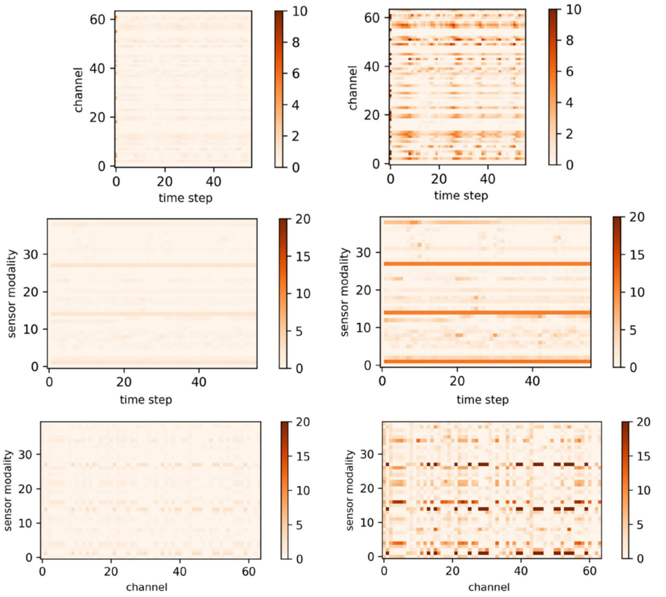


Fig. 3. Visualization of cross-dimension interaction between three attention branches.

module is superior to its plain counterparts. As can be seen from Fig. 3 (Left), the baseline without the triplet attention fails to focus on relevant features between cross-dimension. It is very evident that the triplet attention is able to provide richer activity feature representations due to the use of cross-dimension interaction (Fig. 3 Right vs. Left).

Furthermore, the visualization analysis is provided to evaluate the impact of sensor nodes placed on different body parts of each participant. As shown in Fig. 4, the three main IMUs placed on the wrist, ankle and chest of human body are used to collect various human activities, the attention weights of different sensor modalities are computed. Specifically, for *"nordic walking"* activity, the triplet attention puts a high emphasis on the hand sensor (hand_x), the ankle sensor (ankle_z) and the chest sensor (chest_z). For *"rope jumping"* activity, the triplet attention focuses on the hand sensor (hand_y, hand_z), the ankle sensor (ankle_x) and the chest sensor (chest_y). For *"vacuum cleaning"* activity, it pays much attention to the hand sensor (hand_y), the ankle sensor (ankle_y, ankle_z) and the chest sensor (chest_x, chest_z). In a word, compared with the baseline counterparts, it is more reasonable that triplet attention mechanism can treat different sensor modalities unequally.

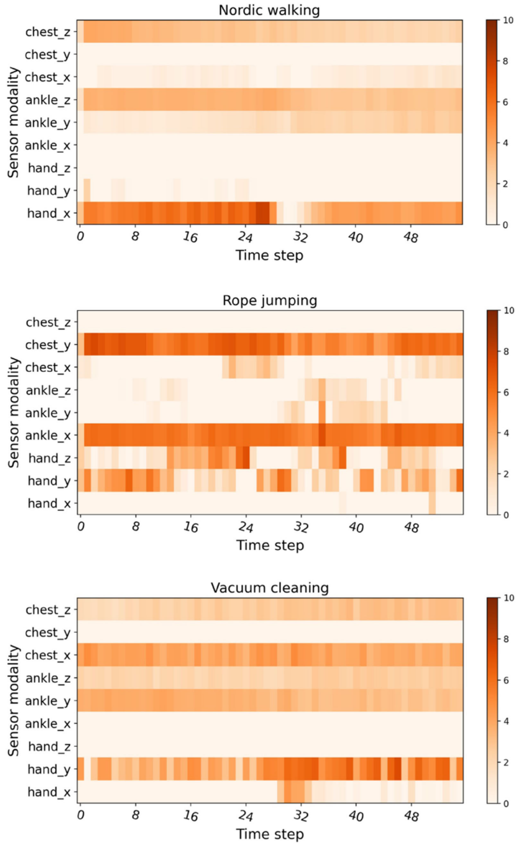


Fig. 4. Visualization of sensor attention on PAMAP2 dataset.

B. Ablation Studies

We further conduct ablation experiments on PAMAP2 dataset to validate the effectiveness of cross-dimension interaction via evaluating the impact of the branches in the triplet attention module. As shown in Table III, the triplet attention with all three branches turned on is denoted as full. *"Channel off"* indicates that the first two branches of the input sensor tensor without

TABLE III
PERFORMANCE FOR DIFFERENT TRIPLET ATTENTION BRANCHES

Model	$F_1(\%)$	Para.(M)
Standard CNN	91.13	0.861
Standard CNN + TA(channel off)	92.09	0.862
Standard CNN + TA(spatial off)	91.69	0.863
Standard CNN + TA(full)	**92.48**	0.863
Equally-sized ResNet	92.58	1.368
Equally-sized ResNet + TA(channel off)	93.01	1.369
Equally-sized ResNet + TA(spatial off)	92.77	1.369
Equally-sized ResNet + TA(full)	**93.20**	1.370

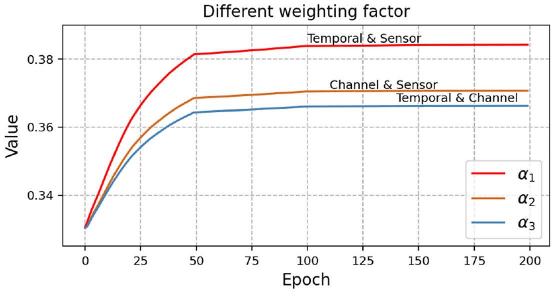


Fig. 5. The impact of different weighting factors.

TABLE IV
THE MEAN $F_1(\%)$ SCORE OF DIFFERENT AVERAGE METHODS

Model	CNN	ResNet
TA(with simple average)	92.19	92.96
TA(with weighted average)	92.48(**+0.29**)	93.21(**+0.25**)

permutation are turned off, which can be seen as a two-attention case. *"Spatial off"* indicates that the third branch, which is involved in permutations of the input sensor tensor, is turned off. It can be seen as a one-attention case. The results show that the triplet attention performs significantly better than one or two attention, as well as its plain counterpart without attention, which is in line with our statement. Here, we treat α_1, α_2 and α_3 as three learnable parameters rather than hyperparameters, whose initial values are 1/3. That is to say, their parameters are learned during training from data sets. Fig. 5 illustrate this learning process on PAMAP2 dataset. Results from Table IV, it can be seen that the learnable parameters are superior to simple averaging. The source code will be released at the website: https://github.com/yinntag/Triple-Cross-domain-Attention-for-HAR.

Actually, the window size has an important effect on activity recognition performance. Fixing an overlap rate, one can use a fixed-length sliding window to segment continuous sensor reading, which may produce continuous samples and each of them may be assigned a specific activity label. As a consequence, sensor signals are divided into windows of a fixed size and with no inter-window gaps, where an overlap between adjoining windows is tolerated in order to preserve the continuity of samples. Though sliding window has been normally utilized to perform segmentation, there is still no clear consensus on how to select an optimal window size. According to our intuition, reducing the window length will be more beneficial for a faster activity recognition, as well as reduced computational cost and energy consumption. Instead, increasing window length are usually used for the recognition of complex activities that last a longer time. We check the performance on PAMAP2 dataset with different window sizes to show the robustness of the proposed method. Results are summarized in Fig. 6. It can be seen that the classification performance evolves non-monotonically as the window size increases, which attains a peak value at 343. The triplet attention is able to reliably produce performance gain on every window size.

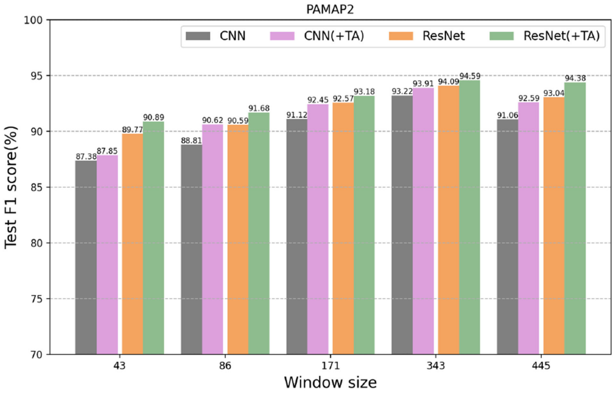


Fig. 6. The test mean $F_1(\%)$ score at different sliding window sizes.

In order to verify the robustness of the proposed method, we perform leave-one-subject-out cross validation on PAMAP2 dataset. Actually, it can be seen as a special case of k-fold cross validation, in which each individual person is treated as a "test" set. In other words, the number of folds should be equal to that of persons. As mentioned above, the PAMAP2 dataset is collected from nine subjects. Thus, in this case, the whole dataset will be divided into 9 folds. The average F_1 score is used as a metric to evaluate the final classification performance. We perform 9 folds, or iterations, of our model. Each time, the model will be trained on 8 subject and tested on the "left out" subject. Results are shown in Table V. It can be clearly seen that the triple attention can reliably produce performance gain over both baselines. Specifically, the triplet attention could produce a significant improvement in the leave-one-subject-out cross validation, which beats the baseline CNN by 1.15%, and ResNet by 0.45% respectively.

C. Actual Implementation of Raspberry Pi

To test the real-time performance of our model on mobile devices, the CNN integrated with triplet attention module is deployed in an embedded system based on Raspberry Pi OS with PyTorch installed. By importing the trained model file to the embedded system, we perform real-time prediction of activities on WISDM dataset. As shown in Fig. 7, the HAR system is deployed into a Raspberry Pi 3B+, which is equipped with an official supported Raspberry Pi operating system. It has a good compatibility with current popular deep learning library

TABLE V
THE MEAN F_1(%) SCORE OF LEAVE-ONE-SUBJECT-OUT EXPERIMENT ON PAMAP2 DATASET

Model	CNN	CNN(+TA)	ResNet	ResNet(+TA)
# Subject_01	90.12	**91.52**	**91.94**	91.89
# Subject_02	84.21	**85.42**	87.67	**88.01**
# Subject_03	96.05	**96.82**	95.90	**97.01**
# Subject_04	96.13	**96.47**	96.26	**96.55**
# Subject_05	93.21	**94.28**	95.01	**96.25**
# Subject_06	82.47	**83.02**	86.52	**87.90**
# Subject_07	70.79	**74.67**	**77.12**	76.45
# Subject_08	95.89	**96.33**	96.01	**96.03**
# Subject_09	92.11	**92.95**	93.36	**93.77**
Average	89.01	**90.16**	91.09	**91.54**

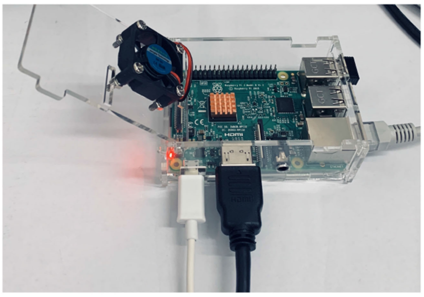
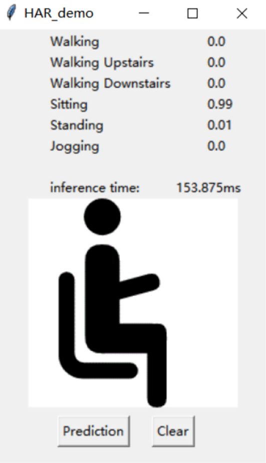


Fig. 7. Actual implementation on Raspberry Pi 3 Model B+ platform.



HAR_demo

Walking 0.0
Walking Upstairs 0.0
Walking Downstairs 0.0
Sitting 0.99
Standing 0.01
Jogging 0.0

inference time: 153.875ms

Fig. 8. The user interface of HAR application with triplet attention.

PyTorch 1.7. The Raspberry Pi is configured to communicate with a laptop computer. A Python program is developed for the HAR application (Fig. 8). For the practical implementation, a 10-second window with an 95% overlap rate is used to segment sensor readings. That is to say, the sliding step length is equal to 500 ms, and the HAR system will wait for 500 ms to read and predict next sample. We measure the inference time over 300 runs and results are shown in Fig. 9. It can be seen that the standard CNN takes around 129 ms per window, while CNN+TA takes 153.9 ms per window, which is far below 500 ms. Thus, this result is in line with our expectations and indicates that the proposed model can easily perform activity inference in a real-time way.

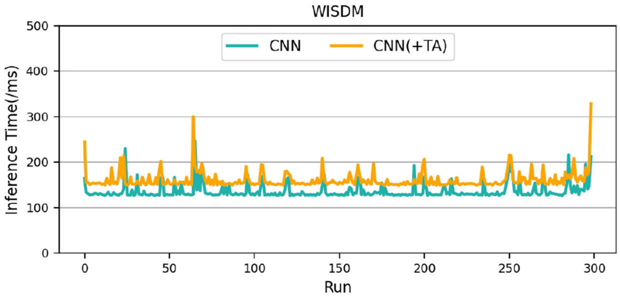


Fig. 9. Inference time of convolutional network with or without triplet attention.

Fig. 10. Snapshots of data collection in real scene.

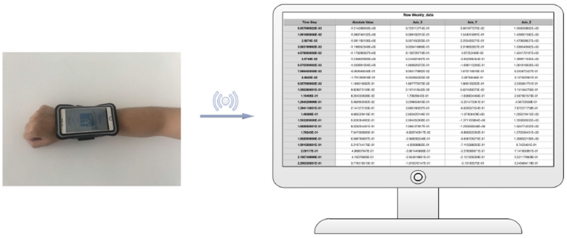


Fig. 11. The demo of collecting and processing raw data.

D. Weakly Supervised Learning

Our method is also evaluated on the weakly supervised dataset, which was collected by placing an iPhone 7 in the right pants pocket of 10 volunteers. Fig. 10 illustrates the data collection process, in which each volunteer performs 5 kinds of activities *("walking," "going upstairs," "going downstairs," "jumping" and "jogging"). "Walking"* is regarded as a background activity, which is distinguished from the rest four target activities. The sensor data is collected at a sampling frequency of 50 Hz. The application called HascLogger is used to record the three-axis accelerometer data of these activities, which produces 76,157 samples. Accordingly, the sensor data can be segmented by a sliding window of 40.96 seconds and 50% overlap. In our experiment, the ratio of training set to test set is 7:3. Fig. 11 presents the software interface where these raw data are collected and processed.

We compare the triplet attention with several state-of-the-art algorithms such as CNN, VGGNet and ResNet on the weakly labeled dataset. It can be seen that the embedded triplet attention module produces the best performance among all the algorithms

TABLE VI
THE CLASSIFICATION PERFORMANCE ON WEAKLY LABELED HAR DATASET

Model		$F_1(\%)$	Para.(M)
Standard CNN	Baseline	89.80	0.48
	+TA	**92.68**	0.48
Equally-sized VGGNet	Baseline	90.72	0.78
	+TA	**92.96**	0.78
Equally-sized ResNet	Baseline	90.93	0.79
	+TA	**93.85**	0.80
Ordóñez *et al,* 2016 [37]		90.94	-
Wang *et al,* 2019 [38]		93.55	-

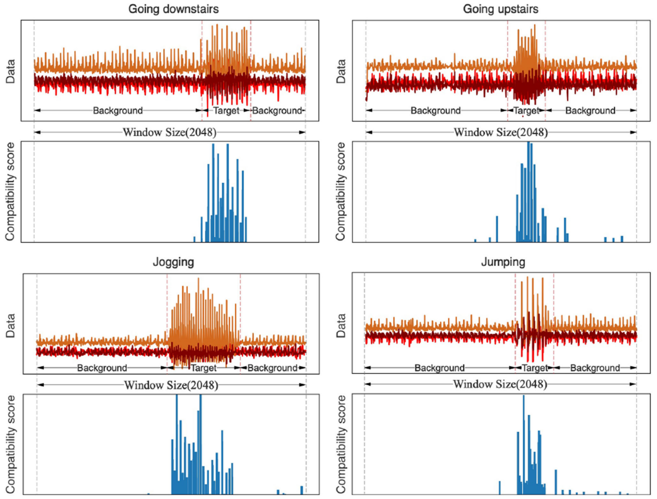


Fig. 12. Some example of location for target activity of the weakly sensor data.

in Table VI. Respectively, the proposed method achieves 2.88%, 2.24% and 2.92% performance gains over all baselines using CNN, VGGNet and ResNet as backbones. At the same time, our method is also superior to DeepConvLSTM [37] by a large margin of 2.91%. Compared with Wang *et al's* work [38], the triplet attention achieves 0.3% performance gain. The results show that cross-dimensional attention is also conducive to enhance the feature representation of weakly supervised learning.

In the final step, the visualizing analysis is provided so as to identify what part of the target signal is the most important along the temporal dimension. For the weakly labeled dataset, every signal window often contains the target activity and the background activity that submerges it, such as *"walking"*, which is different from strictly labeled HAR dataset. The four sensor signal windows, that are roughly labeled as *"jogging," "jumping," "going downstairs" and "going upstairs"*, are shown in Fig. 12. Due to the reason that our triplet attention method can focus on only the interesting part of the target activity and weaken the background activities, it will be more beneficial for ground truth data annotation.

VI. CONCLUSION

In this paper, we focus on learning cross-interaction attention for sensor based HAR task with low model complexity. A new triplet attention module is proposed, which tends to capture the cross-interaction between sensor dimension, temporal dimension, and channel dimensions via building three attention branches. Our experimental results show that the lightweight triplet attention block plays a crucial role in improving the performance of various deep CNN architectures such as the plain CNN and ResNet. Our triplet attention exhibits a good generalization ability for various sensor based HAR tasks. Several ablation experiments including visualization analysis are provided to support our conclusion, which verify the effectiveness of the triplet attention method. We hope this work could motivate future research of attention-based network architecture design in a large variety of practical HAR scenarios.

REFERENCES

[1] Z. Wang, M. Jiang, Y. Hu, and H. Li, "An incremental learning method based on probabilistic neural networks and adjustable fuzzy clustering for human activity recognition by using wearable sensors," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 4, pp. 691–699, Jul. 2012.

[2] M. A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, and H. P. Tan, "Deep activity recognition models with triaxial accelerometers," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 8–13.

[3] A. Akbari and R. Jafari, "Personalizing activity recognition models through quantifying different types of uncertainty using wearable sensors," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 9, pp. 2530–2541, Sep. 2020.

[4] Z. Wang, D. Wu, J. Chen, A. Ghoneim, and M. A. Hossain, "A triaxial accelerometer-based human activity recognition via EEMD-based features and game-theory-based feature selection," *IEEE Sensors J.*, vol. 16, no. 9, pp. 3198–3207, May 2016.

[5] Z. Chen, Q. Zhu, Y. C. Soh, and L. Zhang, "Robust human activity recognition using smartphone sensors via CT-PCA and online SVM," *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 3070–3080, Dec. 2017.

[6] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Comput. Surv.*, vol. 46, no. 3, pp. 1–33, 2014.

[7] M. Zeng *et al.*, "Convolutional neural networks for human activity recognition using mobile sensors," in *Proc. 6th Int. Conf. Mobile Comput. Appl. Serv.*, 2014, pp. 197–205.

[8] B. Meng, X. Liu, and X. Wang, "Human action recognition based on quaternion spatial-temporal convolutional neural network and LSTM in RGB videos," *Multimedia Tools Appl.*, vol. 77, no. 20, pp. 26901–26918, 2018.

[9] X. Li, Y. Wang, B. Zhang, and J. Ma, "PSDRNN: An efficient and effective HAR scheme based on feature extraction and deep learning," *IEEE Trans. Ind. Informat.*, vol. 16, no. 10, pp. 6703–6713, Oct. 2020.

[10] A. Joulin, L. Van Der Maaten, A. Jabri, and N. Vasilache, "Learning visual features from large weakly supervised data," in *Proc. Eur. Conf. Comput. Vis.*, New York, NY, USA: Springer, 2016, pp. 67–84.

[11] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[12] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "A^2-nets: Double attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 352–361.

[13] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[14] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," in *Proc. Neural Inf. Process. Syst. Time Ser. Workshop*, 2015.

[15] W. Gao, L. Zhang, Q. Teng, J. He, and H. Wu, "DanHAR: Dual attention network for multimodal human activity recognition using wearable sensors," *Appl. Soft Comput.*, vol. 111, 2021, Art. no. 107728.

[16] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proc. 21th Int. Eur. Symp. Artif. Neural Netw. Comput. Intell. Mach. Learn.*, vol. 3, pp. 437–442, 2013.

[17] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *Proc. 16th Int. Symp. Wearable Comput.*, 2012, pp. 108–109.

[18] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newslett.*, vol. 12, no. 2, pp. 74–82, 2011.

[19] D. Micucci, M. Mobilio, and P. Napoletano, "Unimib shar: A dataset for human activity recognition using acceleration data from smartphones," *Appl. Sci.*, vol. 7, no. 10, 2017, Art. no. 1101.

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[21] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[22] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019.

[23] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3139–3148.

[24] H. Ma, W. Li, X. Zhang, S. Gao, and S. Lu, "Attnsense: Multi-level attention mechanism for multimodal human activity recognition," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 3109–3115.

[25] M. Zeng *et al.*, "Understanding and improving recurrent networks for human activity recognition by continuous attention," in *Proc. ACM Int. Symp. Wearable Comput.*, 2018, pp. 56–63.

[26] J. He, Q. Zhang, L. Wang, and L. Pei, "Weakly supervised human activity recognition from wearable sensors by recurrent attention learning," *IEEE Sensors J.*, vol. 19, no. 6, pp. 2287–2297, Mar. 2019.

[27] Q. Teng, K. Wang, L. Zhang, and J. He, "The layer-wise training convolutional neural networks using local loss for sensor based human activity recognition," *IEEE Sensors J.*, vol. 20, no. 13, pp. 7265–7274, Jul. 2020.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[29] Z. N. Khan and J. Ahmad, "Attention induced multi-head convolutional neural network for human activity recognition," *Appl. Soft Comput.*, vol. 110, 2021, Art. no. 107671.

[30] A. Ignatov, "Real-time human activity recognition from accelerometer data using convolutional neural networks," *Appl. Soft Comput.*, vol. 62, pp. 915–922, 2018.

[31] Z. Xiao, X. Xu, H. Xing, F. Song, X. Wang, and B. Zhao, "A federated learning system with enhanced feature extraction for human activity recognition," *Knowl.-Based Syst.*, vol. 229, 2021, Art. no. 107338.

[32] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep learning models for real-time human activity recognition with smartphones," *Mobile Netw. Appl.*, vol. 25, no. 2, pp. 743–755, 2020.

[33] K. Walse, R. Dharaskar, and V. Thakare, "Performance evaluation of classifiers on WISDM dataset for human activity recognition," in *Proc. Second Int. Conf. Inf. Commun. Technol. Competitive Strategies*, 2016, pp. 1–7.

[34] D. Ravi, C. Wong, B. Lo, and G.-Z. Yang, "Deep learning for human activity recognition: A. resource efficient implementation on low-power devices," in *Proc. IEEE 13th Int. Conf. Wearable Implantable Body Sensor Netw.*, 2016, pp. 71–76.

[35] R. Janarthanan, S. Doss, and S. Baskar, "Optimized unsupervised deep learning assisted reconstructed coder in the on-nodule wearable sensor for human activity recognition," *Measurement*, vol. 164, 2020, Art. no. 108050.

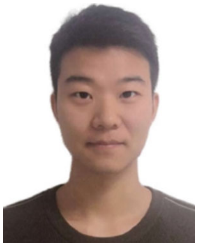
[36] T. Liu, S. Wang, Y. Liu, W. Quan, and L. Zhang, "A lightweight neural network framework using linear grouped convolution for human activity recognition on mobile devices," *J. Supercomput.*, pp. 1–21, 2021.

[37] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[38] K. Wang, J. He, and L. Zhang, "Sequential weakly labeled multiactivity localization and recognition on wearable sensors using recurrent attention networks," *IEEE Trans. Hum.-Mach. Syst.*, vol. 51, no. 4, pp. 355–364, Aug. 2021.

Yin Tang received the B.S. degree from the Hunan University of Engineering, Xiangtan, China, in 2018. He is currently working toward the M.S. degree with Nanjing Normal University, Nanjing, China. His research interests include activity recognition, computer vision, and machine learning.

Lei Zhang received the B.Sc. degree in computer science from Zhengzhou University, Zhengzhou, China, the M.S. degree in pattern recognition and intelligent system from the Chinese Academy of Sciences, Beijing, China, and the Ph.D. degree from Southeast University, Nanjing, China, in 2011. In 2008, he was a Research Fellow with IPAM, UCLA. He is currently an Associate Professor with the School of Electrical and Automation Engineering, Nanjing Normal University, Nanjing, China. His research interests include machine learning, human activity recognition, and computer vision.



Qi Teng received the B.S. degree from the Henan University of Engineering, Zhengzhou, China, in 2017. He is currently working toward the M.S. degree with Nanjing Normal University, Nanjing, China. His research interests include activity recognition, computer vision, and machine learning.

Fuhong Min received the master's degree from the School of Communication and Control Engineering, Jiangnan University, Wuxi, China, in 2003, and the Ph.D. degree from the School of Automation, Nanjing University of Science and Technology, Nanjing, China, in 2007. From 2009 to 2010, she was a Postdoctoral Fellow with the School of Mechanical Engineering, University of Southern Illinois, Carbondale, IL, USA. She is currently a Professor with the School of Electrical and Automation Engineering, Nanjing Normal University, Nanjing, China. Her research interests include circuits and signal processing.

Aiguo Song (Senior Member, IEEE) received the B.S. degree in automatic control and the M.S. degree in measurement and control from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1990 and 1993, respectively, and the Ph.D. degree in measurement and control from Southeast University, Nanjing, China, in 1998. He was an Associate Researcher with Intelligent Information Processing Laboratory, Southeast University. From 1998 to 2000, he was an Associate Professor with the Department of Instrument Science and Engineering, Southeast University. From 2000 to 2003, he was the Director of Robot Sensor and Control Laboratory, Southeast University. From April 2003 to April 2004, he was a Visiting Scientist with the Laboratory for Intelligent Mechanical Systems, Northwestern University, Evanston, IL, USA. He is currently a Professor with the School of Instrument Science and Engineering, Southeast University. His research interests include teleoperation control, haptic display, Internet telerobotics, and distributed measurement systems.