Deformable Convolutional Networks for Multimodal Human Activity Recognition using Wearable Sensors

Shige Xu, Lei Zhang, Wenbo Huang, Hao Wu and Aiguo Song, *Senior Member, IEEE*

***Abstract*—Recent years have witnessed significant success of convolutional neural networks (CNNs) in human activity recognition (HAR) using wearable sensors. Nevertheless, prior works have an obvious drawback. An activity sample may contain heterogeneous sensor modalities from different body parts. Moreover, the significance of each modality will change over time. Because a normal convolution filter usually samples activity data at a fixed regular grid, it is hard to capture salient features of activities along different sensor modalities or time intervals. What is the best filter form for activity recognition still remains a challenging task. In this paper, to resolve this issue, we present a new deformable convolutional network for recognizing human activities from intricate sensory data. Specifically, the learned offsets and the feature amplitudes are added into standard convolution, which can be modulated to allow more free form deformation over the sampling grid for sensory data. Comparing previous results, we achieve state-of-the-art recognition accuracies, *e.g.*, 82.91%, 80.02%, 97.35% and 99.21% respectively on several benchmark HAR datasets including OPPORTUNITY, UNIMIB-SHAR, USC-HAD and WISDM, hence indicating the advantage of the proposed method. The visual analysis is provided, which shows that the deformation could be conditioned on different input activity samples. The receptive field and the sampling locations can be adjusted in an adaptive manner, which leads to a better interpretability for deep model behaviors. Installing PyTorch on a Raspberry Pi 3 B plus system, we evaluate actual run time of the deformable model. The results show that the deformable filter is able to still maintain almost the same inference time, which is very beneficial for activity recognition tasks. Our work can promote further researches by leveraging an inter-modulating information to connect the deformable convolution and attention modules.**

***Index Terms*—Sensor, filter, deformable convolution, activity recognition, deep learning**

I. Introduction

DURING the past decade, there has been rapid development in sensor technology according to several key performance indicators such as better accuracy, smaller size, lower manufacturing cost [1]. These advantages enable a wide range of motion sensors, *e.g.*, accelerometer and gyroscope to be embedded into portable devices such as smart phones or watches, which could provide key enhancements for sensing capabilities in order to make these devices smarter [2]. Compared with camera-based human activity recognition (HAR) that suffers from privacy leakage, sensor technology has become popular in HAR [3]. Typically, an accelerometer tends to measure human's velocity information, while a gyroscope tends to measure rotation information. An Inertial Measurement Unit (IMU) node that forms part of an intelligent instrument may measure and report multimodal sensory information such as human body's specific force, angular velocity, and sometimes orientation. Sensor-based HAR will be convenient to our daily life in many application scenarios including sports tracking [4], healthcare [5], and smart homes [6], *etc*. For example, HAR can be extensively applied for amputee behavior analysis dedicated to enhancing the quality of life [7]. Driving behaviors can be analyzed that lead to safe travel. The daily activity recognition has played a key role in building smarter home environment. In an unobstructive way, HAR also can be used to continuously track daily activities of elderly people and alert their potential falls, which provides a safe elderly-care service. In military applications, precise activity information about soldiers together with locations and health conditions can provide a reliable evaluation for their performance and safety, which is also beneficial to assist their decision behavior in both combat or training scenarios [8]. Through analyzing input data collected from heterogeneous sensors in smart wearables attached to human body, HAR can assist the intelligent computer system to better understand human behavior, which offers a better assistance service to improve their quality of life.

A vast majority of HAR systems split sensor time series into fixed-length windows and classify each window to one activity by a machine learning (ML) algorithm [9]. Over a long time, traditional ML algorithms such as random forest, Bayesian network, and support vector machine have played a major role in addressing HAR problem [10]. However, they usually require time-consuming pre-processing steps to design handcrafted or domain-specific features. During the past decade, deep learning has become a dominant technique in ML community, which could automatically extract discriminative features to reduce the burden of handcrafted features. Especially, due to exceptional performance, convolutional neural networks (CNNs) [11][12] have recently received much attention, which can exploit the local dependency and translation invariance of

The work was supported in part by the National Science Foundation of China under Grant 61962061, the Industry-Academia Cooperation Innovation Fund Projection of Jiangsu Province under Grant BY2016001-02 and the Natural Science Foundation of Jiangsu Province under grant BK20191371. *(Corresponding author: Lei Zhang.(e-mail: leizhang@njnu.edu.cn))*

Shige Xu, Lei Zhang and Wenbo Huang are with School of Electrical and Automation Engineering, Nanjing Normal University, Nanjing, 210023, China.

Hao Wu is with the School of Information Science and Engineering, Yunnan University, China.

Aiguo Song is with the School of Instrument Science and Engineering, Southeast University.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2022.3158427, IEEE Transactions on Instrumentation and Measurement

data. These advantages make it very suitable for the inference of sensor signals.

Thus, there is a growing number of CNN frameworks for HAR, where a set of convolutional layers, subsampling layers, densely connected layers are stacked to learn hierarchical features [11]. In practice, an activity will last a time interval, and there are a few basic movements that could be involved within each activity. Every activity may be a combination of several basic movements. CNNs show a great potential to identify the salient features of activity data. To be specific, the convolutional kernels in the lower layers are in charge of extracting local features to characterize the salience of each basic movement, while the convolutional kernels in the higher layers are in charge of capturing temporal dependency at high-level abstraction to characterize the salience of a combination of several basic movements. Moreover, CNN is competent in modelling activity data by capturing the local connections in multimodal sensory data, hence leading to accurate recognition. Numerous previous survey papers summarize its main advantage and wide popularity in the HAR community [11][12]. However, there are two major drawbacks in HAR with CNN. Firstly, heterogeneous sensor modalities may have different contributions in recognizing various activities. For example, in a three-axial accelerometer (MMA7260, Freescale), the acceleration signals along x, y and z axes characterize upwards/downwards movements, lateral movements and forward/backward movements, respectively. As a result, the x-axial acceleration signals could be helpful to recognize the "jumping" activity, while the y-axial acceleration signals contribute more to distinguish the "turning-left" and "turning-right" activities. Merging unimportant sensor information without considering their difference inevitably introduces substantial noise into sensor signals. Treating them equally without distinction may undermine classification performance. Secondly, different sensor windows may correspond to activities with different scales. For example, different from "walking", the "jumping" activity only shows up in a short time interval rather than in the entire window. However, for normal convolution, the receptive fields (RFs) [13][14] of neurons in the same CNN layer have to share the same size, which limits the CNN's modelling ability. A more flexible filter form is desirable for HAR.

As the above two drawbacks have shown, CNNs are inherently limited to model such variations for activity recognition with wearable sensors. The limitation can be attributed to the fixed geometric structures of normal convolution filters: a normal filter usually samples an input feature map at a regular grid with fixed locations. In most cases, the normal filter within convolution modules is set to 3×3, 5×5 and 7×7, *etc* [13][15]. Thus, it is unrealistic for CNN to track such variations from sensor time series through normal filters. How to design a filter form that can adaptively change according to different activity samples has rarely been explored in activity recognition area. In this paper, we first present a deformable convolution module for enhancing CNNs' capability of modeling HAR with wearable sensors, which allows free deformation over sampling grid within convolution modules (Fig. 1). In comparison with

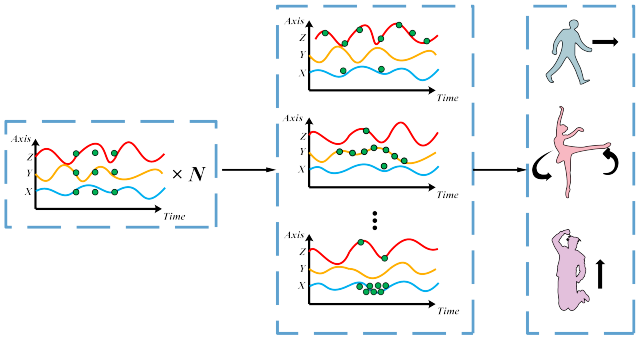

Fig. 1: Flexible filter form according to different activity characteristics

normal filters, the offsets are added to the regular sampling grid, which can be learned in an adaptive manner. In addition, the feature amplitude is modulated, which enables the network to further learn how to transform the input feature map from hybrid sensor signals. Actually, the activity windows are built using sliding window technique, and each window is constructed along the two dimensions that correspond to time-steps and heterogeneous sensor modalities respectively. Unlike image data, these basic geometric transformations will mainly involve the offsets and the modulation of feature amplitudes on different datasets, which are the model parameters to be learned during the training process. To date, there are no related works with deformable filters for activity recognition using wearable sensors, and this paper aims to fill this gap by providing a novel deformable convolution technique to tackle the above challenges. Our main contributions are summarized as follows:

Firstly, this paper introduces a new deformable convolutional neural network for activity recognition, where the offsets and feature amplitudes within standard convolution are simultaneously modulated, hence allowing a free deformation over the sampling grid in an adaptive manner.

Secondly, according to different activity samples fed into deep models, we visually show how the deformable convolution changes on its sampling grid. The receptive field and the sampling locations can be adjusted in an adaptive manner, which leads to better interpretability for recognizing human activities from intricate sensory data.

Finally, lots of experiments are conducted on four public HAR datasets consisting of OPPORTUNITY [16], UNIMIB-SHAR [17], USC-HAD [18] and WISDM [19], where performance comparison and analysis are provided. Furthermore, ablation experiments are performed to analyze the impact of several important hyper-parameters such as kernel size and modulating amplitude, which verify the effectiveness and efficiency of the proposed method. We also measure actual inference time on a Raspberry Pi platform.

The rest of this paper is organized as follows: Section II reviews the related works. Section III illustrates the structure of our proposed deformable CNN for HAR with sensor signals. Section IV presents experimental setup, results and analysis, which indicates the superiority of our proposed model. Finally, conclusions are made in Section V.

II. Related works

During the past decade, CNNs have achieved exceptional performance in visual recognition tasks, which gained a lot of attention in deep learning community. For example, Krizhevsky *et al*. [20] at the earliest time proposed AlexNet, in which a list of convolutional layers and pooling layers are stacked sequentially, yielding significant performance gain over shallow neural networks for image classification. Simonyan *et al*. [21] introduced VGGNet, in which more convolutional layers with smaller filters are stacked, in order to extract discriminative features at a larger receptive field. He *et al*. presented ResNet [22], in which an identity map is used as skip connection skipping two or three layers to generate the deeper network, enabling better gradient propagation in both forward and backward passes. From another perspective, Szegedy *et al*. introduced GoogLeNet [23], in which multiple sized filters, *e.g.*, 1×1, 3×3, and 5×5, are stacked into an Inception module, leading to an efficient deep neural network architecture for computer vision. However, these CNNs and their variants are not suitable for irregular geometric transformations, because a normal filter usually samples an input feature map at a regular grid. In visual recognition tasks, one main challenge lies in that objects may have different scales, viewpoints, and deformations. These CNNs and their variants are not suitable for irregular geometric transformations, because a normal filter usually samples an input feature map at a regular grid. In order to handle such geometric transformations, Dai *et al*. [24][25] have recently designed a novel CNN with flexible filters, which is based on the idea of changing the sampling locations in the modules by adding additional offsets and learning the offsets for the visual tasks, with no need of extra supervision. Zhu *et al*. [26] continued to reform Dai *et al*'s method by adding a learned offset as well as a learned feature amplitude, which enables the network module to implement more free-form sampling in computer vision tasks. Gao *et al*. [27] proposed an idea of Deformable Kernels, which can strengthen regular kernels with richer expressiveness by directly interacting with the effective receptive field of the computation during runtime. Kim *et al*. [28] proposed an efficient CNN architecture, which is called as deformable kernel network (DKN), where the output sets of neighbors and the corresponding weights could be adaptively tuned for each pixel. A weighted average can be computed as the final filtering result for joint image filtering. However, prior to these works mainly focusing on computer vision tasks, the deformable module has been rarely explored in HAR scenarios. At present, it still lacks internal mechanisms to model flexible transformations for complex sensor signals, which are hard to be convolved at a regular sampling grid. As a result, there is a noticeable shortcoming for normal convolution in HAR scenarios.

Recently, there has been another increasing line of research in HAR using wearable sensors, which aims to alleviate the burden of handcrafted features requiring expert knowledge. These researches have employed CNNs to automatically extract features. For example, Zeng *et al*. [29] at the earliest time proposed a shallow CNN in HAR scenarios, in which the input is restricted to accelerometer data. According to the hypothesis that heterogeneous sensor modalities, *e.g.*, accelerometer and gyroscope need to be convolved separately, Yang *et al*. [30] used CNNs with multiple iterations of convolutional and pooling layers combined for feature extraction. When the input is taken from multimodal sensors, Chen *et al*. [31] presented a sophisticated CNN, which includes three convolutional layers with 18, 36, 24 filters, and each is followed by 2×1 max-pooling layers respectively. In order to extract the association between two adjoining pairs of sensor axes, they adopted a 12×2 filter at the first layer. The 12×1 filter is applied in the remaining layers, which is able to capture only the temporal association. Similarly, in order to improve the representation ability of sensor signals before feeding them to CNN, a new signal, referred to as signal image, is generated through applying a Fourier transform on a set of permutations using the sensor axes of the raw signal. Jiang *et al*. [32] further proposed a CNN with two layers, in which they adopt filters of 5×5 followed by 4×4 and 2×2 average-pooling layers respectively. Despite exceptional results, the method proposed by Jiang *et al*. is computationally expensive, which prevents its wide use. To tackle this problem, Ha *et al*. [33] presented an architecture consisting of two convolutional layers with 32 and 64 filters of 3×3 respectively, where a zero-padding technique is adopted after the first convolutional layer to keep different sensor modalities separated. Luo et al. [34] introduce a binarized convolutional network for real-time HAR, in which the dilated convolution is used to enlarge the receptive field and improve its potential capturing capability for time series. This work will effectively reduce latency in resource-constrained mobile devices, which may better support computation-intensive deep models in ubiquitous HAR scenarios. Ma et al. [35] propose a new deep model called AttnSense for multimodal HAR tasks, which combines attention module with a CNN and a Gated Recurrent Units (GRU) network to highlight more important sensor modalities or time intervals. The attention-based deep model can improve the interpretability of deep model behaviors. Ordóñez et al. [36] introduce a novel hybrid network architecture called DeepConvLSTM that inserts LSTM layers into normal convolutional layers, which demonstrates obvious advantages in feature extraction from raw sensory data. Ignatov et al. [37] propose to use CNNs for automatic feature extraction together with handcrafted activity features, which is able better capture contextual information about the global form of sensory data. The proposed solution is evaluated on mobile devices to ensure satisfactory inference time. There is only the fixed geometric structure within standard convolution filters, which limits conventional CNN's modelling capability of irregular geometric transformations for activity recognition tasks. What is the best filter form in HAR scenarios still remains open. In this paper, we first fill the gap via introducing a new deformable convolution technique to address the above challenge.

III. Model

A. Deformable Convolution

In HAR scenarios, it is a critical and challenging task to extract activity features from raw sensor readings. At the first

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2022.3158427, IEEE Transactions on Instrumentation and Measurement

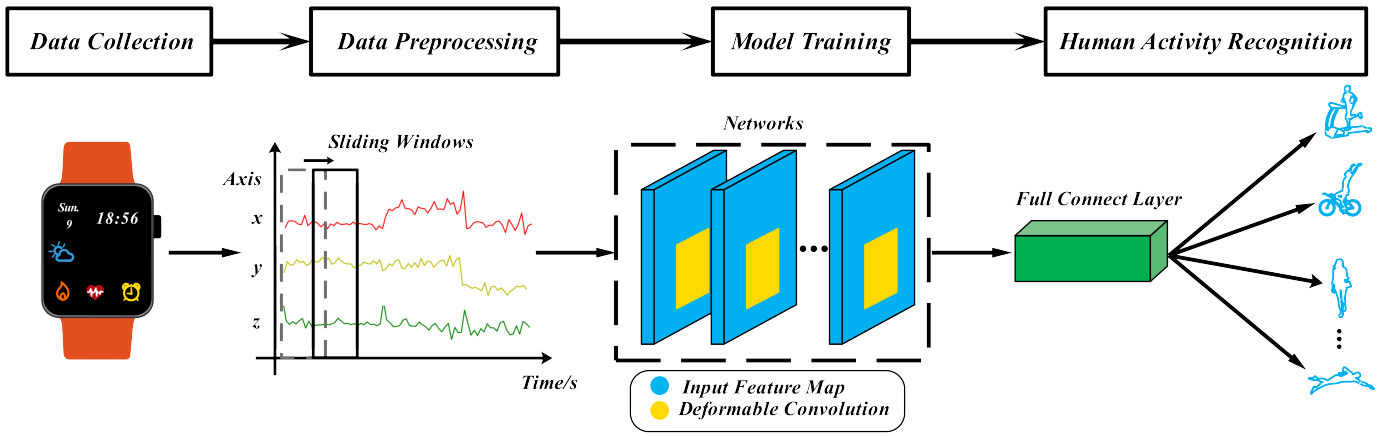

Fig. 2: The overview of deformable convolution (DFC) model on sensor-based Human Activity Recognition

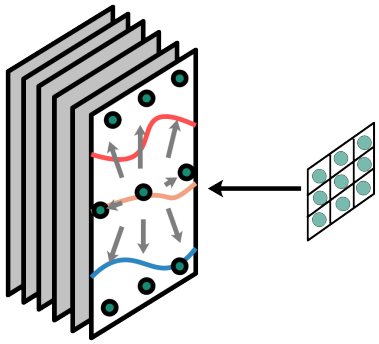

Fig. 3: The sechematic diagram of the deformable convolution (DFC)

stage of a standard activity chain, raw sensor data is collected from one or multiple sensors that are attached to different parts of human body. It is well known that one sensor may produce multiple values (*e.g.*, a triaxial accelerometer generates 3-dimensional acceleration signals typically corresponding to x, y, and z directions), or multiple sensors are jointly used. Therefore, a mathematical notation of raw sensor input D can be formulated as [11][12]:

$$D = \begin{pmatrix} d_1^1 & \cdots & d_1^t \\ \vdots & \ddots & \vdots \\ d_s^1 & \cdots & d_s^t \end{pmatrix} \tag{1}$$

in which s denotes the number of sensor modalities and d^t denotes the corresponding vector value at a time t. Each sensor modality will be sampled at a constant frequency, which produces a multivariate time series. Thus, sensor-based HAR can be treated as a multivariate time series classification problem [12]. Traditional signal processing techniques such as time-frequency transformations or other statistical approaches can be used to generate handcrafted activity features for shallow ML classification algorithms. However, the features extraction procedure is heuristic and needs to be carefully engineered. They are not common or universal to capture discriminative features, which often work well in one specific activity recognition task, but badly in the other tasks. Most prior works [11][38] depend on heuristic hand-crafted features and shallow learning architectures, which is not very effective to find those discriminative features to accurately identify similar activities. We propose a systematic deep learning approach for activity recognition problems, which adopts a deformable convolution network to automatically learn features from raw sensor inputs in a flexible way. Based on the deformable convolution architecture, the automated features can be seen as the high-level abstracted representation from low-level raw sensor signals. In this section, we introduce the structure of deformable convolution, where the filter form is conditioned on a particular input, which then produces corresponding output feature map. Because of the integrated offsets and the feature amplitudes whose parameters can be easily trained end-to-end with standard backpropagation, the proposed deformable convolution is a differentiable module which applies a deformable filter to convolve an input feature map during both forward and backward passes. A vast majority of HAR methods with CNN often focus on convolution modules with regular sampling grid, which need to be designed carefully to avoid the case that the heterogeneous sensor modalities are convolved together [29][30][33]. As a comparison, one main advantage of our deformable convolution lies in that it can adaptively change filter form, which is more beneficial for HAR using wearable sensors.

As illustrated in Fig. 2 and Fig. 3, the normal convolution is composed of two main parts: 1) R: a regular grid used to sample over an input feature map; 2) summation: it is used to sum up the sampled values weighted by w [24][30][31]. Without loss of generality, the output feature map y can be formulated as

$$\begin{aligned} y(p_0) &= \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \\ R &= \{(-1,-1),(-1,0),\cdots,(0,1),(1,1)\} \end{aligned} \tag{2}$$

where p_0 is the location on output while p_n enumerates locations in regular sampling grid R. In order of computation, the deformable filter first regularly convolves on an input feature map, and then learns the offsets that should be applied to the feature map, giving a transformation of R conditional on the input. To be specific, we use offsets Δp to augment the above R, in which Δp is a real number with the unconstrained range. The output y in equation (2) is updated by Δp. We

redefine the output y by equation (3):

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \quad \{\Delta p_n | n = 1, ..., N\}, N = |R|. \tag{3}$$

This grid generator makes R samples on $p_0 + p_n$ irregularly. In essence, the predicted offsets are used to generate an irregular sampling grid, which consists of a set of locations where the input map should be sampled in a flexible form. Since Δp_n is a typical fraction, we further use bilinear interpolation to express equation (3) as follows [24][25][39]:

$$x(p_0) = \sum_q G(q, p) \cdot x(q). \tag{4}$$

In equation (4), p represents an arbitrary fractional location and q enumerates the input 'x' integral location. The bilinear interpolation kernel is denoted by G, which has two dimensions:

$$G(p, q) = g(p_x, q_x) \cdot g(p_y, q_y) \quad g(a, b) = max(0, 1 - |a - b|). \tag{5}$$

In equation (5), there are merely a small number of non-zero q in $G(p,q)$ [39][24]. Thus, equation (4) can be easily computed. Finally, the combination of these parts forms a deformable convolution.

B. Modulated Deformable Convolution

We adopt a modulation technique to enhance the feature extraction capability of deformable convolution across different activity recognition datasets. Due to the use of modulation, the deformable convolution can adjust the offsets, as well as the feature amplitudes simultaneously at different locations. To be specific, as indicated above, the receptive field size can be defined as a sampling grid R. In this paper, without loss of generality, a 3×3 filter is used, which can be expressed as:

$$R = \{(-1, -1), (-1, 0), \cdots, (0, 1), (1, 1)\}. \tag{6}$$

As a result, the modulated deformable filters are formulated as:

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \cdot \Delta \mathrm{m}_n \quad \{\Delta \mathrm{m}_n \in [0, 1],\ \Delta m_n | n = 1, ..., N\}, N = |R| \tag{7}$$

where p_0 is the sampled location on the output feature map y while p_n enumerates locations in regular sampling grid R. Δm is a modulation scalar on the n-th location, which is generated by a Sigmoid activation function. As illustrated in equation (7), its value lies in the interval [0, 1]. Actually, the content information from some locations may have no or less contribution to output. If Δm_n is identical to zero, the value sampled by the modulated deformable filter at n-th location will be zero. That is to say, in an extreme case, the feature amplitude even could be set to zero in order to avoid sampling over such locations, which enables one more freedom to adjust the support regions of deformable convolution. Similarly, equation (7) still utilizes bilinear interpolation to calculate $x(p_0 + p_n + \Delta p_n)$, since $p_0 + p_n + \Delta p_n$ is fractional. Referring to Dai *et al*. 2017 and Jaderberg *et al*. 2015 [25][26][39], we still split the two-dimensional bilinear kernel into two one-dimensional ones, which is formulated as:

$$G(p, q) = g(p_x, q_x) \cdot g(p_y, q_y) \tag{8}$$

where $g(a, b) = max(0, 1 - |a - b|)$.

Overall, for a given input feature map x, the offset and the feature amplitude need to be learned through a separate convolutional layer, which produces the output y with $3N$ channels. The first $2N$ channels are in charge of generating the learned offsets, while the remaining N channels are fed into a Sigmoid layer to produce the modulation scalars. At the training stage, the convolution kernel, the offset and the feature amplitude can be learned simultaneously.

IV. Experiment

The whole experiment is composed of four parts. In part one, we select four publicly available HAR datasets consisting of OPPORTUNITY [16], UNIMIB-SHAR [17], USC-HAD [18] and WISDM [19] for our evaluation. Performance analysis and comparison are presented. In part two, ablation studies are performed to evaluate the influence of several key factors such as the kernel size and the feature amplitude. In part three, we provide visualization analysis to show the changing behavior of deformable convolution through various human activities. In the final part, the actual operation of the deformable model in real-time systems is evaluated, in which the embedded system used in this experiment is Raspberry Pi system. In what follows, we will first detail the datasets used and the experimental setup.

A. Experimental comparison and analysis

1) Datasets: In order to show the effectiveness of deformable convolution, we select four publicly available datasets in our evaluation. Data preprocessing is a crucial step in this activity recognition process. All datasets are involved in human activities in different contexts, which have been recorded via various heterogeneous sensors. Sensor signals are often involved in various human activities in different contexts, which have been recorded via hybrid sensor modalities. Due to various reasons such as sensor malfunction, the activity data collected from sensors inevitably contains noisy data. In numerous prior studies, a low-pass filter is always used to perform additional noise filtering, which can be seen as a smoothing operation for data in the temporal domain [40]. Specifically, as shown in Fig. 4, raw acceleration signals are filtered using a 3rd order Butterworth low-pass filter with a cutoff frequency of 20 Hz to remove noise. This frequency is enough to capture human body motion because 99% of its energy is constrained below 15 Hz [12][41]. Assuming that the gravitational force has only low frequency components, we further separate the acceleration signals into body and gravity acceleration signals by using another Butterworth low-pass filter with a corner frequency of 0.3 Hz. As a result, the produced values will reflect more actual activity effect. By subtracting the mean and dividing by the standard deviation, the heterogeneous sensor values are normalized into zero mean and unit variance. Since standard classification algorithms

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2022.3158427, IEEE Transactions on Instrumentation and Measurement

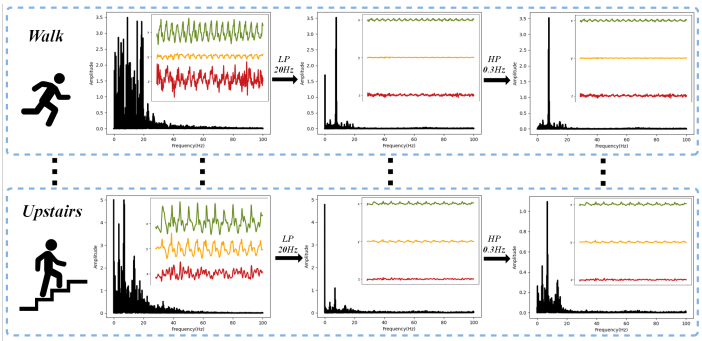

Fig. 4: Unfiltered and filtered signal using 3th order Butterworth filter

TABLE I: The setting of data-preprocessing and experiment setup

	OPPO[16]	UNIMIB[17]	USC[18]	WISDM[19]
Sample Rate (Hz)	30	50	100	20
Subjects	12	30	14	29
Categories	17	17	12	6
Samples	30656	11771	10964	10981
Window Size	64×1	151×1	512×1	200×1
Stride	32	76	256	20
Overlap Rate(%)	50	50	50	90
Z-score Standardization	True	True	True	True
Average-Pooling	4×3	3×1	8×1	4×1
Initial Learning Rate	1×10^{-3}	5×10^{-4}	5×10^{-4}	1×10^{-4}
Decay Period	40	30	30	40
Batch Size	128	256	128	64
Epochs	150	150	150	150

could not be directly adopted to handle raw sensor readings, it needs to be first divided into activity samples by the sliding window technique. To be specific, at a fixed overlap rate, one can slide a fixed-length window over continuous sensor reading to generate continuous samples, where each window may be assigned a specific activity label. As a consequence, the sensor time series is divided into consecutive windows of a fixed size and without inter-window gaps, and an overlap between adjoining windows is tolerated to preserve the continuity of sensor signals. We present the detailed parameter settings of these datasets, such as sampling frequency, window size and overlap rate in Table I. Actually, the window size has an important effect on activity recognition performance. According to prior works [42], reducing the window length will be suitable for faster activity recognition, which also leads to smaller computational costs and energy consumption. Instead, increasing window length will be useful for the recognition of complex activities that last a longer time. So far, there is still no clear consensus on how to select an optimal window size. For fair comparisons, we follow the same parameter settings as adopted in [16] [17] [18] and [19]. We add the detailed statistic of these used datasets in Table I. Moreover, more descriptions about these datasets are provided as follows:

OPPORTUNITY dataset [16]: The dataset called OPPORTUNITY is designed by a European research team, which basically covers 17 kinds of common morning activities in a breakfast scenario. Specifically, 12 volunteer subjects are asked to perform a set of daily morning activities, *e.g.*, "preparing and drinking coffee", "making and eating sandwich", and "cleaning dining table" under a smart home environment. The dataset is collected from a lot of hybrid sensing modalities consisting of accelerometers, gyroscopes, magnetometers, and video cameras, which are integrated into the environment, as well as on human bodies. There are 72 heterogeneous sensors integrated in 15 sensor system networks. This dataset provides a rich playground to evaluate algorithms such as supervised classification, multimodal sensor fusion, sensor network research. In this paper, we select the subset from the OPPORTUNITY challenge, which includes unsegmented sensor recordings from 4 subjects with only on-body sensors. Data is collected at a fixed sampling rate of 30Hz, where each subject performs these asked activities for 5 different runs. The length of sliding window is equal to 1s and the overlap rate is set to 50%, which results in overall 650k samples.

UNIMIB-SHAR dataset [17]:The dataset is built by Daniela *et al.* in University of Milano-Bicocca. The researchers recorded sensor signals through a triaxial accelerometer embedded in a smartphone Samsung Galaxy Nexus I9250, in which the installed Android OS version is 5.1.1, allowing accelerations from $\pm$2g to $\pm$16g. All samples are involved in 17 fine-grained activities consisting of both 9 types of activities of daily living (ADL), *i.e.*, "StandingUpFL", "LyingDownFS", "StandingUpFS", "Running", "SittingDown", "GoingDownS", "GoingUpS", "Walking" and "Jumping", and 8 types of falls, *i.e.*, "Falling-BackSC", " FallingBack", "FallingWithPS", "FallingForw", " FallingLeft", "FallingRight", " HittingObstacle" and "Syncope". All actions are performed by 30 participants whose ages range between 18 and 60 years. The whole dataset was designed for monitoring human activity, especially for detecting the type of falls such as "sideward", "forward", "backward", and "syncope". The window length and overlap rate are set to around 3 seconds and 50% respectively. Data is sampled at a frequency of 50HZ, which provides 11771 acceleration samples.

USC-HAD dataset [18]: The dataset called as University of Southern California Human Activity Dataset (USC-HAD) is collected by 14 participants (7 male and 7 female) aged from 21-49, whose height (cm) and weight (kg) ranged from 160-185 and 43-80 respectively. The sensing devices that contain a MotionNode were fixed to the participants' front right hip and each participant performed 12 well-defined low-level daily activities, *i.e.*, "walking forward", "walking left", "walking right", "walking upstairs", "walking downstairs", "running forward", "jumping", "sitting", "standing", "sleeping", "elevating up" and "elevating down". The dataset is designed as a benchmark for various algorithm comparisons, especially for healthcare scenarios including elder care and health monitoring. Each participant performs 5 runs for each activity. The sampling frequency of sensor signals is 100Hz. The length of the sliding window is equal to 5.12 seconds and the overlap rate is set to 50%, which results in overall 10964 samples.

WISDM dataset [19]: The dataset is collected by the Wireless Sensor Data Mining (WISDM) laboratory in a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2022.3158427, IEEE Transactions on Instrumentation and Measurement

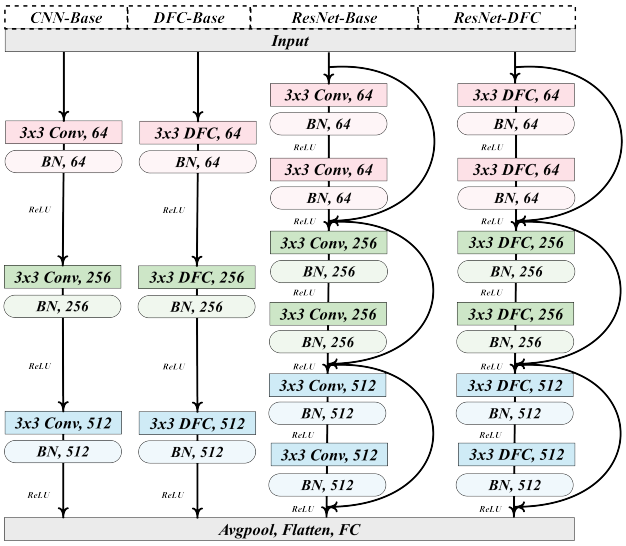

Fig. 5: Brief description for each backbone

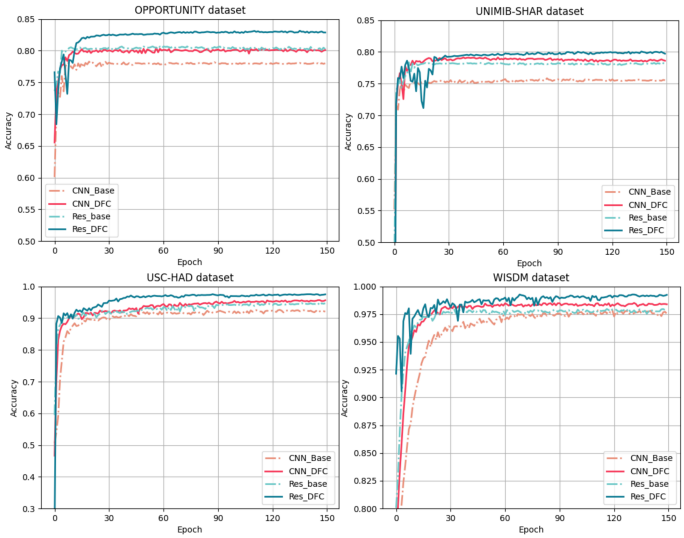

Fig. 6: Performance comparisons on four public datasets

strictly controlled scenario. They used an Android smartphone equipped with a three-axial accelerometer to collect sensor signals. 29 volunteer subjects placed the smartphone in their front leg trouser pockets, and each volunteer subject was asked to perform 6 types of low-level daily activities including "walking", "jogging", "upstairs", "downstairs", "sitting" and "standing", which contribute to 38.6%, 31.2%, 11.2%, 9.1%, 5.5% and 4.4% samples respectively. The data is composed of triaxial (*i.e.*, x, y, z) accelerometer signals collected at a sampling frequency of 20 Hz, where the x, y, and z axes reflect longitudinal, lateral, and forward activity signals respectively. The length of sliding window is equal to 10 seconds and the overlap rate is set to 90%. As a result, the whole WISDM dataset includes 10,981 samples.

2) Experimental setup: In order to evaluate the effectiveness of the proposed deformable convolution against existing approaches, four datasets are analyzed. Table I summarizes several important properties of these datasets during the data preprocessing stage. The datasets are partitioned into three parts: a training set (70%), a validation set (10%) and a test set (20%).The validation set is used to tune the hyperparameters such as kernel size. To compare the relative performance gain caused by the proposed solution, we have utilized two different baseline configurations, *i.e.*, CNN and ResNet, in which C(L_s) means the layer has L_s feature maps. The baseline CNN contains a set of stacked convolutional layers, pooling layers, as well as a softmax layer for final classification probability. To demonstrate the generality ability of deformable convolution, we also use an equally-sized ResNet as our baseline. As shown in Fig. 5, a flowchart is drawn to illustrate the complete network architecture. The details of our network setting are presented in Table I. Among these network structures, the kernel size and corresponding stride within convolution modules are set to 3×3 and 2 respectively. The dynamic learning scheme is employed, in which the learning rate will decay by half after a specific number of epochs. The exact values are shown in Table I as well. Batch normalization is used to help the converge of training process. As shown in Fig. 6, the number of the training epochs is constantly set to 150

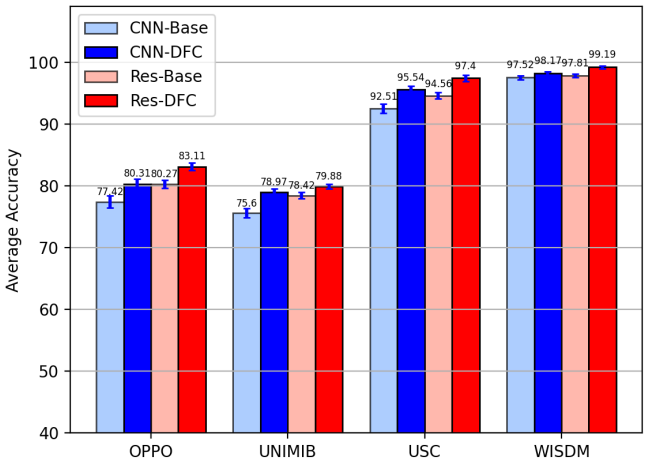

Fig. 7: Average results of five-fold cross-validation

during the whole experiment. An Adam optimizer is chosen to update parameters for the optimization of cross-entropy loss function. In particular, the offset Δp are initially set to zero. In other words, we transform the filter form at the beginning from a normal convolutional filter without offset. The modulation of the feature amplitude Δm is initially set to zero. All the experiments are implemented in Pytorch deep learning library on a server with 4 NVIDIA GeForce RTX 3090 GPUs.

3) Performance comparison and analysis: We also evaluate relative performance gain over both baselines. As shown in Fig. 6, the deformable convolution significantly contributes to the performance gain, when compared with the two baselines across four public datasets. From results in Table II, what could be observed is that on OPPORTUNITY dataset [16], the normal baseline CNN has a classification accuracy of 77.61%, while the deformable CNN acquires a better result with 80.22%, which almost reaches the recognition level of the baseline ResNet with much smaller parameters. The deformable ResNet outperforms its corresponding counterpart

TABLE II: Parameters(M)&Flops(G)&Accuracy(%)

	OPPORTUNITY	UNIMIB-SHAR	USC-HAD	WISDM
CNN-Base	0.192 & 0.012 & 77.61	1.58 & 0.02 & 75.51	1.58 & 0.076 & 92.33	1.56 & 0.027 & 97.61
CNN-DFC	0.193 & 0.012 & **80.22**	1.67 & 0.022 & **79.11**	1.68 & 0.084 & **95.36**	1.67 & 0.029 & **98.30**
ResNet-Base	3.47 & 0.219 & 80.38	6.23 & 0.081 & 78.14	6.23 & 0.303 & 94.42	6.21 & 0.106 & 97.85
ResNet-DFC	3.68 & 0.243 & **82.91**	6.67 & 0.088 & **80.02**	6.66 & 0.335 & **97.35**	6.64 & 0.115 & **99.21**
Deep Learning's Acc (%)	78.90 ([36] 2016)	77.03 ([43] 2018)	97.01 ([32] 2015)	98.81 ([44] 2020)
	74.50 ([45] 2016)	76.67 ([46] 2021)	91.70 ([47] 2020)	98.70 ([48] 2020)
	76.83 ([29] 2014)	75.65 ([49] 2019)	94.06 ([50] 2020)	98.97 ([51] 2021)
	-	72.80 ([52] 2021)	-	93.32 ([37] 2018)
Shallow Learning's Acc (%)	69.94 ([53] 2020)	68.00 ([54] 2021)	86.48 ([55] 2018)	92.98 ([56] 2015)
	68.00 ([57] 2020)	65.74 ([58] 2019)	78.47 ([56] 2015)	93.50 ([59] 2018)
	66.80 ([60] 2016)	-	-	93.95 ([61] 2019)

by an accuracy improvement of 2.53%, with almost no extra cost such as memory and computational overhead. Results from UNIMIB-SHAR dataset [17], we could observe that a recognition rate below 80% is obtained by the normal CNN, which is much lower than those on other public datasets. Nevertheless, the deformable module produces a significant increase in classification performance when their backbone architectures are the same. Specially, there is an increase of 3.6% and 1.88% respectively over the baseline CNN and ResNet models, which is accompanied by an increase with 0.09M and 0.44M in parameter number. It is worthwhile to mention that deformable CNN even performs better than the corresponding baseline ResNet, which yields a 0.97% performance gain with only one-fourth of parameters. According to the results on USC-HAD dataset [18], the deformable module leads to state-of-the-art recognition effect reaching an accuracy of 95.36% when integrated into convolutional architectures, which is around 3.03% higher than the corresponding baseline. Compared with the baseline ResNet with classification accuracy 94.42%, the deformable module refreshes the top result to 97.35%, which causes an increase with only 0.032G FLOPs at computational cost. Finally, the performance comparisons are conducted on WISDM dataset [19]. As far as we have known, this dataset can be easily classified via by simple forward-feedback neural networks. Therefore, results from Table II, the baseline CNN can achieve a pretty high accuracy, *i.e.*, 97.61%, which may be further increased to 97.85% in the case of ResNet. Despite so, the deformable solution is still able to refresh the score to 98.3% and 99.21% respectively, providing further performance improvement over both baselines. In order to further verify the superiority of the proposed deformable convolution, we perform a five-fold cross-validation on the four HAR benchmarks. During the five-fold cross-validation, the original datasets are randomly partitioned into 5 folds, in which each fold is held out in turn and the training is performed on the rest four-fifths. Thus, the learning procedure is executed overall 5 times on different training sets. As a result, the final accuracy can be estimated by averaging the obtained 5 accuracies. Fig. 7 reports the mean accuracy and standard deviation of both baselines *i.e.*, CNN and ResNet evaluated on different datasets. It can be seen that our method is able to reliably improve both baselines due to deformable convolution kernels.

TABLE III: Performance comparisons of different kernel sizes on OPPORTUNITY dataset

Kernel Size	1×1	3×3	5×5	7×7	9×9
Par(M)	0.2	1.69	4.63	9.06	14.95
MemR+W(M)	0.45	7.37	19.11	36.73	60.23
MAdd(M)	4.71	44.42	123.12	241.13	398.49
Flops(M)	2.38	22.24	61.59	120.62	199.33
Time/Epoch(s)	7.33	**3.94**	4.72	7.26	13.04
Avgacc(%)	58.25	80.22	80.31	80.48	80.54

During recent years, some approaches that use deep neural networks have been extensively investigated in the HAR scenario. As far as we have known, many researchers such as Zeng *et al*. [29], Yang *et al*. [30], and Ordóñez *et al*. [36] have tried to adopt deep neural networks to re-attack activity recognition challenges. We compare our results with those of state-of-the-art approaches. Results from Table II, it can be clearly seen that the deep learning methods are significantly superior to shallow ML algorithms that use handcrafted features, which shows the necessity of deep learning in activity recognition tasks. For OPPORTUNITY dataset, the proposed deformable model leads to a significant increase of 6.08% when compared with the convolutional architecture by Zeng *et al*. [29], as well as a 3.01% accuracy improvement compared with DeepConvLSTM by Ordóñez *et al*. [36]. For UNIMIB-SHAR dataset, we could observe a clear 5% gap between the best (proposed deformable model) and worst (CNN-Base), and the deformable convolution produces a performance gain of 2.99% and 3.35% respectively over Li *et al*'s [43] hybrid deep-learning architecture and Liu *et al*'s [46] linear grouped convolution. On USC-HAD dataset, the deformable method surpasses Singh *et al*'s [50] result using self-attention method and Bi *et al*'s [47] consequences with dynamic active learning by 3.29% and 5.65% respectively. In the case of WISDM dataset, the deformable model achieves a 0.24% performance improvement over Xiao *et al*'s [51] federated learning method and is also superior to the other two methods by 5.89% and 0.51% respectively (Ignatov *et al*. [37] and Noori *et al*. [48]), which demonstrates a good generalization ability.

B. Ablation experiments

1) Kernel size selection: This part aims to better understand how kernel size (*i.e.*, N) affects classification performance by the proposed deformable convolution in HAR. To investigate

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2022.3158427, IEEE Transactions on Instrumentation and Measurement

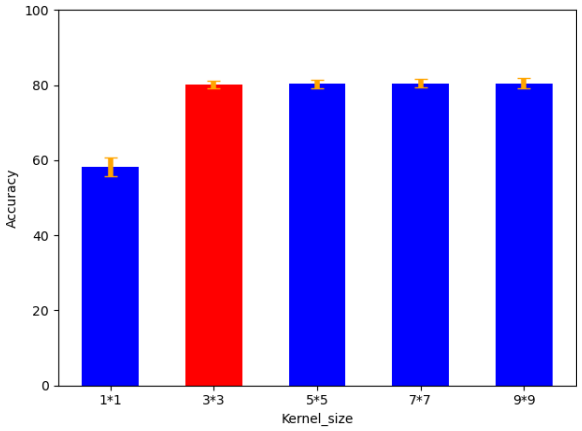

Fig. 8: Kernel size Comparisons on OPPORTUNITY dataset

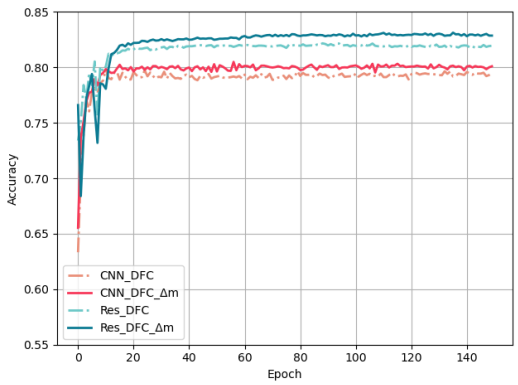

Fig. 9: Performance Comparisons on OPPORTUNITY dataset with modulation Δm

what are the best kernel size on OPPORTUNITY dataset [16], the average test accuracy is illustrated in Fig. 8 with respect to different kernel sizes within it. Taking into account three important factors such as accuracy, parameter number, and FLOPs, we summarize the comparison results in Table III. It can be seen that the deformable convolution with 3×3 kernels can strike a better tradeoff between classification performance and resource consumption. For instance, achieving 80.22% average accuracy calculated after 5 runs, 3×3 kernels need much fewer parameters and FLOPs than larger kernels with 5×5, 7×7 and 9×9. When using larger kernels, the accuracy only increases slightly, but inevitably leads to a rapid increase in memory and computation burden. The observation is in line with the basic design principle of CNN, in which smaller kernels usually show better performance than larger ones [21].

2) The impact of modulating the feature amplitude: The task of this part is to analyze the effect of Δm on performance improvement. We conduct this ablation study with two baseline network structures on OPPORTUNITY dataset [16]. Fig. 9 displays test accuracy curves with both baselines without modulation mechanism. In addition, we perform performance comparison via adding modulation on the feature amplitude. It can be observed that adding Δm can produce a significant performance gain over two baselines.

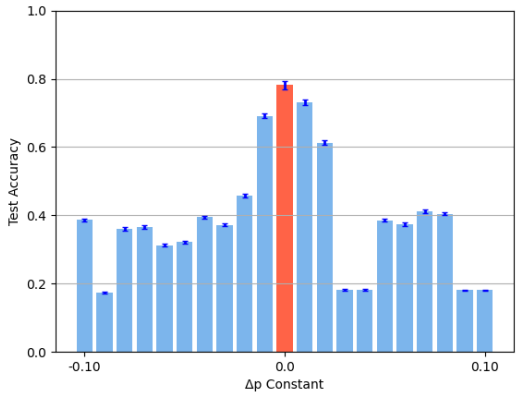

Fig. 10: Performance comparisons when initializing Δp in the range of -0.1 to 0.1 (interval 0.01)

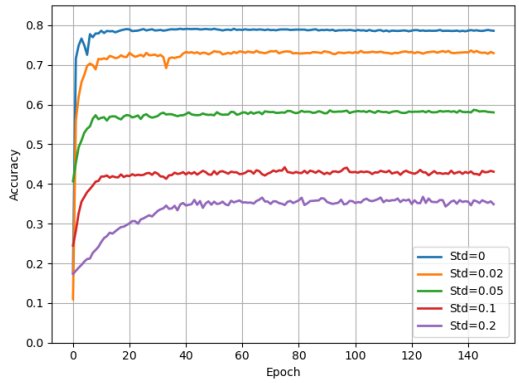

Fig. 11: Performance comparisons when initializing Δp by using random Gaussian noise with different variances

As indicated above, the primary intention of deformation is to change sampling locations of the input feature map. Modulating the feature amplitude, we can further use bilinear interpolation method to obtain a brand-new value for each feature map. The modulation parameter Δm generated by a Sigmoid function can produce one more freedom to change filter structure, which enables the convolutional filter to evolve in a more flexible form to infer human activities.

3) The initial values of Δp and Δm: In this part, we investigate how to set the initial value of Δp during the training process on UNIMIB-SHAR dataset [17]. We evaluate the impact of the initial value by increasing Δp from -0.1 to 0.1 with a fixed step length of 0.01. Fig. 10 shows that the classification performance first increases and then decreases, which attain a maximum value when Δp is set to zero. Generally speaking, due to the use of the stochastic optimization algorithm, *i.e.*, stochastic gradient descent, the weights of artificial neural networks are usually initialized to small random numbers. However, as indicated above, the value of Δp is initially set to zero. In order to show its advantage, we perform comparison experiments. From results in Fig. 11, it can be seen that the initial value zero is obviously superior to small random numbers generated by Gaussian noise with different standard variance. That is to say, the deformable convolution initially starts to transform from a regular sampling grid. We next evaluate the impact of the initial value of Δm. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2022.3158427, IEEE Transactions on Instrumentation and Measurement

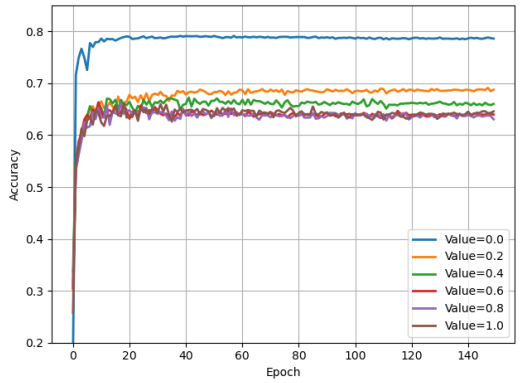

Fig. 12: Performance comparisons on DFC model with different initial values of Δm

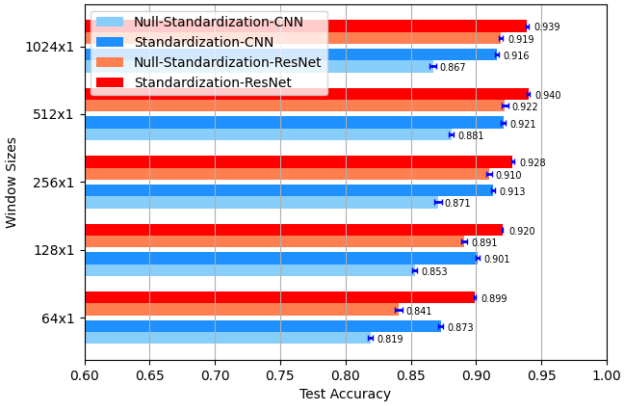

Fig. 13: Evaluation of pre-processing on recognition effectiveness

modulation Δm is distributed over the range [0, 1], which are scalars generated by a Sigmoid layer. In other words, the modulated amplitude Δm are learned parameters. Without loss of generality, we plot the test accuracy curves with different initial values, *i.e.*, Δm=0, 0.2, 0.4, 0.6, 0.8 and 1. As shown in Fig. 12, the classification accuracy rapidly decays as the initial value of Δm increases. In order to modulate the feature amplitude with Δm, we can acquire satisfactory results via setting its initial value to zero.

4) Data-preprocess evaluation: In order to evaluate the impact of data preprocessing on the obtained results, we perform the ablation experiment on USC-HAD dataset [18] with different window lengths, *e.g.*, 64, 128, 256, 512, 1024. Results are shown in Fig. 13. Fixing the overlap rate to 50%, it can be seen that the test accuracies evolve non-monotonically as the window length increases, which attains a peak value at 512. We could obtain an optimal test accuracy of 92.2% ($\pm$0.25%) for CNN and 94.0% ($\pm$0.09%) for ResNet respectively. If the window length is set to 1024, there is a slight drop in test accuracy. Overall, when the window length is equal to 512, the deformable convolution can strike a better trade-off between classification performance and resource consumption, *i.e.*, inference speed and memory cost. On the other hand, it can be clearly seen that the test accuracies with Z-score standardization are significantly superior to those without Z-score standardization, which indicates that data standardization that normalizes heterogeneous sensors signals plays an indispensable role in HAR.

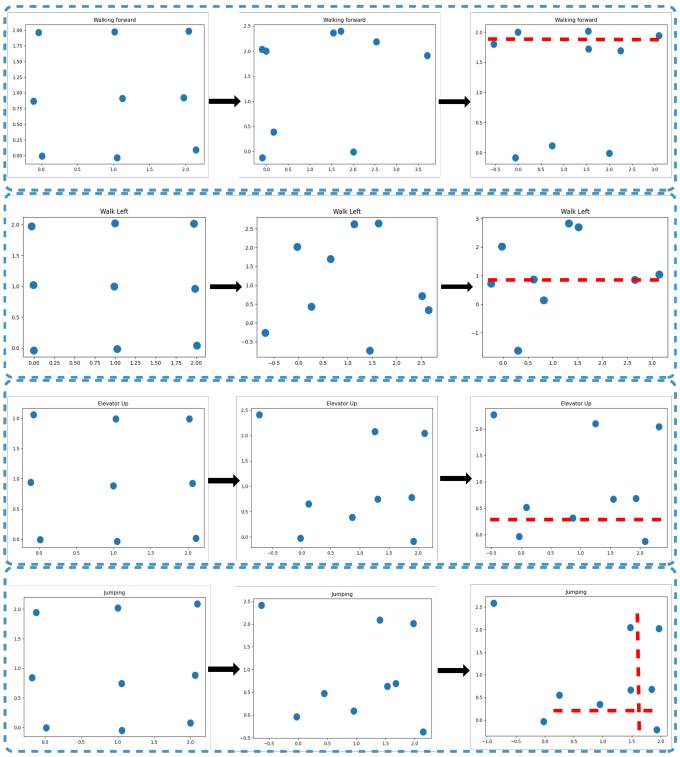

Fig. 14: The evolving process of a 3×3 sampling grid about four actions ("walking", "walking left", "elevator up" and "jumping") on USC-HAD dataset, where x and y axes correspond to time and sensor (acc-x, acc-y and acc-z from bottom to top) dimension in each subplot respectively

C. Visualization

1) Visualization of the offsets: Actually, these filters do change their filter shape rather than filter size while deforming. The filter size is an important hyperparameter, which has a direct fluence on the number of weight parameters in ConvNets. Large filters have been rarely used in order to keep this number low and avoid overfitting. Smaller filters should be preferably selected. The ConvNets tend to first adopt small filters and then implicitly increase the receptive field size by gradually increasing network depth and meanwhile reducing resolution via pooling operation. Thus, in this paper, we treat the kernel size as hyperparameters, which are set to 3×3 smaller kernels. On USC-HAD dataset [18], we visually analyze how the deformable filter evolves during a training stage. A sensing device called MotionNode containing a three-axial accelerometer is attached to each volunteer subject's front right hip for data collection, in which x axis that points to the ground is orthogonal to the plane formed by y and z axes. To be specific, the acceleration signal along x axis represents upwards and downwards movements in a vertical direction, while the signals along z and y axes are an indicator of forward/backward movements and lateral movements respectively. Because accelerometers often play an important role in recognizing daily human activities, we only

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2022.3158427, IEEE Transactions on Instrumentation and Measurement

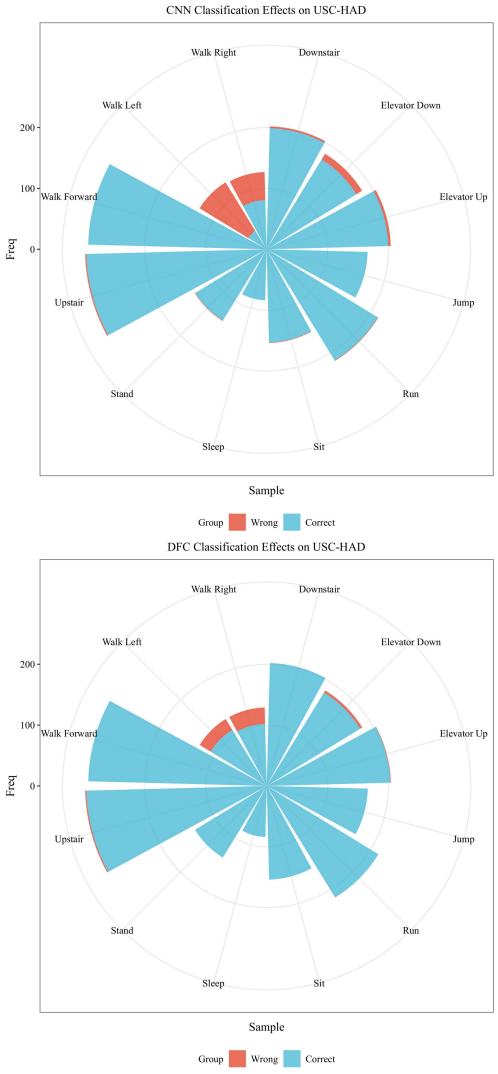

Fig. 15: Classification Comparisons on USC-HAD dataset between DFC and CNN-base

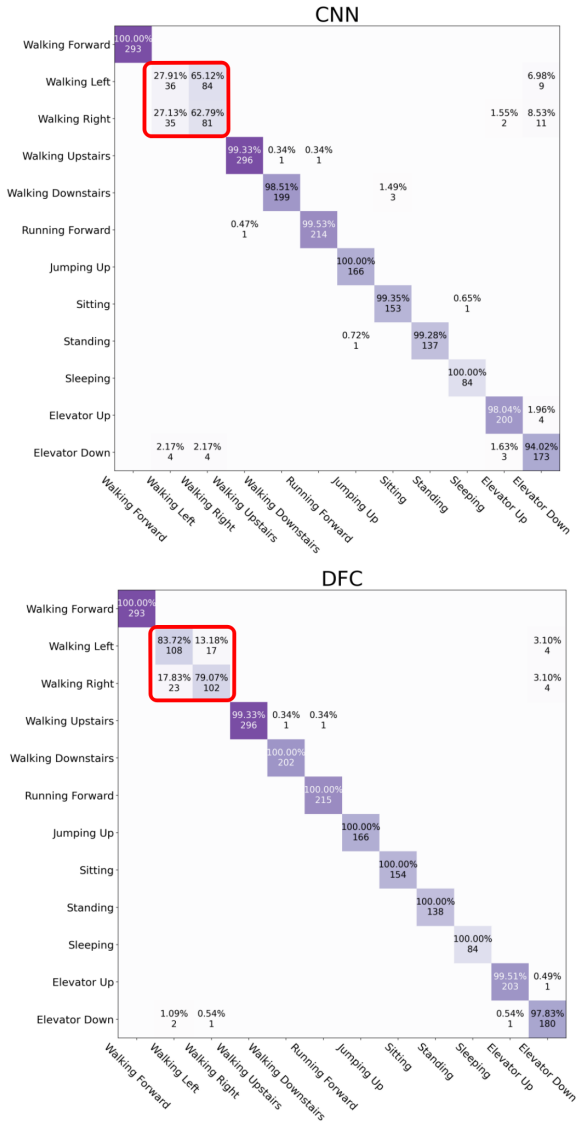

Fig. 16: Confusion matrices on USC-HAD dataset

show the deformation over the three raw acceleration signals. Fig. 14 shows four types of activities labeled as "walking forward", "walking left", "elevator up", "jumping". Instead of a regular sampling grid, it can be observed that for the signals of "walking forward" and "elevator up" activities the sampling locations within the filter tend to concentrate on z and x axes respectively (*e.g.*, forward/backward direction and vertical direction), while there are nearly no sampling locations distributed over y axis (lateral direction). On the contrary, in the case of "walk left", the proposed deformable model focuses more on y axis, *i.e.*, lateral direction, which corresponds to the "left" direction. For the "jumping" activity that has salient features with upwards and downwards movements, it could be seen that most sampling locations concentrate on x axis. In addition, we could clearly observe that these sampling locations demonstrate an obvious trend to aggregate over a shorter time span. As we have known, the "jumping" activity only occurs in a short time interval rather than in the entire sample window. The observation results suggest that the deformable filter can flexibly model irregular geometric transformations for sensor signals via tracking changes from various input activities, which agrees well with common intuition.

2) Visualization of classification effect with deformable convolutions: Fig. 15 shows the recognition comparison between DFC and CNN in the case of USC-HAD dataset [18]. The area of each sector in the figure denotes the number of samples for each activity, in which the light blue and red part correspond to the correct and wrong classification samples respectively. In comparison with the baseline CNN, it can be observed that the deformable convolution produces a much lower error rate in "Walk Right" and "Walk Left" activities, while in other cases it achieves almost 100% classification accuracy.

3) Confusion matrice: To clearly show what classes the data was classified into, the confusion matrices for standard convolution and deformable convolution are computed respectively on USC-HAD dataset [18]. As shown in Fig. 16, it can be seen that there is a large number of misclassified samples that occur between "Walking Left" and "Walking Right". This is due to that their signal waveforms could be very similar, which are very hard to discriminate. Due to irregular geometric

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2022.3158427, IEEE Transactions on Instrumentation and Measurement

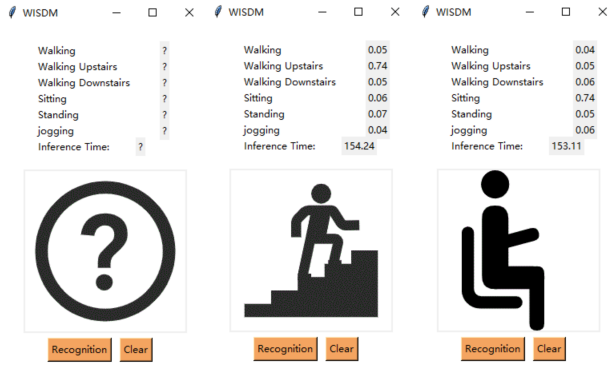

Fig. 17: The user interface of the application software

TABLE IV: Inference time on Raspberry Pi system

Model	Inference Time(ms)
CNN(Baseline)	140-161
CNN-DFC	153-172

transformations, the deformable convolution may lead to an obvious performance improvement, which reduces the number of misclassifications from 84 and 35 to 17 and 23 respectively. The results indicate that our method is able to better capture discriminative features for activity recognition.

D. Deformable convolution algorithm's deployment on Raspberry Pi

Besides the classification performance, we have to measure the actual inference time of the deformable model in embedded systems, due to the limit of computational overhead. For simplicity and without loss of generality, the real-time HAR measuring systems should be implemented in three steps: 1) train our network with collected training sensor data from WISDM dataset; 2) import this network into an embedded system; 3) run trained network on the embedded system to read real-time data and output the prediction. Due to its advantages in price and system compatibility, the embedded measurement system used in this experiment is the Raspberry Pi 3 B plus, equipped with ARM Cortex-A53 and 1GB LPDDR2 SDRAM. We evaluate the proposed deformable model via installing PyTorch on the 32-bit Raspberry PI OS. In order to perform the real-time prediction, we perform timing when the model is loaded and starts to output a prediction. Fig. 17 illustrates the user interface of the application software. Table IV summarizes the inference time with two network structures which include CNN and DFC-CNN. Fig. 18 illustrates the measured inference time of both architectures over 400 runs. During the prediction process, it can be seen that each sliding window takes around 140-161ms to complete one prediction by CNN. In the case of deformable convolution, the inference speed reaches 153-172ms per window. There is no significant increase in the inference time caused by the deformable convolutional network. The deformable model can easily perform real-time prediction on this embedded measuring platform.

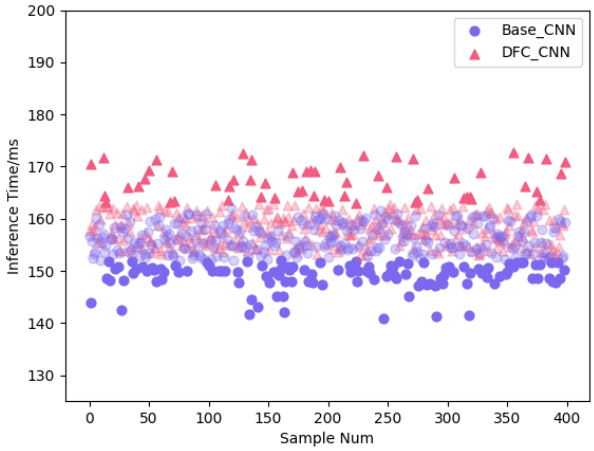

Fig. 18: Inferrence time distribution (400 samples)

V. Conclusion

In this paper, we first present a deformable convolutional network by modulating the offset and the feature amplitude within normal filters, which produces a flexible grid to sample sensor signals for improving the effectiveness of CNN in HAR scenarios. The deformable network demonstrates state-of-the-art performance on various HAR benchmarks. In addition, we also analyze several meaningful factors such as the offset, modulating the feature amplitude as well as their initializations, and visually validate the effective adaption of filter form, which leads to a better understanding of its mechanism. Finally, we show that the inference time obtained from a Raspberry Pi 3 B plus system, which is in line with the real-time requirements for HAR on resource-limited embedded systems. Deformable convolution still has a great potential to be excavated in HAR scenario. Despite the success of deep learning in ubiquitous HAR scenarios, the inner mechanism has not been fully revealed. It is necessary to explore the key factors that influence deep model decisions. Thus, the interpretability of deep decision behaviors has become a hot research topic. To resolve this issue, attention mechanism has been popular, which not only highlights which time interval or sensor modality that really matters, but also tells us which body part that contributes to identifying a specific activity. In a future study, we will investigate how to combine the deformable convolution with attention mechanism to achieve higher recognition performance, as well as better interpretability of deep model decisions for activity recognition tasks. Specifically, we can make a reformulation of deformable convolution by incorporating a new channel-wise attention-based modulation mechanism and then stack more such deformable convolutions into deep networks, which could further intensify the control of sampling over a broader range of feature levels.

References

[1] Y. Zhang, G. Tian, S. Zhang, and C. Li, "A knowledge-based approach for multiagent collaboration in smart home: From activity recognition to guidance service," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 2, pp. 317–329, 2019.

[2] T. Tuncer, F. Ertam, S. Dogan, and A. Subasi, "An automated daily sports activities and gender recognition method based on novel multikernel local diamond pattern using sensor signals," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 12, pp. 9441–9448, 2020.

[3] M. Abbas and R. L. B. Jeannes, "Exploiting local temporal characteristics via multinomial decomposition algorithm for real-time activity recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[4] Z. Chen, C. Jiang, S. Xiang, J. Ding, M. Wu, and X. Li, "Smartphone sensor-based human activity recognition using feature fusion and maximum full a posteriori," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 7, pp. 3992–4001, 2019.

[5] S. Xia, L. Chu, L. Pei, Z. Zhang, W. Yu, and R. C. Qiu, "Learning disentangled representation for mixed-reality human activity recognition with a single imu sensor," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–14, 2021.

[6] G. Panahandeh, N. Mohammadiha, A. Leijon, and P. Händel, "Continuous hidden markov model for pedestrian activity classification and gait analysis," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 5, pp. 1073–1083, 2013.

[7] W. Huang, L. Zhang, W. Gao, F. Min, and J. He, "Shallow convolutional neural networks for human activity recognition using wearable sensors," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[8] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE communications surveys & tutorials*, vol. 15, no. 3, pp. 1192–1209, 2012.

[9] G. Okeyo, L. Chen, H. Wang, and R. Sterritt, "Dynamic sensor data segmentation for real-time knowledge-driven activity recognition," *Pervasive and Mobile Computing*, vol. 10, pp. 155–172, 2014.

[10] 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 Transactions on Information Technology in Biomedicine*, vol. 16, no. 4, pp. 691–699, 2012.

[11] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.

[12] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 54, no. 4, pp. 1–40, 2021.

[13] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 510–519.

[14] W. Gao, L. Zhang, W. Huang, F. Min, J. He, and A. Song, "Deep neural networks for sensor-based human activity recognition using selective kernel convolution," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.

[15] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, Y. Cun *et al.*, "Learning convolutional feature hierarchies for visual recognition," *Advances in neural information processing systems*, vol. 23, pp. 1090–1098, 2010.

[16] R. Chavarriaga, H. Sagha, A. Calatroni, S. T. Digumarti, G. Tröster, J. d. R. Millán, and D. Roggen, "The opportunity challenge: A benchmark database for on-body sensor-based activity recognition," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 2033–2042, 2013.

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

[18] M. Zhang and A. A. Sawchuk, "Usc-had: a daily activity dataset for ubiquitous activity recognition using wearable sensors," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 2012, pp. 1036–1043.

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

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[24] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[25] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *arXiv preprint arXiv:1506.02025*, 2015.

[26] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9308–9316.

[27] H. Gao, X. Zhu, S. Lin, and J. Dai, "Deformable kernels: Adapting effective receptive fields for object deformation," *arXiv preprint arXiv:1910.02940*, 2019.

[28] B. Kim, J. Ponce, and B. Ham, "Deformable kernel networks for joint image filtering," *International Journal of Computer Vision*, vol. 129, no. 2, pp. 579–600, 2021.

[29] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," in *6th International Conference on Mobile Computing, Applications and Services*. IEEE, 2014, pp. 197–205.

[30] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition." in *Ijcai*, vol. 15. Buenos Aires, Argentina, 2015, pp. 3995–4001.

[31] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE International conference on image processing (ICIP)*. IEEE, 2015, pp. 168–172.

[32] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1307–1310.

[33] S. Ha and S. Choi, "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors," in *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016, pp. 381–388.

[34] F. Luo, S. Khan, Y. Huang, and K. Wu, "Binarized neural network for edge intelligence of sensor-based human activity recognition," *IEEE Transactions on Mobile Computing*, 2021.

[35] H. Ma, W. Li, X. Zhang, S. Gao, and S. Lu, "Attnsense: Multi-level attention mechanism for multimodal human activity recognition." in *IJCAI*, 2019, pp. 3109–3115.

[36] 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.

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

[38] X. Zhou, W. Liang, I. Kevin, K. Wang, H. Wang, L. T. Yang, and Q. Jin, "Deep-learning-enhanced human activity recognition for internet of healthcare things," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6429–6438, 2020.

[39] D. Costarelli, M. Seracini, and G. Vinti, "A comparison between the sampling kantorovich algorithm for digital image processing with some interpolation and quasi-interpolation methods," *Applied Mathematics and Computation*, vol. 374, p. 125046, 2020.

[40] I. Suarez, A. Jahn, C. Anderson, and K. David, "Improved activity recognition by using enriched acceleration data," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015, pp. 1011–1015.

[41] A. Keshavarzian, S. Sharifian, and S. Seyedin, "Modified deep residual network architecture deployed on serverless framework of iot platform based on human activity recognition application," *Future Generation Computer Systems*, vol. 101, pp. 14–28, 2019.

[42] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas, "Window size impact in human activity recognition," *Sensors*, vol. 14, no. 4, pp. 6474–6499, 2014.

[43] F. Li, K. Shirahama, M. A. Nisar, L. Köping, and M. Grzegorzek, "Comparison of feature learning methods for human activity recognition using wearable sensors," *Sensors*, vol. 18, no. 2, p. 679, 2018.

[44] 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 Journal*, vol. 20, no. 13, pp. 7265–7274, 2020.

[45] N. Y. Hammerla, S. Halloran, and T. Plötz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2016, pp. 1533–1540.

[46] 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," *The Journal of Supercomputing*, pp. 1–21, 2021.

[47] H. Bi, M. Perello-Nieto, R. Santos-Rodriguez, and P. Flach, "Human activity recognition based on dynamic active learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 922–934, 2020.

[48] F. M. Noori, M. Riegler, M. Z. Uddin, and J. Torresen, "Human activity recognition from multiple sensors data using multi-fusion representations and cnns," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2, pp. 1–19, 2020.

[49] H. Cho and S. Yoon, "Applying singular value decomposition on accelerometer data for 1d convolutional neural network based fall detection," *Electronics Letters*, vol. 55, no. 6, pp. 320–322, 2019.

[50] S. P. Singh, M. K. Sharma, A. Lay-Ekuakille, D. Gangwar, and S. Gupta, "Deep convlstm with self-attention for human activity decoding using wearable sensors," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 8575–8582, 2020.

[51] 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," *Knowledge-Based Systems*, vol. 229, p. 107338, 2021.

[52] K. Wang, J. He, and L. Zhang, "Sequential weakly labeled multiactivity localization and recognition on wearable sensors using recurrent attention networks," *IEEE Transactions on Human-Machine Systems*, 2021.

[53] S. S. Alia, P. Lago, and S. Inoue, "Mcomat: a new performance metric for imbalanced multi-layer activity recognition dataset," in *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, 2020, pp. 232–237.

[54] F. Daghero, C. Xie, D. J. Pagliari, A. Burrello, M. Castellano, L. Gandolfi, A. Calimera, E. Macii, and M. Poncino, "Ultra-compact binary neural networks for human activity recognition on risc-v processors," in *Proceedings of the 18th ACM International Conference on Computing Frontiers*, 2021, pp. 3–11.

[55] H. Kwon, G. D. Abowd, and T. Plötz, "Adding structural characteristics to distribution-based accelerometer representations for activity recognition using wearables," in *Proceedings of the 2018 ACM international symposium on wearable computers*, 2018, pp. 72–75.

[56] A. Sivakumar, R. Anirudh, and P. Turaga, "Geometric compression of orientation signals for fast gesture analysis," in *2015 Data Compression Conference*. IEEE, 2015, pp. 423–432.

[57] Z. Liu, L. Yao, L. Bai, X. Wang, and C. Wang, "Spectrum-guided adversarial disparity learning," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 114–124.

[58] A. Ferrari, D. Micucci, M. Mobilio, and P. Napoletano, "Hand-crafted features vs residual networks for human activities recognition using accelerometer," in *2019 IEEE 23rd International Symposium on Consumer Technologies (ISCT)*. IEEE, 2019, pp. 153–156.

[59] W. Lu, F. Fan, J. Chu, P. Jing, and S. Yuting, "Wearable computing for internet of things: A discriminant approach for human activity recognition," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2749–2759, 2018.

[60] H. Gjoreski, J. Bizjak, M. Gjoreski, and M. Gams, "Comparing deep and classical machine learning methods for human activity recognition using wrist accelerometer," in *Proceedings of the IJCAI 2016 Workshop on Deep Learning for Artificial Intelligence, New York, NY, USA*, vol. 10, 2016, p. 970.

[61] H. Nematallah, S. Rajan, and A.-M. Cretu, "Logistic model tree for human activity recognition using smartphone-based inertial sensors," in *2019 IEEE SENSORS*. IEEE, 2019, pp. 1–4.

Shige Xu received the B.S. degree from Nanjing Normal University, Nanjing, China, in 2020. He is currently pursuing the M.S. degree with Nanjing Normal University. His research interests include activity recognition, computer vision, and machine learning.

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

Wenbo Huang received the B.S. degree from Nanjing University of Technology, Nanjing, China, in 2019. He is currently pursuing the M.S. degree with Nanjing Normal University. His research interests include activity recognition, computer vision, and machine learning.

Hao Wu received the Ph.D. degree in computer science from Huazhong University of Science and Technology, Wuhan, China, in 2007. Now, he is an associate professor at School of Information Science and Engineering, Yunnan University, China. He has published more than 50 papers in peer-reviewed international journals and conferences. He has also served as reviewers and PC members for many venues. His research interests include natural language processing, recommender systems and service computing.

Aiguo Song received the Ph.D. degree in measurement and control from Southeast University, Nanjing, China, in 1996. He is currently a Professor with the School of Instrument Science and Engineering, Southeast University. His current research interests include teleoperation, haptic display, the Internet Telerobot- ics, distributed measurement systems, and machine learning. Dr. Song is also the Chair of the China Chapter of the IEEE Robotics and Automation Society.