Journal of Automation and Intelligence 1 (2022) 100007

Contents lists available at ScienceDirect

Journal of Automation and Intelligence

journal homepage: www.keaipublishing.com/en/journals/journal-of-automation-and-intelligence/

Deep learning and transfer learning for device-free human activity recognition: A survey☆

Jianfei Yang, Yuecong Xu, Haozhi Cao, Han Zou, Lihua Xie *

The School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Keywords:
Human activity recognition
Deep learning
Transfer learning
Domain adaptation
Action recognition
Device-free

ABSTRACT

Device-free activity recognition plays a crucial role in smart building, security, and human–computer interaction, which shows its strength in its convenience and cost-efficiency. Traditional machine learning has made significant progress by heuristic hand-crafted features and statistical models, but it suffers from the limitation of manual feature design. Deep learning overcomes such issues by automatic high-level feature extraction, but its performance degrades due to the requirement of massive annotated data and cross-site issues. To deal with these problems, transfer learning helps to transfer knowledge from existing datasets while dealing with the negative effect of background dynamics. This paper surveys the recent progress of deep learning and transfer learning for device-free activity recognition. We begin with the motivation of deep learning and transfer learning, and then introduce the major sensor modalities. Then the deep and transfer learning techniques for device-free human activity recognition are introduced. Eventually, insights on existing works and grand challenges are summarized and presented to promote future research.

1. Introduction

Human Activity Recognition (HAR) [1] refers to the art of identifying human activities using Artificial Intelligence (AI) from the gathered human data by various kinds of sensor sources [2], which has been developing rapidly in the past few decades. Successful HAR applications include video surveillance [3], person identification (e.g. gait recognition) [4], smart home automation (e.g. gesture recognition) [5] and human–computer interaction [6]. Current solutions for HAR mainly rely on various sensors which categorize HAR into two types: device-based HAR and device-free HAR. Device-based HAR leverages wearable sensors, such as the Inertial Measurement Unit (IMU) equipped in mobile phones. A common example is the gait counting algorithm based on the XYZ-accelerometer. Device-free HAR does not require the user to take sensors. Device-free sensors are deployed in the surrounding environment and can detect human motions, mainly including video cameras, IoT-enabled Radio-Frequency (RF), and WiFi. As device-free HAR is more ubiquitous and convenient, its research based on various sensors has ballooned. Hence, this paper mainly focuses on device-free HAR and its deep learning solutions. Similar to device-based HAR, device-free HAR is also formulated as a pattern recognition problem. Traditional machine learning algorithms play a vital role in the early-stage research on device-free HAR. By extracting hand-crafted features, classic classifiers, such as support vector machine, random forest, and Naive Bayes, have made significant results. Though these methods yield excellent results in controlled environments or small datasets, these models cannot deal with real-world scenarios that are more complex and dynamic. Such defects are partially caused by the limited human knowledge that is leveraged for feature engineering. Hand-crafted features cannot generalize well when confronting complicated scenarios such as the heterogeneity of human activity. Furthermore, compared to device-based sensors, environmental factors can make a larger impact on device-free sensors such as background clutter in videos. Traditional machine learning methods could not deal with these challenges.

Recently, deep learning techniques have evolved promptly into a powerful tool for HAR. Empowered by back-propagation, deep models are able to learn robust features automatically by designing an objective function such as the cross-entropy loss of the HAR classification problem. In this fashion, deep learning has achieved remarkable performances in computer vision, natural language processing, and big data analytics. Device-free HAR systems are revamped to obtain more fine-grained results as they embrace deep learning models. Moreover, different from classic machine learning, the merits of deep learning are better reflected in a variety of more realistic learning scenarios with less

☆ This work is supported by NTU Presidential Postdoctoral Fellowship, "Adaptive Multimodal Learning for Robust Sensing and Recognition in Smart Cities" project fund, in Nanyang Technological University, Singapore.

* Corresponding author.
E-mail addresses: yang0478@e.ntu.edu.sg (J. Yang), xuyu0014@e.ntu.edu.sg (Y. Xu), haozhi001@e.ntu.edu.sg (H. Cao), zouh0005@ntu.edu.sg (H. Zou), elhxie@ntu.edu.sg (L. Xie).

https://doi.org/10.1016/j.jai.2022.100007
Received 8 August 2022; Received in revised form 18 August 2022; Accepted 22 October 2022

data and label, such as unsupervised learning, transfer learning, few-shot learning, online learning, and incremental learning, which enables robust HAR systems to perform at lower annotation costs.

To overcome the lack of annotated data in practice, transfer learning is widely used in device-free HAR with massive public datasets released. Deep learning models are pre-trained on existing datasets and then fine-tuned on specific applications or down-streaming tasks. Nevertheless, the domain shift caused by environmental dynamics and datasets biases severely hinders the performances of deep models. Domain adaptation, as an important transfer learning technique, aims to transfer knowledge from a well-known source domain to an unlabeled target domain by bridging the gap of domain shift. As shown in Fig. 1, it helps deep HAR models adapt to more challenging situations such as HAR in the dark [7] and cross-site WiFi human sensing [8], which plays a crucial role in real-world device-free HAR applications.

Although there have been some surveys in deep learning [9,10] for device-based HAR or visual HAR [11,12], there has been no specific survey that investigates device-free HAR based on deep learning and transfer learning. As far as we know, this is the *first* article that summarizes the recent progress of deep learning and transfer learning for ubiquitous device-free HAR systems. This survey introduces how device-free HAR system works and how deep learning and transfer learning promote device-free HAR applications. By summarizing existing literature, we put forward the directions of the future research and prospect. The papers in this survey are selected using Semantic Scholar and Google Scholar with the keywords of activity recognition, visual action recognition, transfer learning, deep learning, and device-free. We mainly consider the high-impact papers in the top-tier conferences including MobiCom, MobiSys, Ubicomp, PerCom, CVPR, ICCV, ECCV, NeurIPS, and peer-reviewed journals. For the dataset papers, we search academic datasets in Mendeley Data, Github, and IEEE Dataport. The hard criteria ensure the quality of papers in this review.

The rest of the paper is organized as follows. Section 2 introduces the background of device-free HAR, how deep learning is utilized and why transfer learning takes effect. In Section 3, we summarize the main sensor modalities for device-free HAR systems. In Section 4, we review the deep learning-based HAR approaches, while the transfer learning-based approaches are reviewed in Section 5. Related datasets are also introduced. Then we summarize the existing works and present some insights in Section 6. The future directions are discussed in Section 7. The paper is concluded in Section 8.

2. Background

2.1. Device-free activity recognition

The objective of human activity recognition is to understand and predict human behaviors according to specific requirements. Device-free HAR leverages various sensors that are deployed in the surrounding environment. For a specific timestamp t, a device-free sensor capture the human motions as

$$m_t = \{d_t^1, d_t^2, \dots, d_t^D\}, \tag{1}$$

where m_t is the sensor data at time t, d_t^i denotes the ith dimension of the sensor and D denotes the total number of dimensions of m_t. For example, D is the number of RGB pixels for camera-based HAR [11] while D is the number of subcarriers for WiFi-based HAR [13]. Each sample of human activities may last for a period of time τ, which is written as

$$x^i = \{m_1, m_2, \dots, m_\tau\}, \tag{2}$$

where the dimension of a sample x^i is $x \in \mathbb{R}^{D\times\tau}$. A device-free HAR dataset $\mathcal{X}$ consists of N samples such that

$$\mathcal{X} = \{x^i, y^i\}_{i=1}^N, \tag{3}$$

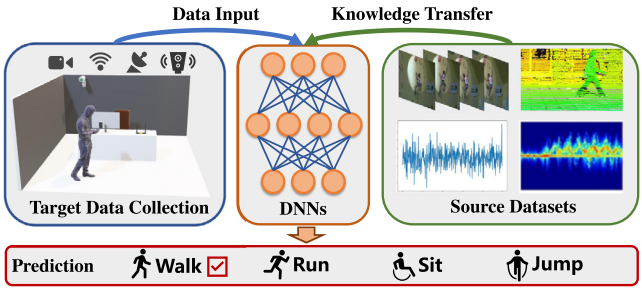

Fig. 1. An illustration of device-free HAR using deep neural networks and transfer learning skills. Deep neural networks have the strong capacity to recognize human activities from various modalities of data. Nevertheless, deep models may degrade due to the distribution difference between training and testing phase. Unsupervised domain adaptation leverages unlabeled target data to perform knowledge transfer, achieving significant improvement for practical HAR scenarios.

where y^i is the activity label of x^i and $y^i \in \{Y|Y = 1, 2, \dots, K\}$. K is the number of activities of the label space. For the HAR task, we aim to build a model $\mathcal{G}$ that can predict the activity label y based on the sensor data x

$$y = \mathcal{G}(x). \tag{4}$$

The sample-level HAR problem can be extended to frame-level classification or activity segmentation problems. Here we only focus on the sample-level HAR and review its recent progress.

2.2. Pros and cons of deep learning-based HAR

Compared to conventional machine learning, deep learning-based HAR generates better results and helps recognize more complex human activities. The merits of deep learning HAR approaches are as follows:

(1) Deep learning extracts more robust features using massive data. Though human-expert knowledge helps us design effective hand-crafted features for HAR [14], these features may not hold good generalization ability when confronting data under dynamic environments or more complex human activities, such as gym exercise that contains a series of irregular human motions. Deep learning methods learn the feature space that mostly contributes to the task automatically by novel network architectures and back-propagation [15], which overcomes the shortcomings of hand-crafted features [10].
(2) Deep learning can still learn representations in an unsupervised manner. Recent progress on unsupervised learning based on mutual information and contrastive learning [16,17] tackles the problem of the requirement of a large amount of well-annotated data that is costly.

Nevertheless, in real-world applications, deep models still encounter challenges:

(1) Deep learning models still require sufficient labeled HAR data to obtain a good classifier. In practice, such data is either expensive to annotate or hard to collect.
(2) For device-free HAR, the dynamic and complex environments can degrade the deep learning performance [5,8]. Even though deep learning features are robust, statistical learning approaches still follow the assumption that the training and testing data are independent and identically distributed. This might be broken when the environmental dynamics change the HAR data distribution. For example, in visual HAR, the training samples mostly come from the ideal condition with good illumination conditions and clear human targets, but in the real world, the testing scenario might be at night and the targets are severely occluded.

2.3. Why transfer learning?

To deal with the challenges of the lack of tremendous annotated data for specific tasks, transfer learning comes into existence [18]. The most common technique is pre-training [19] and fine-tuning. Based on existing public datasets, one can pre-train the model in advance, and then re-use the parameters of the feature extractor (e.g. deep convolutional neural network [20]) to further train a classifier in that feature space. This is also applied to device-free HAR field [21]. With pre-trained parameters, deep models only require a small amount of annotated data for fine-tuning.

Pre-training has achieved remarkable improvement for the scenario where we face the shortage of labeled HAR data, yet it does not apply to all applications. The scope of its usage should rely on the assumption that the pre-training data is similar to the testing scenario, and transferring knowledge across distinct domains is not effective and even leads to negative transfer [22,23]. For example, the existing visual HAR datasets are mostly collected during the daytime with good illumination, but the target task is to conduct HAR in the dark [7]. Another example is the cross-site WiFi-based HAR [24,25]. Part of such difference can be formulated as the distribution discrepancy (i.e. domain shift) that is tackled by domain adaptation approaches [26]. Domain Adaptation (DA) can transfer knowledge from a label-rich source domain to a label-scarce or unlabeled target domain. With DA, deep HAR models can adapt to various unseen scenarios without expert annotations.

In the following sections, we begin with kinds of device-free sensor modalities, and then review the recent progress of deep learning and transfer learning models for robust HAR.

3. Device-free sensor modalities

Various device-free sensor modalities have been developed to provide the different granularity of sensing solutions for human activity recognition. The main characteristics of these technologies (*e.g.* the cost, granularity, and the preservation of privacy) are summarized in Table 1.

Table 1
Overview of major sensing technologies.

Technology	Granularity	Cost	Privacy-preserving
Camera	High	Moderate	No
Infrared	Low	Low	Yes
RF	Middle	Moderate	Yes
Radar	Middle	Moderate	Yes
Lidar	High	High	No
ES	Low	Low	Yes
WiFi	High	Low	Yes

3.1. Cameras

Nowadays, cameras are everywhere as CCTV is widely deployed for security. Massive images and videos can be captured via cameras, and with the development of deep learning, accurate visual HAR systems have been developed [27]. Visual sensing has a high granularity of sensing materials but also arouses privacy concerns such as HAR in hospitals or smart homes. Furthermore, illumination and occlusion also severely affect the performance of visual sensing under some specific circumstances [7,28], which requires other sensing techniques for a complement.

3.2. Lidar

Lidar is capable of measuring distances to a target by illuminating the target with laser light and then measuring the reflected light. Using laser return times and wavelengths, 3D representations of the target can be obtained. Its sensing granularity is very high, which enables lidar-based object recognition [29] and HAR [30]. Despite high-dimensional Lidar data, deep learning models have sufficient capacity to capture discriminative features for HAR [31]. Nevertheless, Lidar is expensive and the processing of its data requires high computing resources, which makes it not suitable for edge-side HAR applications such as HAR in smart homes.

3.3. Radar

Radar-based HAR is more appealing since it is cost-effective and privacy-preserving [32]. Different from Lidar that employs laser light, radar leverages radio waves based on transceiver antennas, such as mmWave radar [33]. The point cloud can be obtained after a series of transformations [34]. Radar can detect human motions for a further distance, and it is not affected by illumination or weather. However, the data granularity is not as high as lidar, which brings more challenges to model design.

3.4. Pyroelectric infrared (PIR)

Pyroelectric infrared sensor (PIR) can detect human motions in a designated area. A surface electric charge is generated by a PIR sensor when it is exposed to heat in the form of infrared radiation. The range of infrared wavelength is from 700 nm to 1 mm and this is longer than visible light, which makes PIR-based sensing less intrusive than visual sensing. PIR is widely used in HAR applications in the smart buildings including occupancy detection [35] and intruder detection [36]. However, the sensing granularity of PIR is not high, so it does not have to rely on deep learning and cannot be employed for fine-grained activity recognition or other high-level tasks.

3.5. Environment sensors (ES)

Environment sensors mainly include light, wind, air quality, temperature, humidity, and CO_2 sensors. Being deployed in an indoor environment, they usually reflect human living comfort. ES is usually utilized for improving thermal comfort and saving energy [37]. The data collected by ES is also used to predict the occupancy number [38]. Naturally, as more people lead to more heat and CO_2 concentration, simple statistical learning methods such as Linear Discriminant Analysis (LDA) and Random Forest (RF) can produce satisfactory results. The drawback of ES is that the large-scale deployment of ES brings high costs and ES cannot provide high-resolution data for fine-grained sensing.

3.6. Radio frequency (RF)

Radio Frequency Identification (RFID) belongs to radio-based sensing [39]. Other radio-based methods include ZigBee radio and WiFi radio. Human activities are captured by RF signals in a passive manner. Traditional statistical learning models such as K-Nearest Neighbors (KNN) achieve satisfactory performance for RF-based activity recognition. Deploying multiple RF sensors enables device-free HAR for simultaneously conducted activities [40]. The disadvantage of RF-sensors is that the sensing range is rather limited so RF sensors should be used in a large scale, which may lead to higher cost and additional deployment processes.

3.7. Wireless local area network (WiFi)

Nowadays, WiFi access points have been deployed in most of the commercial and residual buildings, and nearby every IoT device, such as soundbar, TV, thermostats, and power switch, are equipped with a WiFi module. Recently, Channel State Information (CSI) is extracted from the physical layer of wireless communications, which reflects the

situations of the multi-path propagation of wireless signals in indoor environments. WiFi-based sensing is conducted in a device-free manner, and the granularity of CSI data is moderately high. Recent literature has witnessed many successful employments of CSI measurements for various HAR applications, such as device-free indoor localization [41], action recognition [42,43], gesture recognition [8], human identification [4], smoking detection [44] and crowd counting [45,46]. However, due to the individual heterogeneity and environmental dynamics, deep models may capture intrinsic noises in CSI data, which decreases the performance.

3.8. Ultra-wideband (UWB)

The ultra-wideband refers to the radio communication with a large effective bandwidth that is larger than 500MHz, which means that it can transmit enormous data at a short distance. Meanwhile, UWB radar is empowered by a number of rapid short pulses that occupy the entire bandwidth. Furthermore, it is not sensitive to the multi-path effect due to the high-time resolution. UWB-based HAR is becoming more and more popular for smart home automation and indoor localization services [47,48]. The Apple iPhone 11 has already been equipped with UWB beacons [49].

4. Deep learning for device-free HAR

4.1. Deep neural networks

Inspired by the biological neural systems [50], deep learning methods based on Deep Neural Networks (DNNs) have been proposed to extract data features in an automatic manner. Among the various types of DNNs, Convolutional Neural Networks (CNNs) yield exceptional performances in image-based tasks, such as image classification and object detection. CNNs consist of multiple layers which include convolution, pooling, activation, and fully connected layers. One of the pioneering works is LeNet [51], which has been successfully applied to hand-written digit classification. Over the past few years, deeper and more complex CNNs [52–55] have been subsequently introduced and achieved satisfying results on large-scale datasets such as ImageNet [56]. For HAR application, 1D or 2D CNN can extract spatial and temporal features, but the limitation is the size of the convolution kernel which only focuses on local patterns. If the length of patterns is long such as temporal data, then CNN may not be effective enough.

To capture both long-term and short-term patterns, Recurrent Neural Network (RNN) was proposed [20]. Long Short-Term Memory (LSTM) [57] is a specialized design of RNN, which uses multiple gate units to forget or memorize specific signals. LSTM is suitable for representation learning of sequence data, which is widely used in natural language processing [58] and temporal data modeling [59]. However, LSTM requires sufficient training data and the computational complexity of its training is high when compared to CNN. For CSI data, when long-term patterns are important, LSTM is a panacea that learns to memorize these patterns from the temporal axis. Then we review the application of CNN and LSTM for HAR with recently advanced architectures (e.g. transformers) is also included.

4.2. Applications for camera-based HAR

Visual HAR is normally performed through videos, which could be viewed as a collection of multiple images placed sequentially across time. As one of the most commonly used DNNs, CNNs have been applied broadly for visual HAR. From the type of CNNs utilized, deep learning methods for visual HAR could be generally grouped into two categories: 2D-CNN based methods [60,61] and 3D-CNN [62–64] based methods.

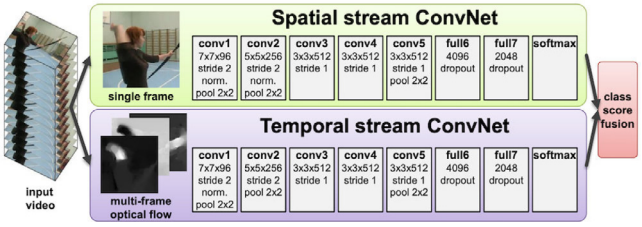

Fig. 2. Typical structure of a two-stream network.
Source: Simonyan et al. [65].

4.2.1. HAR feature learning with independent spatial and temporal feature extraction

Early deep learning methods for action recognition, such as those proposed in [60,65] utilize 2D-CNNs to extract features from videos. In both methods, a video frame is sampled and used as input to a 2D-CNN for feature extraction. Meanwhile, previous researchers suggest that human would also process videos in a two-stream manner: the Ventral Stream processes object attributes such as object appearances and object colors; while the Dorsal Stream processes the motions and locations of the object [66]. Inspired by such research, a separate stream is added to extract temporal features embedded in videos, utilizing optical flow [60,65] computed using algorithms such as TV-L1 [67]. The structure of the spatial stream and the temporal stream is usually similar or even identical (e.g., in [65]), yet they are trained separately. The softmax scores of each individual stream are combined with a late fusion strategy. A typical structure of the two-stream network is presented in Fig. 2.

Subsequently, multiple networks have been proposed to improve on the early two-stream networks. One improvement concerns the fusion strategy between the spatial and temporal streams [60], where the features extracted from each stream are fused before obtaining the softmax scores; another improvement is proposed in ST-ResNet [68], where the ResNet [54] is used as the feature extraction backbone. The Temporal Segment Network (TSN) [61] improves from previous work by segmenting videos into clips where the spatial and temporal features are extracted, the overall video feature is obtained through segmental consensus fusion. DOVF [69] extends from TSN by employing a two-stage classification strategy, while TRN [70] segments videos into clips of multiple temporal scales. Alternatively, handcrafted features such as iDT are aggregated with two-stream networks to achieve better performance [71]. Meanwhile, computation of optical flow requires high computational power and large storage resource. Additionally, optical flow requires pre-computation, which prohibits fully end-to-end training. To address such limitations, subsequent works propose to estimate optical flow through a trainable neural network. FlowNet [72] which learns optical flow from synthetic ground truth data is one example. Subsequently, MotionNet [73] estimates optical flow through the prediction of successive frames, LMoF [74] constructs a learnable directional filtering layer, while TVNet [75] unfolds the TV-L1 [67] algorithm and formulates it with a neural network. Rep-Flow [76] extends TVNet by constructing a convolutional layer for optical flow estimation.

There are also other methods built based on 2D-CNNs while avoiding the use of optical flow. A typical strategy involves the use of Recurrent Neural Networks (RNNs) and their variants for modeling temporal features in the form of sequence information. One seminal work is the Long-term Recurrent Convolutional Network (LRCN) [77], which extracts features of each frame with 2D-CNNs, while the overall video feature is modeled through a Long Short-Term Memory [57] (LSTM)-style RNN. More recently, Shi et al. proposed ShuttleNet [78] which is constructed by Gated Recurrent Units [79] (GRU) that are loop connected.

4.2.2. HAR feature learning with joint spatiotemporal feature extraction

A more direct approach towards applying deep learning to visual HAR is to extract temporal features jointly with spatial features, applying convolutional operations on both spatial and temporal dimensions. The 3D-CNN is first introduced in [80], performed on videos preprocessed with a hardwired layer. The convolutional kernels used in [80] are 3D kernels, where the filters are extended along the temporal dimension. Empirically, 3D-CNN outperforms 2D-CNN without optical flow by a noticeable margin. Subsequently, a slow fusion strategy is proposed in [81] that fuses video features obtained from multiple clips progressively utilizing 3D-CNN. Further, C3D [82] is introduced as a generic video feature extractor, with full video frames as input while employing homogeneous convolutional kernels. With larger and deeper networks such as VGG [52], ResNet [54] and ResNext [55] introduced and achieved outstanding performances, this progress is also employed in 3D-CNNs for visual HAR. I3D [62], 3D-ResNet [83] and 3D-ResNext [84] are built by expanding the convolutional kernels of their 2D-CNN counterparts to the temporal dimension and are all deeper and larger 3D-CNNs compared to C3D.

Though 3D-CNNs are able to model temporal features with spatial features jointly without optical flow, their parameter size is much larger than their 2D counterparts, resulting in increase computation and difficulty for training. To address such issues, I3D [62] propose to initialize 3D-CNNs by inflating weights of 2D-CNNs trained for image classification. Meanwhile, R(2+1)D [83] proposes to improve 3D-CNN efficiency by separating spatial convolution operation with temporal convolution operation. This strategy is shared by S3D [85] and P3D [86]. Alternatively, Channel-Separated Convolutional Network (CSN) [87] shows that the efficiency and effectiveness of 3D-CNNs could also be improved by performing convolution operation across channels separately. MFNet [88] shares such a strategy, while including multiplexer modules to facilitate information flow across channels. More recently, SlowFast Network [89] is proposed to include a slow and fast pathway for modeling spatial and temporal semantics separately. Further, correlation information embedded within videos could be combined with 3D-CNNs to further boost performance, as suggested in ACTF [63] which utilizes inter-frame regional correlation, as well as KPSEM [64], which utilizes correlation between spatiotemporal key points.

4.2.3. Self-attention for HAR feature learning

Recently, there has been a rise of research interest in utilizing self-attention as a means of feature extraction through correlation within the input data. Self-attention has been proven effective in Natural Language Processing (NLP), and has subsequently been utilized in visual HAR. One seminal work concerns the Non-Local network (NLNet) [90], which expands the self-attention into spatiotemporal features and extracts long-range spatiotemporal correlation features. Variants of the NLNet improves from the perspectives of network generalization [91, 92] and network efficiency [93,94]. The above models still require CNN as a backbone for feature extraction. With the recent success of pure self-attention-based networks such as the Transformer [95], such a strategy has also been applied to visual HAR. Works in [96–98] expand the Transformer to suit the spatiotemporal features of videos, achieving notable performances without CNN.

4.2.4. Benchmark datasets

The development of various benchmark datasets is a key driving force for the rapid progress of deep learning methods in visual HAR. To evaluate the performance of the various methods presented in Section 4.2, a number of datasets are established. Earlier vision-based action recognition benchmark datasets include KTH [99], Weizmann [100], and IXMAS [101]. In general, these datasets contain a relatively small number of action classes and are collected offline without using public available videos. Performances on these previous datasets are mostly saturated, partly due to their small scale. Larger datasets such as Hollywood2 [102], Olympic Sports dataset [103], HMDB51 [104], UCF50 [105] and UCF101 [106] are subsequently introduced, where videos are collected from public video platforms such as YouTube. Both the HMDB51 and UCF101 are still considered challenging benchmarks and progress have been made constantly throughout the past decade.

More recently, larger-scale datasets have been further introduced to include more classes and videos. These include the Kinetics dataset [107] which is one of the largest datasets in terms of the number of action categories. It has become the primary choice in action recognition studies thanks to its scale. Meanwhile the Something–Something (V1/V2) dataset [108] are introduced as a benchmark with more fine-grained actions (e.g. “Pushing something from right to left" instead of ‘Pushing"), and allows methods to develop a fine-grained understanding of actions. Subsequently, the scale of visual HAR datasets have further increased, with the launch of “mega-video" datasets such as Moments in Time [109]. The introduction of these larger datasets help push the boundaries of vision-based action recognition, and lead to the proposal of more sophisticated models. However, large datasets require large storage and are time-consuming without powerful computation tools (i.e., GPUs or TPUs). To accommodate such limitations, down-scaled datasets such as the MiniKinetics [85] are introduced with less action categories and videos.

Recently, there has been a rapid increase of research interest with regards to computer vision tasks in adverse environments, such as face recognition in the dark [110–112]. The rise in research for visual tasks under adverse environments has been further expanded to the video domain. For visual HAR in adverse environments, the ARID [7] is introduced as the first public dataset for HAR in the dark. Empirical results have shown the inability of current deep learning methods to cope with visual HAR in adverse environments, and suggest additional techniques to be applied. Table 2 summarizes the main characteristics (number of action classes and videos, collection method and year of publication) of all the datasets mentioned above.

4.3. Applications for IoT-enabled HAR

4.3.1. Deep learning enabled HAR systems

Different from visual HAR, other modalities of data, collected from IoT sensors mentioned in Section 3, do not have delicate granularity, leading to smaller data dimensions. Normally speaking, very deep models are not instructed for IoT-enabled HAR systems. It is seen that deep learning approaches have been successfully applied to most of the device-free sensor modalities. For RF-based HAR, Chen et al. propose a lightweight CNN model based on point-wise grouped convolution and depth-wise separable convolutions [113]. Meanwhile, LSTM is leveraged for HAR based on infrared motion sensors [114]. As LSTM has a strong capacity to extract temporal data from sequences, it is also applied to HAR based on environmental sensors [115] that has low granularity but more sensing perspectives. For mmWave radar, both its doppler signatures and generated point cloud can be fed into deep HAR models [32]. Doppler signatures such as micro-Doppler spectrogram [116] can be regarded as image and processed by 2D-CNN, while point cloud data requires special designs of deep models, such as PointNet and Graph Neural Network (GNN) [117]. Meng et al. utilize CNN, GNN, and PointNet for Radar-based gait recognition and achieve 90% accuracy [118]. UWB-based HAR systems rely on high-resolution Channel Impulse Responses (CIR) [119], which achieves high recognition accuracy of complex activities based on 2D-CNN. WiFi CSI data is also an estimation of CIR but it is affected by multi-path effects [120]. Even so, WiFi-based HAR systems still have high recognition performance and lower cost based on purely CNN models [8] or Long-Term Recurrent Convolutional Network (LRCN) [121]. Deep learning has enabled many WiFi-based HAR applications including occupancy detection [46], gesture recognition [8,121–123], person identification [4], crowd counting [124,125] and sign language classification [126].

Table 2
Overview of current visual HAR datasets.

Dataset	# Classes	# Videos	Collection method	Year	Download link
KTH [99]	6	2,391	Offline	2004	Website
Weizmann [100]	10	90	Offline	2007	Website
IXMAS [101]	11	1,148	Offline	2007	Website
Hollywood2 [102]	12	3,669	Online	2009	Website
Olympic Sports [103]	16	800	Online	2010	Website
HMDB51 [104]	51	6,849	Online	2011	Website
UCF50 [105]	50	6,676	Online	2012	Website
UCF101 [106]	101	13,320	Online	2012	Website
Something–Something v1 [108]	174	108,499	Offline	2017	Website
Something–Something v2 [108]	174	220,847	Offline	2018	Website
Kinetics-400 [107]	400	160,000	Online	2017	Website
MiniKinetics [85]	200	85,000	Online	2018	Website
Moments in Time [109]	339	1,000,000	Online	2019	Website
ARID [7]	11	3,784	Offline	2020	Website

Table 3
Overview of IoT-enabled HAR datasets (h: hours; frs: frames).

Dataset	Sensor modality	Sampling rate	# Classes	# Subject	# Samples	Year	Download link
Kasteren [134]	RF	NA	8	1	245	2008	Website
SBR-WiFi [135]	WiFi	1000 Hz	6	6	720	2017	Website
SignFi [126]	WiFi	200 Hz	276	5	8280	2018	Website
WiAR [136]	WiFi	30 Hz	16	10	4800	2018	Website
Widar-Gait [137]	WiFi	1000 Hz	16	16	12000	2019	Website
mmGait [118]	mmWave Radar	NA	95	95	30 h	2019	Website
LboroHAR [132]	Lidar, RGBD	NA	9	16	136710 frs	2019	NA
CI4R [138]	mmWave Radar	NA	11	6	2640	2020	Website
IR-UWB [139]	UWB	NA	6	8	4230	2020	Website
LAMAR [133]	Lidar	NA	7	3	NA	2020	NA

A more important factor is the limitation of computational complexity due to the constraint of edge devices. Two solutions are provided. The first one is to design and employ lightweight models with fewer parameters and computation burdens, which could be achieved by model compression [127]. For example, compressive sensing is leveraged for RF-based HAR to learn a compact and de-noised representation [128]. Network pruning also contributes to accelerating deep HAR models by cutting redundant connections, which plays an important role in high-resolution modalities (i.e. point cloud), such as radar-based and Lidar-based HAR [129,130]. The second scheme is to transmit the data to a cloud server for model inference. However, higher communication traffic is demanded for edge HAR devices. This could be possibly solved by the auto-encoder and quantization. For instance, EfficientFi is developed for large-scale wireless sensing, which consists of a quantized autoencoder that compresses and decodes the HAR sensing data, and a joint training classifier that conducts the activity recognition [131]. The discrete feature space learned by the autoencoder is communication-friendly for cloud centralized computing.

4.3.2. Benchmark datasets

Deep learning for IoT-enabled HAR systems is not as thriving as that for visual HAR. This is partially caused by the lack of high-quality annotated HAR datasets for various modalities. We extensively survey these datasets from the perspectives of sensor modality, activity number, sample number, sampling rate, subject number, and collection year. Some datasets that have not been divided into samples are released, so the sample number is filled by their collection duration. The datasets are summarized in Table 3. Two of these datasets are not publicly available, which may require access from their authors [132,133]

5. Transfer learning for device-free HAR

5.1. Transfer learning and domain adaptation

Though deep learning has a strong fitting capacity, it still requires massive labeled data that is hard and expensive to collect for some tasks. Pre-training deals with this problem by learning prior knowledge on the existing dataset and then reusing the parameters for downstream tasks. However, it still cannot generalize to a domain whose distribution is different from the training set. The reason lies in a major assumption in statistical learning algorithms that the training and future testing data share the same feature space and thus have the same distribution. In many real-world scenarios, this assumption does not hold. This could be tackled by domain adaptation that aims to transfer knowledge from a label-rich source domain to a label-scarce or even unlabeled target domain [140]. Denote $\mathbb{P}$ the source distribution and $\mathbb{Q}$ the target distribution. The problem is that $p(x) \neq q(x)$. For the same task, the assumption of domain adaptation is that the conditional output distribution is invariant, *i.e.* $p(y|x) = q(y|x)$, where x, y are the input data and output labels, respectively [18]. As such, most domain adaptation algorithms deal with the divergence of distributions between the training and future testing data for the same task.

For HAR applications, DA tackles two important issues: lack of data in the testing scenario and the dataset bias between the training and testing data. In a word, DA bridges the gap between the well-trained model and a new test scenario. For example, in wireless sensing, the sensor data is a superposition of human motions and environments. The environmental changes are detrimental to the performance of HAR model [124], which requires us to apply DA. Currently, the main streams of DA include statistics-based DA [141], adversarial DA [142], semantic DA [143] and entropy minimization [144]. The DA scenarios include closed-set DA [145], partial DA [146], open-set DA [147], multi-source DA [148] and source-free DA [149,150]. Here we review the pre-training and DA applications for visual and IoT-enabled HAR systems.

5.2. Applications for cross-domain visual HAR

5.2.1. Pre-training and fine-tuning

Early works for cross-domain visual HAR attempt to retrieve the success of the pre-training in the image domain in a fully supervised manner. Specifically, networks are first pre-trained on large-scaled labeled datasets to learn general representations and subsequently fine-tuned on target datasets. Since videos are constructed by multiple

image frames along the temporal dimension, there are two alternatives for fully supervised pre-training: one is to directly leverage the pre-trained model from the image domain for video action recognition, while the other is to pre-train the models on early large-scale video datasets (e.g. Kinetics-400 [107], Sports-1M [81]). One of the primary works [62] compares the difference between these two pre-training alternatives and suggests that pre-training on large-scale video datasets may lead to more robust temporal features, achieving better performance than the image pre-trained strategy. [84] further expands the scope to deep models and justifies that pre-training strategy leads to exceptional improvement when using deeper models. With emerging larger scales of labeled datasets and deeper networks, pre-training on large-scale video datasets becomes a popular strategy in the following works [85,89] to boost the performance on different target datasets.

Pretraining on large-scale datasets in a fully supervised manner brings consistent improvement, while it requires resource-expensive manual annotations to generate label information. To overcome this drawback, recent methods pay more attention to self-supervised learning which profits from the more accessible unlabeled data during the pre-training stage. More specifically, the core of self-supervised learning is to design pretext tasks where supervision signals are automatically generated based on the characteristic of data itself. The network can therefore learn effective representation without any manual annotation during the pre-training stage and subsequently transfer the knowledge to the fine-tuning stage. Based on the design of pretext tasks, previous self-supervised methods for visual HAR can be categorized into three types: (i) dense prediction, (ii) spatio-temporal reasoning, and (iii) contrastive learning.

Self-supervised methods based on dense prediction require networks to predict the low-level information of videos, including video frames [151,152] and frame patches [153–155]. Inspired by the generative networks, [151] proposed a dense prediction pretext task that requires the network to predict the future frames given recent frames as input. [152] further extends this pre-text tasks by introducing extra optical flow input and foreground–background disentangle. Following works attempt to elaborate different modalities as input or prediction target, such as 3D videos [156] and RGB-D data [157]. Inspired by the recent success of dense prediction with Transformer in image domain, [158] adopts Transformer to predict Histograms of Oriented Gradients (HOG) of masked patches and achieves state-of-the-art performance, even outperforming the SOTA fully-supervised HAR methods. Despite its outstanding performance, the dense prediction methods require an additional generator head, leading to extra computational cost. Additionally, it might be less efficient to directly predict low-level information, since these high-frequency details provide limited contributions towards the high-level tasks (i.e. HAR).

Instead of directly predicting low-level information, spatio-temporal reasoning methods propose to generate supervision signals through correlations of videos, such as temporal orders [159–161] and playback rates [162–164]. Inspired by the sequential relationships of video along the temporal dimension, [159] first proposes to leverage the temporal order of videos as the self-supervised supervision signal. More specifically, several clips are sampled from the same raw videos and only clips constructed with consecutive frames are considered as positive samples. The pretext task is therefore designed as a classification problem, where the network is trained to identify the positive samples based on its semantic understanding of the videos. Similarly, [165] proposes a temporal verification task based on sorting a series of shuffled clips, where the network is directly trained to predict the correct temporal order of the input clips. To further extract more effective spatio-temporal features, [161] first adopts 3D CNN rather than 2D CNN to perform temporal order prediction. In addition to temporal orders, recent works propose to leverage playback rates of videos as supervision signals. Specifically, Yao et al. [163] propose the first self-supervised pretext task called Playback Rate Prediction (PRP) to learn long-short-term video representations. Given a raw video, PRP samples video clips with various sample rates, resulting in short video clips with different playback rates. The network is trained to perform two sub-tasks, including playback rate prediction and video clip reconstruction. [162] simplifies the design of playback rate prediction by replacing the computationally expensive generative sub-task by contrastive learning. [164] further boosts the performance by combining playback rate prediction with other temporal augmentation methods, where the network is trained to predict not only the playback rate but also the type of augmentation method applied to augment the input. Compared to dense prediction methods, spatio-temporal reasoning methods are more efficient since they usually can be regarded as classification tasks that do not require additional modules to perform, while they are usually outperformed by methods based on dense prediction and contrastive learning.

Contrastive learning has achieved great progress during the past decade, progressively surpassing other self-supervised methods and even fully supervised methods. The core of contrastive learning is to maximize the mutual information between positive samples, which are usually generated as samples from the same raw video under different augmentations. For instance, [162] attempts to align spatio-temporal representations of the same action or same context as an additional sub-task, so that the network can extract mutual representation from video under different augmentations. While sharing a similar optimization objective (i.e speed prediction) with PRP [163], the introduction of contrastive learning brings significant improvement of performance. To fully leverage contrastive learning for video representation extraction, [166] proposed Sequence Contrastive Learning (SeCO) which conducts contrastive learning from spatial, spatio-temporal and sequential perspectives, outperforming previous image-pretrained fully supervised models on UCF101 [106] and HMDB51 [104]. The following work [167] further improves the performance by adopting more elaborate temporal augmentations and deeper networks. In addition to using simple visual input, some recent works [168,169] propose to leverage both audio and visual information of videos to perform self-supervised learning in a multi-modal manner. This could be a possible direction to leverage various modalities for more accurate HAR in the future.

5.2.2. Cross-domain visual HAR with video-based domain adaptation

In recent years, there has been a rise of research interest in video-based DA (VDA) for cross-domain visual HAR, which requires few or no labels for visual HAR in the target domain. Yet, despite the rise in interest, there are relatively few works on VDA compared to its counterpart in image-based DA. One major factor for the lack of relevant research is the fact that videos contain data with more modalities compared to images, which includes both temporal and spatiotemporal correlation features, thus complicating the overall adaptation process. As one of the primary works, Jamal et al. introduces Action Modeling on Latent Subspace (AMLS) and Deep Adversarial Action Adaptation (DAAA) [173] for VDA. In particular, the AMLS models target domain videos as a sequence of points on a latent subspace while adaptive kernels are learned between source and target domain points on the latent subspace, while DAAA is an end-to-end adversarial-based framework that adapts adversarial-based domain adaptation methods to videos by utilizing 3D-CNNs as feature generators.

Subsequent VDA approaches mostly adopt an adversarial DA approach, and improve on DAAA by focusing on improving source and target video alignment along the temporal direction. TA^{3}N [171] is proposed to align the video features along the temporal direction by applying attention mechanisms to video segments sampled across the temporal direction, whose features are extracted with TRN [70], thus resulting in the dynamic alignment of the different video segments. Similarly, Temporal Co-attention Network (TCoN) [174] is further proposed to align the distributions of video features using a novel cross-domain co-attention mechanism across the temporal dimension. Meanwhile, SAVA [175] is proposed to align video feature by utilizing the auxiliary task of clip order prediction [176]. Further, as an important part of video feature, the spatiotemporal correlation is also

Table 4
Comparison of current cross-domain HAR benchmark datasets.

Dataset	# Classes	# Train videos	# Test videos	Source of data	DA scenario	Year	Website
UCF-HMDB$_{small}$ [170]	5	832	339	UCF50, HMDB51	Closed-set	2014	Website
UCF-Olympic [170]	6	851	294	UCF50, Olympic Sports	Closed-set	2014	Website
UCF-HMDB$_{full}$ [171]	12	2278	931	UCF50, HMDB51	Closed-set	2019	Website
Kinetics-Gameplay [171]	12	46003	3995	Kinetics-600, Gameplay	Closed-set	2019	Website
EPIC Kitchens [172]	8	7935	2159	EPIC Kitchens	Closed-set	2020	Website
HMDB-ARID [7]	11	3058	1153	HMDB51, ARID	Closed-set	2021	Website
UCF-HMDB$_{partial}$ [146]	14	2304	476	UCF101, HMDB51	Partial	2021	Website
MiniKinetics-UCF [146]	45	20996	1106	MiniKinetics, UCF101	Partial	2021	Website
HMDB-ARID$_{partial}$ [146]	10	2712	540	HMDB51, ARID	Partial	2021	Website
Daily-DA [148]	8	16295	2654	ARID, HMDB51, Moments in Time, Kinetics-600	Multi-source/Closed-set	2021	Website
Sports-DA [148]	23	36003	4712	UCF101, Sports-1M, Kinetics-600	Multi-source/Closed-set	2021	Website

aligned in ACAN [145], which results in significant improvements. Besides aligning features with pure RGB input, MM-SADA [172] further exploits the multi-modal nature of videos by incorporating optical flow during feature alignment while adopting a self-supervision alignment approach across the different modalities. To further align features, contrastive learning [177,178] has recently been adopted, with contrastive loss obtained across both source and target video features as well as across features from different modalities.

The above VDA approaches all assume a scenario where the source and target domains share the same label space, denoted as the closed-set DA. Such a scenario may not apply in real-world applications, given the fact that domain adaptation is often used in scenarios where the source label space differs from the target one. For example, partial DA [146] concerns the case where the source label space subsumes the target one, whereas open-set DA [147] assumes the opposite. Meanwhile, multi-source DA [148] relaxes the assumption that source data are sampled from a single domain and match a uniform data distribution. More recently, with greater importance attached to data privacy, source-free DA [149,150] has been proposed where only the source model is provided to the target domain for adaption. All the aforementioned scenarios relax certain constraint assumed by the closed-set DA, and are more challenging with additional techniques required to avoid a negative transfer.

5.2.3. Cross-domain HAR datasets

Besides the complexity of aligning video data across the different domains, the lack of VDA research is also partly contributed by the fact that there are very limited cross-domain datasets available. Two primary cross-domain video datasets are the UCF-Olympic [170] dataset, built across UCF50 [105] and the Olympic Sports dataset [103]; and the UCF-HMDB$_{small}$ [170] dataset, built across UCF50 [105] and HMDB51 [104]. Both cross-domain datasets are of very small scale. To further facilitate research on VDA, larger cross-domain video datasets are introduced, with UCF-HMDB$_{full}$ [171] being the most commonly used benchmark dataset. Subsequently, several cross-domain datasets are proposed with extensive scale, including Kinetics-Gameplay [171] which bridges real-world videos with virtual-world videos, and EPIC Kitchens [172], an imbalanced cross-domain dataset. More recently, the HMDB-ARID dataset [145] is proposed to bridge videos across different illuminations (normal and dark environments) with significantly larger domain shifts compared to prior datasets.

To facilitate research on VDA in scenarios besides closed-set DA, several novel datasets have been proposed recently. Among which, UCF-HMDB$_{partial}$, MiniKinetics-UCF, and HMDB-ARID$_{partial}$ are series of datasets built towards partial DA [146]. For multi-source DA, the Daily-DA and Sports-DA datasets are proposed which are both constructed from a variety of public visual HAR datasets. The different datasets focus on diverse aspects in evaluating VDA methods, including large domain shifts and effectiveness on large-scale videos. It should be noted that both multi-source DA datasets could also be used in a closed-set DA scenario by limiting the source domain to a single domain. The major attributes (number of action classes, number of train and test videos, source of data, etc.) of the aforementioned datasets are outlined in Table 4. The number of train and test videos includes all domains associated.

5.3. Applications for cross-site HAR based on IoT sensors

For IoT-enabled sensor modality such as Lidar, activity data is the superposition of environment and human motions. Even though deep models have been trained with many human motions, they cannot foresee unseen environments and thus their performances may degrade. Worse, since high-quality large-scale HAR datasets are not common for other IoT-enabled sensor modalities (e.g. radar, lidar, WiFi, etc.), pre-training technique is not widely applied. Some recent works explore this direction and validate the effectiveness of transfer learning in WiFi-based HAR [179] and radar-based HAR [180]. However, the diversity of collection environments and the scale cannot stand in comparison with visual HAR datasets, such as Kinetic. As a result, these HAR models may only generalize well in a single domain, but fail in the face of new environments [8].

To deal with the issue of dynamic environments [25], domain adaptive HAR methods have been developed. For example, a pioneer research studies how domain adaptation facilitates cross-domain WiFi-based gesture recognition by adversarial DA [5,181]. More DA techniques are utilized to enhance the model robustness to different environments, such as entropy minimization [137]. Then DA also deals with the dataset bias caused by device settings and human heterogeneity for radar-based HAR [182] and Lidar-based HAR [133]. As cross-domain HAR scenarios are more challenging than standard DA benchmarks, they also promote the development of domain adaptation algorithms [23], such as DIRT-T [183] and CADA [184]. Though DA enables deep HAR models to adapt to new environments in an unsupervised manner, it still suffers from complicated computations. MobileDA is then proposed for efficient domain adaptation in the edge by knowledge distillation [185]. Though DA for HAR just starts to bloom, it is inevitable for robust HAR systems to integrate DA before super large data comes into existence for one modality.

6. Summary and discussion

We summarize the recent progress of deep learning and transfer learning for device-free HAR from three perspectives: deep architecture design, improvement by transfer learning, and noise interference.

6.1. Deep architecture design

Deep HAR models usually consist of a feature extractor and a classifier. The feasible classifiers include SVM, random forest, multilayer perceptron (MLP), etc. Whereas the feature extractor part should be able to extract both spatial and temporal features from sensor data, thus more crucial. From fruitful literature, it is seen that convolution layers are widely used for spatial information extraction. CNNs can act on image pixels or sensor data at one timestamp, and generate semantic feature maps with lower dimensions. For most sensor modalities, 2D-CNNs have already yielded satisfactory HAR results with friendly computational overhead. However, 2D-CNNs may not consider the temporal dynamics among sensor frames. Recurrent neural networks accept inputs from raw or feature sequences and extract short-term

or long-term patterns. 3D-CNN can complete spatial–temporal feature extraction simultaneously but it is limited by the convolution kernel size. With the development of deep architectures, transformers and non-local blocks with self-attention mechanisms further boost feature learning, achieving better results. In summary, for deep HAR model design, we advise to begin with 2D-CNN, and add on RNN or transformers according to the difficulty of tasks and the availability of computational resources.

6.2. Improvement by transfer learning

Pre-trained techniques have made significant progresses on visual HAR with several high-quality annotated datasets. For downstream HAR tasks, pre-trained parameters construct a discriminative prior and thus alleviate the difficulty of learning with limited data. Domain adaptation further enables model to adapt to new domains using unlabeled data, which enhances the viability of deep HAR models in the real world. In the transfer learning field, there are still some limitations. The first limitation lies in the data modality. Few efforts are made for non-video modalities due to the lack of massive datasets. The reason is that videos of human activities are easily obtained in the Internet but other modalities of data are expensive to collect. For a specific modality, a large-scale high-quality dataset contributes to a robust HAR model construction more than model revamps. It is expected that relevant industries and labs collect such datasets together to mitigate the apprehension of building HAR models with insufficient data. The second limitation is the out-of-distribution and issue that is the intrinsic shortcoming for deep models. Domain generalization has been developed to deal with the problem, and it is believed that meta learning also helps increase the generalization ability of deep learning models with a small amount of data. It is expected that deep HAR models should have better generalization ability.

6.3. Noise interference

Three categories of noises lead to the degrading performance of existing HAR models. The first one is the intrinsic sensor noise, especially for IoT passive sensors. For example, high-frequency noises and non-existent wave reflection occur for WiFi CSI and mmWave radar, respectively. Such noises shall be eliminated by pre-processing steps, relying on signal processing techniques such as Fourier or wavelet transformation. The second kind of noise stems from human heterogeneity. Different persons can perform different motions for the same activity category, resulting in challenges as deep models may never see such patterns. The third category of noise is the environmental dynamics, such as bad illumination that may dominate HAR sensor data. For visual HAR, this could be alleviated by image de-noising and enhancement, while for other modalities, the physical model helps grasp the dominant component that might be human activities, such as the Fresnel Zone [186] in radar sensing [187]. The last two noises can also be partially solved by unsupervised domain adaptation if the model can access to new data.

7. Grand challenge

(1) Learning with limited data. HAR data collection is expensive for the sake of volunteers and environments. How do we learn robust HAR models with limited data? Recent progress on deep learning aims to deal with such problems by zero/few-shot learning, meta-learning, and unsupervised learning. Zero/few-shot learning enables deep models to learn features from very few or none of the labeled samples, which conforms with the gesture recognition scenario [8]. Meta-learning aims to train a model on a variety of learning tasks and then learn new tasks using a small number of samples [188]. Learning cross HAR tasks helps deep models generalize well on more scenarios. Unsupervised learning resolves the situation when HAR data is easily collected without annotation [17], which will have an important impact after device-free sensors are deployed on a large scale.

(2) Lightweight HAR model. Most sensor modalities of HAR data are composed of high feature dimensions, and hence they require deep models that consume expensive computation resources. Building lightweight HAR models is also imperative considering the HAR task at the edge device or computation-constrained platforms. To this end, deep model compression techniques (e.g. network pruning [189], quantization, and distillation [190]) should be incorporated into HAR model design.

(3) Integration with physical model and interpretability. A huge advantage of deep models is that they can extract features automatically regardless of the lack of expert knowledge. However, physical or biological models could help us build more robust networks if available [191]. For WiFi-based respiration, physical model motivated completes better results that deep models have not attained [187]. It is expected that robust HAR systems could be developed by integrating physical and deep models. In contrast, if deep models achieve excellent performance on a HAR task, shall we explore the model interpretability and find more intuition?

(4) Multimodal HAR model. Deep multi-modal learning contributes to more sensible ways of combining and connecting two kinds of features, which overcomes the shortcomings of each modality. Imagine that if a camera-based HAR method is integrated with WiFi that can detect activities through walls or under dark circumstances, two modalities learn from each other's strength and close the HAR gap together. For example, Zou et al. integrate both visual and CSI data for accurate multi-modality activity recognition [192]. In the future, robust HAR systems should operate on multiple modalities.

(5) Security issue and federated HAR. Human activities are closely related to privacy. When HAR techniques are applied to smart home automation, such information might be divulged and utilized for illegal purposes. Furthermore, it is more dangerous to manipulate HAR data to tamper with the recognition results of HAR models, which is named adversarial attack [193]. To enhance the security of HAR systems, researches on federated HAR framework [194], HAR model robustness [195], and IoT cyber-security [196] constitute potential solutions.

(6) Complex activities. Many activities are not single motion but a series of meaningful behaviors. While existing deep models can recognize separate human activities, it is still tough to recognize complex activities and understand a kind of human behavior according to a series of human activities. Such fine-grained HAR can contribute more to other disciplines of research, such as Alzheimer's disease diagnosis and behavioral economics.

8. Conclusion

Device-free human activity recognition (HAR) plays a vital role in ubiquitous computing and mobile computing. This paper reviews the recent progress of deep learning and transfer learning techniques for device-free HAR models. We firstly compare various sensor modalities and then introduce deep HAR solutions. Superior to conventional methods, deep HAR models achieve robust results by automatic feature learning. Transfer learning further empowers deep models to work with limited data and adapt to new environments. We summarize the current research efforts and then propose grand challenges from the perspectives of HAR scenarios, learning schemes, system security, interpretability, and high-level applications. We hope that these conclusive and prospective directions can facilitate future HAR research, which helps build a robust, seamless, secure, and efficient HAR system for more significant applications.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] E. Kim, S. Helal, D. Cook, Human activity recognition and pattern discovery, IEEE Pervasive Comput. 9 (1) (2009) 48–53.

[2] N. Gupta, S.K. Gupta, R.K. Pathak, V. Jain, P. Rashidi, J.S. Suri, Human activity recognition in artificial intelligence framework: A narrative review, Artif. Intell. Rev. (2022) 1–54.

[3] W. Lin, M.-T. Sun, R. Poovandran, Z. Zhang, Human activity recognition for video surveillance, in: 2008 IEEE International Symposium on Circuits and Systems, IEEE, 2008, pp. 2737–2740.

[4] H. Zou, Y. Zhou, J. Yang, W. Gu, L. Xie, C. Spanos, WiFi-based human identification via convex tensor shapelet learning, in: AAAI Conference on Artificial Intelligence, 2018, pp. 1711–1719.

[5] H. Zou, J. Yang, Y. Zhou, L. Xie, C.J. Spanos, Robust WiFi-enabled device-free gesture recognition via unsupervised adversarial domain adaptation, in: 2018 27th International Conference on Computer Communication and Networks, ICCCN, IEEE, 2018, pp. 1–8.

[6] K. Lai, J. Konrad, P. Ishwar, A gesture-driven computer interface using kinect, in: 2012 IEEE Southwest Symposium on Image Analysis and Interpretation, IEEE, 2012, pp. 185–188.

[7] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, S. See, Arid: A new dataset for recognizing action in the dark, in: International Workshop on Deep Learning for Human Activity Recognition, Springer, 2021, pp. 70–84.

[8] J. Yang, H. Zou, Y. Zhou, L. Xie, Learning gestures from WiFi: A siamese recurrent convolutional architecture, IEEE Internet Things J. 6 (6) (2019) 10763–10772.

[9] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine recognition of human activities: A survey, IEEE Trans. Circuits Syst. Video Technol. 18 (11) (2008) 1473–1488.

[10] J. Wang, Y. Chen, S. Hao, X. Peng, L. Hu, Deep learning for sensor-based activity recognition: A survey, Pattern Recognit. Lett. 119 (2019) 3–11.

[11] H.F. Nweke, Y.W. Teh, M.A. Al-Garadi, U.R. Alo, Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges, Expert Syst. Appl. 105 (2018) 233–261.

[12] L.M. Dang, K. Min, H. Wang, M.J. Piran, C.H. Lee, H. Moon, Sensor-based and vision-based human activity recognition: A comprehensive survey, Pattern Recognit. 108 (2020) 107561.

[13] J. Yang, H. Zou, H. Jiang, L. Xie, Device-free occupant activity sensing using WiFi-enabled IoT devices for smart homes, IEEE Internet Things J. 5 (5) (2018) 3991–4002.

[14] Z. Chen, L. Zhang, Z. Cao, J. Guo, Distilling the knowledge from handcrafted features for human activity recognition, IEEE Trans. Ind. Inform. 14 (10) (2018) 4334–4342.

[15] Y. LeCun, D. Touresky, G. Hinton, T. Sejnowski, A theoretical framework for back-propagation, in: Proceedings of the 1988 Connectionist Models Summer School, Vol. 1, 1988, pp. 21–28.

[16] M. Gutmann, A. Hyvärinen, Noise-contrastive estimation: A new estimation principle for unnormalized statistical models, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[17] X. Chen, H. Fan, R. Girshick, K. He, Improved baselines with momentum contrastive learning, 2020, arXiv preprint arXiv:2003.04297.

[18] S.J. Pan, Q. Yang, et al., A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.

[19] K. He, R. Girshick, P. Dollár, Rethinking imagenet pre-training, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4918–4927.

[20] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436.

[21] L. Wang, W. Li, W. Li, L. Van Gool, Appearance-and-relation networks for video classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1430–1439.

[22] M. Wang, W. Deng, Deep visual domain adaptation: A survey, Neurocomputing 312 (2018) 135–153.

[23] J. Yang, H. Zou, Y. Zhou, Z. Zeng, L. Xie, Mind the discriminability: Asymmetric adversarial domain adaptation, in: European Conference on Computer Vision, Springer, 2020, pp. 589–606.

[24] D. Wang, J. Yang, W. Cui, L. Xie, S. Sun, Multimodal CSI-based human activity recognition using GANs, IEEE Internet Things J. (2021).

[25] J. Zhang, Z. Tang, M. Li, D. Fang, P. Nurmi, Z. Wang, CrossSense: Towards cross-site and large-scale WiFi sensing, in: Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, 2018, pp. 305–320.

[26] J. Wang, Y. Chen, L. Hu, X. Peng, S.Y. Philip, Stratified transfer learning for cross-domain activity recognition, in: 2018 IEEE International Conference on Pervasive Computing and Communications, PerCom, IEEE, 2018, pp. 1–10.

[27] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, X. Liu, A survey on deep learning for human activity recognition, ACM Comput. Surv. 54 (8) (2021) 1–34.

[28] G. Friedrich, Y. Yeshurun, Seeing people in the dark: Face recognition in infrared images, in: International Workshop on Biologically Motivated Computer Vision, Springer, 2002, pp. 348–359.

[29] H. Gao, B. Cheng, J. Wang, K. Li, J. Zhao, D. Li, Object classification using CNN-based fusion of vision and LiDAR in autonomous vehicle environment, IEEE Trans. Ind. Inform. 14 (9) (2018) 4224–4231.

[30] J. Roche, V. De-Silva, J. Hook, M. Moencks, A. Kondoz, A multimodal data processing system for LiDAR-based human activity recognition, IEEE Trans. Cybern. (2021).

[31] F. Luo, S. Poslad, E. Bodanese, Temporal convolutional networks for multiperson activity recognition using a 2-D LiDAR, IEEE Internet Things J. 7 (8) (2020) 7432–7442.

[32] X. Li, Y. He, X. Jing, A survey of deep learning-based human activity recognition in radar, Remote Sens. 11 (9) (2019) 1068.

[33] Y. Wang, H. Liu, K. Cui, A. Zhou, W. Li, H. Ma, m-Activity: Accurate and real-time human activity recognition via millimeter wave radar, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 8298–8302.

[34] A.D. Singh, S.S. Sandha, L. Garcia, M. Srivastava, Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar, in: Proceedings of the 3rd ACM Workshop on Millimeter-Wave Networks and Sensing Systems, 2019, pp. 51–56.

[35] R.H. Dodier, G.P. Henze, D.K. Tiller, X. Guo, Building occupancy detection through sensor belief networks, Energy Build. 38 (9) (2006) 1033–1043.

[36] M. Moghavvemi, L.C. Seng, Pyroelectric infrared sensor for intruder detection, in: 2004 IEEE Region 10 Conference TENCON 2004, Vol. 500, IEEE, 2004, pp. 656–659.

[37] J.D. Hewlett, M. Manic, C.G. Rieger, WESBES: A wireless embedded sensor for improving human comfort metrics using temporospatially correlated data, in: 2012 5th International Symposium on Resilient Control Systems, IEEE, 2012, pp. 31–36.

[38] L.M. Candanedo, V. Feldheim, Accurate occupancy detection of an office room from light, temperature, humidity and CO 2 measurements using statistical learning models, Energy Build. 112 (2016) 28–39.

[39] S. Wang, G. Zhou, A review on radio based activity recognition, Digit. Commun. Netw. 1 (1) (2015) 20–29.

[40] S. Sigg, S. Shi, Y. Ji, RF-based device-free recognition of simultaneously conducted activities, in: Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication, 2013, pp. 531–540.

[41] X. Wang, L. Gao, S. Mao, S. Pandey, CSI-based fingerprinting for indoor localization: A deep learning approach, IEEE Trans. Veh. Technol. 66 (1) (2017) 763–776.

[42] W. Wang, A.X. Liu, M. Shahzad, K. Ling, S. Lu, Device-free human activity recognition using commercial WiFi devices, IEEE J. Sel. Areas Commun. 35 (5) (2017) 1118–1131.

[43] J. Yang, H. Zou, H. Jiang, L. Xie, CareFi: Sedentary behavior monitoring system via commodity WiFi infrastructures, IEEE Trans. Veh. Technol. 67 (8) (2018) 7620–7629.

[44] X. Zheng, J. Wang, L. Shangguan, Z. Zhou, Y. Liu, Smokey: Ubiquitous smoking detection with commercial WiFi infrastructures, in: Computer Communications, IEEE INFOCOM 2016-the 35th Annual IEEE International Conference on, IEEE, 2016, pp. 1–9.

[45] W. Xi, J. Zhao, X.-Y. Li, K. Zhao, S. Tang, X. Liu, Z. Jiang, Electronic frog eye: Counting crowd using WiFi, in: IEEE INFOCOM 2014-IEEE Conference on Computer Communications, IEEE, 2014, pp. 361–369.

[46] H. Zou, Y. Zhou, J. Yang, W. Gu, L. Xie, C. Spanos, Freedetector: Device-free occupancy detection with commodity WiFi, in: 2017 IEEE International Conference on Sensing, Communication and Networking, SECON Workshops, IEEE, 2017, pp. 1–5.

[47] K. Bouchard, J. Maitre, C. Bertuglia, S. Gaboury, Activity recognition in smart homes using UWB radars, Procedia Comput. Sci. 170 (2020) 10–17.

[48] L. Cheng, A. Zhao, K. Wang, H. Li, Y. Wang, R. Chang, Activity recognition and localization based on UWB indoor positioning system and machine learning, in: 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON, IEEE, 2020, pp. 0528–0533.

[49] Apple invents iBeacon Version 2 using ultra-wide band radio technology, 2019, https://www.patentlyapple.com/patently-apple/2019/01/apple-invents-ibeacon-version-2-using-ultrawide-band-radio-technology.html. (Online; Accessed 4 September 2019).

[50] A.H. Marblestone, G. Wayne, K.P. Kording, Toward an integration of deep learning and neuroscience, Front. Comput. Neurosci. 10 (2016) 94.

[51] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.

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

[53] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: IEEE CVPR 2015, 2015, pp. 1–9.

[54] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[55] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500.

[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.

[57] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[58] J.P. Chiu, E. Nichols, Named entity recognition with bidirectional LSTM-CNNs, Trans. Assoc. Comput. Linguist. 4 (2016) 357–370.

[59] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal LSTM with trust gates for 3D human action recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 816–833.

[60] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1933–1941.

[61] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks for action recognition in videos, IEEE Trans. Pattern Anal. Mach. Intell. 41 (11) (2018) 2740–2755.

[62] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[63] Y. Xu, J. Yang, K. Mao, J. Yin, S. See, Exploiting inter-frame regional correlation for efficient action recognition, Expert Syst. Appl. 178 (2021) 114829.

[64] H. Cao, Y. Xu, J. Yang, K. Mao, J. Yin, S. See, Effective action recognition with embedded key point shifts, Pattern Recognit. 120 (2021) 108172.

[65] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, 2014, arXiv preprint arXiv:1406.2199.

[66] M.A. Goodale, A.D. Milner, Separate visual pathways for perception and action, Trends Neurosci. 15 (1) (1992) 20–25.

[67] C. Zach, T. Pock, H. Bischof, A duality based approach for realtime TV-L 1 optical flow, in: Joint Pattern Recognition Symposium, Springer, 2007, pp. 214–223.

[68] R. Christoph, F.A. Pinz, Spatiotemporal residual networks for video action recognition, Adv. Neural Inf. Process. Syst. (2016) 3468–3476.

[69] Z. Lan, Y. Zhu, A.G. Hauptmann, S. Newsam, Deep local video feature for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1–7.

[70] B. Zhou, A. Andonian, A. Oliva, A. Torralba, Temporal relational reasoning in videos, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 803–818.

[71] L. Wang, Y. Qiao, X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4305–4314.

[72] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, T. Brox, Flownet: Learning optical flow with convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2758–2766.

[73] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, D. Katabi, Through-wall human pose estimation using radio signals, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7356–7365.

[74] W. Li, D. Chen, Z. Lv, Y. Yan, D. Cosker, Learn to model blurry motion via directional similarity and filtering, Pattern Recognit. 75 (2018) 327–338.

[75] L. Fan, W. Huang, C. Gan, S. Ermon, B. Gong, J. Huang, End-to-end learning of motion representation for video understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6016–6025.

[76] A. Piergiovanni, M.S. Ryoo, Representation flow for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9945–9953.

[77] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.

[78] Y. Shi, Y. Tian, Y. Wang, W. Zeng, T. Huang, Learning long-term dependencies for action recognition with a biologically-inspired deep network, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 716–725.

[79] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014, arXiv preprint arXiv:1412.3555.

[80] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2012) 221–231.

[81] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.

[82] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.

[83] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A closer look at spatiotemporal convolutions for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459.

[84] K. Hara, H. Kataoka, Y. Satoh, Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet? in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6546–6555.

[85] S. Xie, C. Sun, J. Huang, Z. Tu, K. Murphy, Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 305–321.

[86] Z. Qiu, T. Yao, T. Mei, Learning spatio-temporal representation with pseudo-3D residual networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5533–5541.

[87] D. Tran, H. Wang, L. Torresani, M. Feiszli, Video classification with channel-separated convolutional networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5552–5561.

[88] Y. Chen, Y. Kalantidis, J. Li, S. Yan, J. Feng, Multi-fiber networks for video recognition, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 352–367.

[89] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6202–6211.

[90] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

[91] K. Yue, M. Sun, Y. Yuan, F. Zhou, E. Ding, F. Xu, Compact generalized non-local network, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 6511–6520.

[92] Y. Chen, Y. Kalantidis, J. Li, S. Yan, J. Feng, A^2-Nets: Double attention networks, 2018, arXiv preprint arXiv:1810.11579.

[93] Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, GCNET: Non-local networks meet squeeze-excitation networks and beyond, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.

[94] Y. Xu, H. Cao, J. Yang, K. Mao, J. Yin, S. See, PNL: Efficient long-range dependencies extraction with pyramid non-local module for action recognition, Neurocomputing 447 (2021) 282–293.

[95] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[96] R. Girdhar, J. Carreira, C. Doersch, A. Zisserman, Video action transformer network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 244–253.

[97] Y. Zhang, X. Li, C. Liu, B. Shuai, Y. Zhu, B. Brattoli, H. Chen, I. Marsic, J. Tighe, VIDTR: Video transformer without convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13577–13587.

[98] D. Neimark, O. Bar, M. Zohar, D. Asselmann, Video transformer network, 2021, arXiv preprint arXiv:2102.00719.

[99] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: A local SVM approach, in: Proceedings of the 17th International Conference on Pattern Recognition, 2004, Vol. 3, ICPR 2004, IEEE, 2004, pp. 32–36.

[100] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, IEEE Trans. Pattern Anal. Mach. Intell. 29 (12) (2007) 2247–2253.

[101] D. Weinland, E. Boyer, R. Ronfard, Action recognition from arbitrary views using 3D exemplars, in: 2007 IEEE 11th International Conference on Computer Vision, IEEE, 2007, pp. 1–7.

[102] M. Marszalek, I. Laptev, C. Schmid, Actions in context, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 2929–2936.

[103] J.C. Niebles, C.-W. Chen, L. Fei-Fei, Modeling temporal structure of decomposable motion segments for activity classification, in: European Conference on Computer Vision, Springer, 2010, pp. 392–405.

[104] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: A large video database for human motion recognition, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2556–2563.

[105] K.K. Reddy, M. Shah, Recognizing 50 human action categories of web videos, Mach. Vis. Appl. 24 (5) (2013) 971–981.

[106] K. Soomro, A.R. Zamir, M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, 2012, arXiv preprint arXiv:1212.0402.

[107] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset, 2017, arXiv preprint arXiv:1705.06950.
[108] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al., The "something something" video database for learning and evaluating visual common sense, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5842–5850.
[109] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S.A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfruend, C. Vondrick, et al., Moments in time dataset: One million videos for event understanding, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1–8.
[110] T. Chen, W. Yin, X.S. Zhou, D. Comaniciu, T.S. Huang, Total variation models for variable lighting face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 28 (9) (2006) 1519–1524.
[111] H. Shim, J. Luo, T. Chen, A subspace model-based approach to face relighting under unknown lighting and poses, IEEE Trans. Image Process. 17 (8) (2008) 1331–1341.
[112] H. Han, S. Shan, X. Chen, W. Gao, A comparative study on illumination preprocessing in face recognition, Pattern Recognit. 46 (6) (2013) 1691–1699.
[113] Z. Chen, C. Cai, T. Zheng, J. Luo, J. Xiong, X. Wang, RF-based human activity recognition using signal adapted convolutional neural network, IEEE Trans. Mob. Comput. (2021).
[114] D. Singh, E. Merdivan, I. Psychoula, J. Kropf, S. Hanke, M. Geist, A. Holzinger, Human activity recognition using recurrent neural networks, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, 2017, pp. 267–274.
[115] S. Chung, J. Lim, K.J. Noh, G. Kim, H. Jeong, Sensor data acquisition and multimodal sensor fusion for human activity recognition using deep learning, Sensors 19 (7) (2019) 1716.
[116] X. Li, Y. He, F. Fioranelli, X. Jing, Semisupervised human activity recognition with radar micro-doppler signatures, IEEE Trans. Geosci. Remote Sens. (2021).
[117] P. Gong, C. Wang, L. Zhang, Mmpoint-GNN: Graph neural network with dynamic edges for human activity recognition through a millimeter-wave radar, in: 2021 International Joint Conference on Neural Networks, IJCNN, IEEE, 2021, pp. 1–7.
[118] Z. Meng, S. Fu, J. Yan, H. Liang, A. Zhou, S. Zhu, H. Ma, J. Liu, N. Yang, Gait recognition for co-existing multiple people using millimeter wave sensing, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, no. 01, 2020, pp. 849–856.
[119] J. Maitre, K. Bouchard, C. Bertuglia, S. Gaboury, Recognizing activities of daily living from UWB radars and deep learning, Expert Syst. Appl. 164 (2021) 113994.
[120] Y. Xie, Z. Li, M. Li, Precise power delay profiling with commodity WiFi, in: Proceedings of the 21st Annual International Conference on Mobile Computing and Networking, ACM, 2015, pp. 53–64.
[121] H. Zou, Y. Zhou, J. Yang, H. Jiang, L. Xie, C.J. Spanos, Deepsense: Device-free human activity recognition via autoencoder long-term recurrent convolutional network, in: 2018 IEEE International Conference on Communications, ICC, IEEE, 2018, pp. 1–6.
[122] H. Zou, Y. Zhou, J. Yang, W. Gu, L. Xie, C. Spanos, Poster: WiFi-based device-free human activity recognition via automatic representation learning, in: Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking, ACM, 2017, pp. 606–608.
[123] H. Zou, Y. Zhou, J. Yang, H. Jiang, L. Xie, C.J. Spanos, WiFi-enabled device-free gesture recognition for smart home automation, in: 2018 IEEE 14th International Conference on Control and Automation, ICCA, IEEE, 2018, pp. 476–481.
[124] H. Zou, Y. Zhou, J. Yang, C.J. Spanos, Device-free occupancy detection and crowd counting in smart buildings with WiFi-enabled IoT, Energy Build. 174 (2018) 309–322.
[125] H. Zou, Y. Zhou, J. Yang, W. Gu, L. Xie, C. Spanos, Freecount: Device-free crowd counting with commodity WiFi, in: GLOBECOM 2017-2017 IEEE Global Communications Conference, IEEE, 2017, pp. 1–6.
[126] Y. Ma, G. Zhou, S. Wang, H. Zhao, W. Jung, Signfi: Sign language recognition using WiFi, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2 (1) (2018) 1–21.
[127] S. Chen, W. Wang, S.J. Pan, Metaquant: Learning to quantize by learning to penetrate non-differentiable quantization, Adv. Neural Inf. Process. Syst. 32 (2019) 3916–3926.
[128] L. Yao, Q.Z. Sheng, X. Li, T. Gu, M. Tan, X. Wang, S. Wang, W. Ruan, Compressive representation for device-free activity recognition with passive RFID signal strength, IEEE Trans. Mob. Comput. 17 (2) (2017) 293–306.
[129] D. Hao, J. Tian, D. Yongpeng, X. Zhuo, A compact human activity classification model based on transfer learned network pruning, in: IET International Radar Conference, Vol. 2020, IET IRC 2020, IET, 2020, pp. 1488–1492.
[130] J. Guo, J. Liu, D. Xu, JointPruning: Pruning networks along multiple dimensions for efficient point cloud processing, IEEE Trans. Circuits Syst. Video Technol. (2021).
[131] J. Yang, X. Chen, H. Zou, D. Wang, Q. Xu, L. Xie, EfficientFi: Towards large-scale lightweight WiFi sensing via CSI compression, IEEE Internet Things J. (2022).
[132] M. Moencks, V. De Silva, J. Roche, A. Kondoz, Adaptive feature processing for robust human activity recognition on a novel multi-modal dataset, 2019, arXiv preprint arXiv:1901.02858.
[133] M.A.U. Alam, F. Mazzoni, M.M. Rahman, J. Widberg, LAMAR: LiDAR based multi-inhabitant activity recognition, in: MobiQuitous 2020-17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, 2020, pp. 1–9.
[134] T. Van Kasteren, A. Noulas, G. Englebienne, B. Kröse, Accurate activity recognition in a home setting, in: Proceedings of the 10th International Conference on Ubiquitous Computing, 2008, pp. 1–9.
[135] S. Yousefi, H. Narui, S. Dayal, S. Ermon, S. Valaee, A survey on behavior recognition using WiFi channel state information, IEEE Commun. Mag. 55 (10) (2017) 98–104.
[136] L. Guo, L. Wang, J. Liu, W. Zhou, B. Lu, HuAc: Human activity recognition using crowdsourced WiFi signals and skeleton data, Wirel. Commun. Mob. Comput. 2018 (2018).
[137] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, Z. Yang, Zero-effort cross-domain gesture recognition with Wi-Fi, in: Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services, 2019, pp. 313–325.
[138] S.Z. Gurbuz, M.M. Rahman, E. Kurtoglu, T. Macks, F. Fioranelli, Cross-frequency training with adversarial learning for radar micro-Doppler signature classification (Rising Researcher), in: Radar Sensor Technology XXIV, Vol. 11408, International Society for Optics and Photonics, 2020, p. 114080A.
[139] Z. Zhengliang, Y. Degui, Z. Junchao, T. Feng, Dataset of human motion status using IR-UWB through-wall radar, J. Syst. Eng. Electron. 32 (5) (2021) 1083–1096.
[140] S.J. Pan, J.T. Kwok, Q. Yang, et al., Transfer learning via dimensionality reduction, in: AAAI, 2008.
[141] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, in: International Conference on Machine Learning, 2015, pp. 97–105.
[142] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: ICML, 2015, pp. 1180–1189.
[143] S. Xie, Z. Zheng, L. Chen, C. Chen, Learning semantic representations for unsupervised domain adaptation, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 80, PMLR, Stockholmsmässan, Stockholm Sweden, 2018, pp. 5423–5432.
[144] Y. Grandvalet, Y. Bengio, Semi-supervised learning by entropy minimization, in: Advances in Neural Information Processing Systems, 2005, pp. 529–536.
[145] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, S. See, Aligning correlation information for domain adaptation in action recognition, 2021, arXiv preprint arXiv:2107.04932.
[146] Y. Xu, J. Yang, H. Cao, Z. Chen, Q. Li, K. Mao, Partial video domain adaptation with partial adversarial temporal attentive network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9332–9341.
[147] P.P. Busto, A. Iqbal, J. Gall, Open set domain adaptation for image and action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2018) 413–429.
[148] Y. Xu, J. Yang, H. Cao, K. Wu, M. Wu, R. Zhao, Z. Chen, Multi-source video domain adaptation with temporal attentive moment alignment, 2021, arXiv preprint arXiv:2109.09964.
[149] J. Liang, D. Hu, J. Feng, Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation, in: International Conference on Machine Learning, PMLR, 2020, pp. 6028–6039.
[150] J. Liang, D. Hu, Y. Wang, R. He, J. Feng, Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer, IEEE Trans. Pattern Anal. Mach. Intell. (2021).
[151] N. Srivastava, E. Mansimov, R. Salakhutdinov, Unsupervised learning of video representations using LSTMs, in: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Vol. 37, ICML '15, JMLR.org, 2015, pp. 843–852.
[152] C. Vondrick, H. Pirsiavash, A. Torralba, Generating videos with scene dynamics, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS '16, Curran Associates Inc., Red Hook, NY, USA, 2016, pp. 613–621.
[153] X. Li, S. Liu, S. De Mello, X. Wang, J. Kautz, M.-H. Yang, Joint-task self-supervised learning for temporal correspondence, Adv. Neural Inf. Process. Syst. 32 (2019).
[154] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, C. Feichtenhofer, Masked feature prediction for self-supervised visual pre-training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14668–14678.
[155] D. Yuan, X. Chang, P.-Y. Huang, Q. Liu, Z. He, Self-supervised deep correlation tracking, IEEE Trans. Image Process. 30 (2020) 976–985.

[156] C. Gan, B. Gong, K. Liu, H. Su, L.J. Guibas, Geometry guided convolutional neural networks for self-supervised video representation learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5589–5597.

[157] Z. Luo, B. Peng, D.-A. Huang, A. Alahi, L. Fei-Fei, Unsupervised learning of long-term motion dynamics for videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2203–2212.

[158] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, C. Feichtenhofer, Masked feature prediction for self-supervised visual pre-training, 2021, arXiv preprint arXiv: 2112.09133.

[159] I. Misra, C.L. Zitnick, M. Hebert, Shuffle and learn: Unsupervised learning using temporal order verification, in: European Conference on Computer Vision, Springer, 2016, pp. 527–544.

[160] B. Fernando, H. Bilen, E. Gavves, S. Gould, Self-supervised video representation learning with odd-one-out networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3636–3645.

[161] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, Y. Zhuang, Self-supervised spatiotemporal learning via video clip order prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10334–10343.

[162] J. Wang, J. Jiao, Y.-H. Liu, Self-supervised video representation learning by pace prediction, in: European Conference on Computer Vision, Springer, 2020, pp. 504–521.

[163] Y. Yao, C. Liu, D. Luo, Y. Zhou, Q. Ye, Video playback rate perception for self-supervised spatio-temporal representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020.

[164] S. Jenni, G. Meishvili, P. Favaro, Video representation learning by recognizing temporal transformations, 2020, arXiv preprint arXiv:2007.10730.

[165] H.-Y. Lee, J.-B. Huang, M. Singh, M.-H. Yang, Unsupervised representation learning by sorting sequences, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 667–676.

[166] T. Yao, Y. Zhang, Z. Qiu, Y. Pan, T. Mei, SeCo: Exploring sequence supervision for unsupervised representation learning, in: 35th AAAI Conference on Artificial Intelligence, 2021.

[167] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, Y. Cui, Spatiotemporal contrastive video representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6964–6974.

[168] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelovic, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, A. Zisserman, Self-supervised MultiModal versatile networks, NeurIPS 2 (6) (2020) 7.

[169] P. Morgado, N. Vasconcelos, I. Misra, Audio-visual instance discrimination with cross-modal agreement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12475–12486.

[170] W. Sultani, I. Saleemi, Human action recognition across datasets by foreground-weighted histogram decomposition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 764–771.

[171] M.-H. Chen, Z. Kira, G. AlRegib, J. Yoo, R. Chen, J. Zheng, Temporal attentive alignment for large-scale video domain adaptation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6321–6330.

[172] J. Munro, D. Damen, Multi-modal domain adaptation for fine-grained action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 122–132.

[173] A. Jamal, V.P. Namboodiri, D. Deodhare, K. Venkatesh, Deep domain adaptation in action space, in: BMVC, Vol. 2, 2018, p. 4.

[174] B. Pan, Z. Cao, E. Adeli, J.C. Niebles, Adversarial cross-domain action recognition with co-attention, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 11815–11822.

[175] J. Choi, G. Sharma, S. Schulter, J.-B. Huang, Shuffle and attend: Video domain adaptation, in: European Conference on Computer Vision, Springer, 2020, pp. 678–695.

[176] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, Y. Zhuang, Self-supervised spatiotemporal learning via video clip order prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10334–10343.

[177] X. Song, S. Zhao, J. Yang, H. Yue, P. Xu, R. Hu, H. Chai, Spatio-temporal contrastive domain adaptation for action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9787–9795.

[178] D. Kim, Y.-H. Tsai, B. Zhuang, X. Yu, S. Sclaroff, K. Saenko, M. Chandraker, Learning cross-modal contrastive features for video domain adaptation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13618–13627.

[179] Y. Zhang, Y. Zheng, K. Qian, G. Zhang, Y. Liu, C. Wu, Z. Yang, WIDAR3. 0: Zero-effort cross-domain gesture recognition with Wi-Fi, IEEE Trans. Pattern Anal. Mach. Intell. (2021).

[180] Y. Kim, J. Park, T. Moon, Classification of micro-Doppler signatures of human aquatic activity through simulation and measurement using transferred learning, in: Radar Sensor Technology XXI, Vol. 10188, International Society for Optics and Photonics, 2017, p. 101880V.

[181] H. Zou, J. Yang, Y. Zhou, C.J. Spanos, Joint adversarial domain adaptation for resilient WiFi-enabled device-free gesture recognition, in: 2018 17th IEEE International Conference on Machine Learning and Applications, ICMLA, IEEE, 2018, pp. 202–207.

[182] Y. Lang, Q. Wang, Y. Yang, C. Hou, D. Huang, W. Xiang, Unsupervised domain adaptation for micro-Doppler human motion classification via feature fusion, IEEE Geosci. Remote Sens. Lett. 16 (3) (2018) 392–396.

[183] R. Shu, H.H. Bui, H. Narui, S. Ermon, A DIRT-T approach to unsupervised domain adaptation, in: Proc. 6th International Conference on Learning Representations, 2018.

[184] H. Zou, Y. Zhou, J. Yang, H. Liu, H.P. Das, C.J. Spanos, Consensus adversarial domain adaptation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 5997–6004.

[185] J. Yang, H. Zou, S. Cao, Z. Chen, L. Xie, MobileDA: Towards edge domain adaptation, IEEE Internet Things J. (2020).

[186] J. Lindsey, The fresnel zone and its interpetive significance, Leading Edge 8 (10) (1989) 33–39.

[187] D. Wu, Y. Zeng, F. Zhang, D. Zhang, WiFi CSI-based device-free sensing: From fresnel zone model to CSI-ratio model, CCF Trans. Pervasive Comput. Interact. (2021) 1–15.

[188] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1126–1135.

[189] S. Chen, W. Wang, S.J. Pan, Cooperative pruning in cross-domain deep neural network compression, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, 2019, pp. 2102–2108.

[190] A. Polino, R. Pascanu, D.-A. Alistarh, Model compression via distillation and quantization, in: 6th International Conference on Learning Representations, 2018.

[191] H.A. Elmarakeby, J. Hwang, D. Liu, S.H. AlDubayan, K. Salari, C. Richter, T.E. Arnoff, J. Park, W.C. Hahn, E. Van Allen, Biologically informed deep neural network for prostate cancer classification and discovery, 2020, BioRxiv.

[192] H. Zou, J. Yang, H. Prasanna Das, H. Liu, Y. Zhou, C.J. Spanos, Wifi and vision multimodal learning for accurate and robust device-free human activity recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[193] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, 2017, arXiv preprint arXiv:1706.06083.

[194] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, H. Yu, Federated learning, Synth. Lect. Artif. Intell. Mach. Learn. 13 (3) (2019) 1–207.

[195] D. Hendrycks, K. Lee, M. Mazeika, Using pre-training can improve model robustness and uncertainty, in: International Conference on Machine Learning, PMLR, 2019, pp. 2712–2721.

[196] Y. Lu, L. Da Xu, Internet of Things (IoT) cybersecurity research: A review of current research topics, IEEE Internet Things J. 6 (2) (2018) 2103–2115.

Jianfei Yang received the B.Eng. from the School of Data and Computer Science, Sun Yat-sen University in 2016, and the Ph.D. degree from Nanyang Technological University (NTU), Singapore in 2021. He received the best Ph.D. thesis award from NTU. He used to work as a senior research engineer at the University of California, Berkeley. His research focuses on Artificial Intelligence of Things (AIoT), such as wireless sensing and computer vision based on deep learning and transfer learning. He won many International AI challenges in computer vision and interdisciplinary research fields. Currently, he is a Presidential Postdoctoral Research Fellow and an independent principal investigator at NTU.

Yuecong Xu received the B.Eng. from the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore in 2017, and the Ph.D. degree from Nanyang Technological University (NTU), Singapore in 2021. He was the receiver of the Nanyang President's Graduate Scholarship. His research focuses on video understanding and analysis based on deep learning and transfer learning. He was the co-organizer of the UG2+ Challenge for Computational Photography and Visual Recognition, held in conjunction with CVPR 2021 and CVPR 2022. Currently, he is a Research Scientist at the Institute for Infocomm Research, A*STAR, Singapore and a lecturer at NTU.

Haozhi Cao received his B.Eng. from the School of Electrical Engineering and Automation, Wuhan University in 2019, and his M.Eng. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore in 2021. He is currently a Ph.D. student in the School of Electrical and Electronic Engineering, NTU. He is also working as a Research Associate at Centre for Advanced Robotics Technology (CARTIN), NTU. His research interests include deep learning with applications in video understanding, transfer learning and multi-modal learning.

Han Zou received the B.Eng. (First Class Honors) and Ph.D. degrees in Electrical and Electronic Engineering from the Nanyang Technological University, Singapore, in 2012 and 2016, respectively. He is currently a Postdoctoral Scholar with the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley, CA, USA. His research interests include ubiquitous computing, statistical learning, signal processing and data analytics with applications in occupancy sensing, indoor localization, smart buildings and Internet of Things.

Lihua Xie received the B.E. and M.E. degrees in electrical engineering from Nanjing University of Science and Technology in 1983 and 1986, respectively, and the Ph.D. degree in electrical engineering from the University of Newcastle, Australia, in 1992. Since 1992, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he is currently a professor and served as the Head of Division of Control and Instrumentation from July 2011 to June 2014. He held teaching appointments in the Department of Automatic Control, Nanjing University of Science and Technology from 1986 to 1989 and Changjiang Visiting Professorship with South China University of Technology from 2006 to 2011.

Dr Xie's research interests include robust control and estimation, networked control systems, multi-agent control and unmanned systems. He has served as an editor of IET Book Series in Control and an Associate Editor of a number of journals including IEEE Transactions on Automatic Control, Automatica, IEEE Transactions on Control Systems Technology, and IEEE Transactions on Circuits and Systems-II. Dr Xie is a Fellow of Academy of Engineering Singapore, Fellow of IEEE, Fellow of IFAC, and Fellow of Chinese Automation Association.