Human Action Recognition from Various Data Modalities: A Review

Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu

Abstract—Human Action Recognition (HAR) aims to understand human behavior and assign a label to each action. It has a wide range of applications, and therefore has been attracting increasing attention in the field of computer vision. Human actions can be represented using various data modalities, such as RGB, skeleton, depth, infrared, point cloud, event stream, audio, acceleration, radar, and WiFi signal, which encode different sources of useful yet distinct information and have various advantages depending on the application scenarios. Consequently, lots of existing works have attempted to investigate different types of approaches for HAR using various modalities. In this paper, we present a comprehensive survey of recent progress in deep learning methods for HAR based on the type of input data modality. Specifically, we review the current mainstream deep learning methods for single data modalities and multiple data modalities, including the fusion-based and the co-learning-based frameworks. We also present comparative results on several benchmark datasets for HAR, together with insightful observations and inspiring future research directions.

Index Terms—Human Action Recognition, Deep Learning, Data Modality, Single Modality, Multi-modality.

✦

1 INTRODUCTION

HUMAN Action Recognition (HAR), i.e., recognizing and understanding human actions, is crucial for a number of real-world applications. It can be used in visual surveillance systems [1] to identify dangerous human activities, and autonomous navigation systems [2] to perceive human behaviors for safe operation. Besides, it is important for a number of other applications, e.g., video retrieval [3], human-robot interaction [4], and entertainment [5].

In the early days, most of the works focused on using RGB or gray-scale videos as input for HAR [6], due to their popularity and easy access. Recent years have witnessed an emergence of works [7], [8], [9], [10], [11], [12], [13], [14], [15], [16] using other data modalities, such as skeleton, depth, infrared sequence, point cloud, event stream, audio, acceleration, radar, and WiFi for HAR. This is mainly due to the development of different kinds of accurate and affordable sensors, and the distinct advantages of different modalities for HAR depending on the application scenarios.

Specifically, according to the visibility, data modalities can be roughly divided into two categories, namely, visual modalities and non-visual modalities. As shown in Table 1, RGB, skeleton, depth, infrared sequence, point cloud, and event stream are visually "intuitive" for representing human actions, and can be seen as visual modalities. Generally, visual modalities are very effective for HAR. Among them, RGB video data is the most common data type for HAR, which has been widely used in surveillance and monitoring systems. Skeleton data encodes the trajectories of human body joints. It is succinct and efficient for HAR when the action performing does not involve objects or scene context. Point cloud and depth data, which capture the 3D structure and distance information, are popularly used for HAR in robot navigation and self-driving applications. In addition, infrared data can be utilized for HAR in even dark environments, while the event stream keeps the foreground movement of the human subjects and avoids much visual redundancy, making it suitable for HAR as well.

Meanwhile, audio, acceleration, radar, and WiFi, etc. are non-visual modalities, i.e., they are not visually "intuitive" for representing human behaviors. Nevertheless, these modalities can also be used for HAR in some scenarios which require the protection of privacy of subjects. Among them, audio data is suitable for locating actions in the temporal sequences, while acceleration data can be adopted for fine-grained HAR. Besides, as a non-visual modality, radar data can even be used for through-wall HAR. More details of different modalities are discussed in Section 2.

Single modality-based HAR has been extensively investigated in the past decades [6], [17], [18], [19], [20], [21], [22]. However, since different modalities have different strengths and limitations for HAR, the fusion of multiple data modalities and the transfer of knowledge across modalities to enhance the accuracy and robustness of HAR, have also received great attention recently [23], [24]. More specifically, fusion means the combination of the information of two or more modalities to recognize actions. For example, audio data can serve as complementary information of the visual modalities to distinguish the actions "putting a plate" and "putting a bag" [25]. Besides fusion, some other methods have also exploited co-learning, i.e., transferring knowledge across different modalities to strengthen the robustness of HAR models. For example, in [26], the skeleton data serves as an auxiliary modality enabling the model to extract more discriminative features from RGB videos for HAR.

- Z. Sun and J. Liu are with SUTD, Singapore.
  E-mail: sunembrace@whu.edu.cn; jun_liu@sutd.edu.sg.
- Q. Ke is with The University of Melbourne, Australia.
  E-mail: qiuhong.ke@unimelb.edu.au.
- H. Rahmani is with Lancaster University, UK.
  E-mail: h.rahmani@lancaster.ac.uk.
- M. Bennamoun is with The University of Western Australia, Australia.
  E-mail: mohammed.bennamoun@uwa.edu.au.
- G. Wang is with Alibaba Group, China.
  E-mail: wanggang@ntu.edu.sg.

Considering the significance of using different single modalities for HAR and also leveraging their complementary characteristics for fusion and co-learning-based HAR, we review existing HAR methods from the perspective of data modalities. Specifically, we review the mainstream deep learning architectures that use single data modalities for HAR, and also the methods taking advantage of multiple modalities for enhanced HAR. Note that in this survey, we mainly focus on HAR from short and trimmed video segments, i.e., each video contains only one action instance. We also report and discuss the benchmark datasets. The main contributions of this review are summarized as follows.

(1) To the best of our knowledge, this is the first survey paper that comprehensively reviews the HAR methods from the perspective of various data modalities, including RGB, depth, skeleton, infrared sequence, point cloud, event stream, audio, acceleration, radar, and WiFi.

(2) We comprehensively review the multi-modality-based HAR methods, and categorize them into two types, namely, multi-modality fusion-based approaches and cross-modality co-learning-based approaches.

(3) We focus on reviewing the more recent and advanced deep learning methods for HAR, and hence provide the readers with the state-of-the-art approaches.

(4) We provide comprehensive comparisons of existing methods and their performance on several benchmark datasets (e.g., Tables 2, 3, 4, 5), with brief summaries and insightful discussions.

The rest of this paper is organized as follows. Section 2 reviews HAR methods using different single data modalities. Section 3 introduces multi-modality HAR methods. The benchmark datasets are listed in Section 4. Finally, potential future development of HAR is discussed in Section 5.
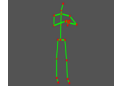
2 SINGLE MODALITY
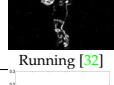
Different modalities can have different characteristics with corresponding advantages and disadvantages, as shown in Table 1. Hence numerous works [7], [8], [9], [10], [11], [12], [13], [14], [15], [16] exploited various modalities for HAR. Below we review the methods that use single data modalities, including RGB, skeleton, depth, infrared, point cloud, event stream, audio, acceleration, radar, and WiFi, for HAR.

2.1 RGB MODALITY

The RGB modality generally refers to images or videos (sequences of images) captured by RGB cameras which aim to recreate what human eyes see. RGB data is usually easy to collect, and it contains rich appearance information of the captured scene context. RGB-based HAR has a wide range of applications, such as visual surveillance [1], autonomous navigation [2], and sport analysis [36]. However, action recognition from RGB data is often challenging, owing 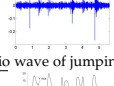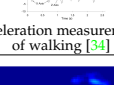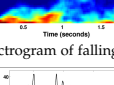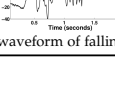to the variations of backgrounds, viewpoints, scales of humans, and illumination conditions. Besides, RGB videos have generally large data sizes, leading to high computational costs when modeling the spatio-temporal context for HAR.

In this subsection, we review the methods which use the RGB modality for HAR. Specifically, since videos contain the temporal dynamics of human motions that are often crucial

TABLE 1
Action samples of different data modalities (with pros and cons).

Modality		Example	Pros	Cons
Visual Modality	RGB	Hand-waving [27]	· Provide rich appearance information · Easy to obtain and operate · Wide range of applications	· Sensitive to viewpoint · Sensitive to background · Sensitive to illumination
	3D Skeleton	Looking at watch [28]	· Provide 3D structural information of subject pose · Simple yet informative · Insensitive to viewpoint · Insensitive to background	· Lack of appearance information · Lack of detailed shape information · Noisy
	Depth	Mopping floor [29]	· Provide 3D structural information · Provide geometric shape information	· Lack of color and texture information · Limited workable distance
	Infrared Sequence	Pushing [30]	· Workable in dark environments	· Lack of color and texture information · Susceptible to sunlight
	Point Cloud	Bending over [31]	· Provide 3D information · Provide geometric shape information · Insensitive to viewpoint	· Lack of color and texture information · High computational complexity
	Event Stream	Running [32]	· Avoid much visual redundancy · High dynamic range · No motion blur	· Asynchronous output · Spatio-temporally sparse · Capturing device is relatively expensive
Non-visual Modality	Audio	Audio wave of jumping [33]	· Easy to locate actions in temporal sequence	· Lack of appearance information
	Acceleration	Acceleration measurements of walking [34]	· Can be used for fine-grained HAR · Privacy protecting · Low cost	· Lack of appearance information · Capturing device needs to be carried by subject
	Radar	Spectrogram of falling [35]	· Can be used for through-wall HAR · Insensitive to illumination · Insensitive to weather · Privacy protecting	· Lack of appearance information · Capturing device is relatively expensive
	WiFi	CSI waveform of falling [35]	· Simple and convenient · Privacy protecting · Low cost	· Lack of appearance information · Sensitive to environments · Noisy

for HAR, most of the existing works focused on using videos to handle HAR [37], and only a few approaches utilized static images [38], [39], [40]. Therefore, we focus, in the following, on reviewing RGB video-based HAR methods.

In the pre-deep learning era, many hand-crafted feature-based approaches, including the space-time volume-based methods [41], space-time interest point (STIP)-based methods [42], and trajectory-based methods [43], were well-designed for RGB video-based HAR. Recently, with the great progress of deep learning techniques, various deep learning architectures have also been proposed. Due to the strong representation capability and superior performance of deep learning-based methods, current mainstream research in this field focuses on designing different types of deep learning frameworks. Consequently, we review, in the following, the advanced deep learning works for RGB-based HAR, which can be mainly divided into three categories, namely, two-stream 2D Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and 3D CNN-based methods, as shown in Table 2. In particular, Section 2.1.1 mainly reviews the two-stream methods and their extensions (e.g., multi-stream architectures), which use 2D CNNs as their backbone models. Section 2.1.2 mainly

reviews RGB-based HAR methods which mainly employ RNN models together with 2D CNNs as feature extractors. In Section 2.1.3, we review the 3D CNN-based methods and their extensions.

2.1.1   Two-Stream 2D CNN-Based Methods

As the name suggests, the two-stream 2D CNN framework generally contains two 2D CNN branches taking different input features extracted from the RGB videos for HAR, and the final result is usually obtained through fusion strategies, as shown in Figure 1(a). In this section, we review the classic two-stream methods [44], [45] and also their extensions [46].

As a classic two-stream framework, Simonyan and Zisserman [44] proposed a two-stream CNN model consisting of a spatial network and a temporal network. More specifically, given a video, each individual RGB frame and multi-frame-based optical flows were fed to the spatial stream and temporal stream, respectively. Hence, appearance features and motion features were learned by these two streams for HAR. Finally, the classification scores of these two streams were fused to generate the final classification result. In another classic work, Karpathy et al. [45] fed low-resolution RGB frames and high-resolution center crops to two separate streams to speed up the computation. Different fusion strategies were investigated to model the temporal dynamics in videos. Several studies endeavored to extend and improve over these classic two-stream CNNs, and they are reviewed below.

To produce better video representations for HAR, Wang et al. [47] fed multi-scale video frames and optical flows to a two-stream CNN to extract convolutional feature maps, which were then sampled over the spatio-temporal tubes centered at the extracted trajectories. Finally, the resulting features were aggregated using Fisher Vector representation [48] followed by a SVM for HAR. Chéron et al. [49] used the positions of human body joints to crop multiple human body parts from the RGB and optical flow images, which were passed through a two-stream network for feature extraction followed by a SVM classifier for HAR.

There are also several other works [50], [51], [52], [53], which extended the two-stream framework to extract long-term video-level information for HAR. Wang et al. [50] divided each video into three segments and processed each segment with a two-stream network. The classification scores of the three segments were then fused by an average pooling method to produce the video-level prediction. Instead of fusing the scores of the segments as in [50], Diba et al. [52] aggregated the features of the segments by element-wise multiplication. Girdhar et al. [51] extracted features from sampled appearance and motion frames based on a two-stream framework, and used the vocabulary of "action words" to aggregate the features into a single video-level representation for classification. Feichtenhofer et al. [53] multiplied the appearance residual features with the motion information at the feature level for gated modulation. Zong et al. [54] extended the two-stream CNN in [53] to a three-stream CNN by adding the motion saliency stream to better capture the salient motion information. Bilen et al. [46] constructed dynamic images from RGB sequences and optical flow sequences to summarize the long-term global appearance and motion via rank pooling [55]. The dynamic

images together with the original RGB images and optical flow information were then fed to a multi-stream framework for HAR. Dynamic images have also been adopted by some other works as representations of videos for HAR [56], [57], [58]. Besides, the two-stream network has also been extended to a two-stream Siamese network to extract features from the frames before an action happens (precondition) and the frames after the action (effect), and the action was then represented as a transformation between the two sets of features [59].

To tackle the high computational cost of computing accurate optical flow, some works [60] aimed to mimic the knowledge of the flow stream during training, in order to avoid the usage of optical flow during testing. Zhang et al. [60] proposed a teacher-student framework, which transfers the knowledge from the teacher network trained on optical flow data to the student network trained on motion vectors which can be obtained from compressed videos without extra calculation. Specifically, the knowledge was transferred by using the soft labels produced by the teacher model as an additional supervision to train the student network. Different from [60], Piergiovanni and Ryoo [61] proposed a trainable flow layer which captures the motion information without the need of computing optical flows.

There have also been some works extending the two-stream CNN architectures in other aspects. Wang et al. [62] found that many two-stream CNNs were relatively shallow, and thus, they designed very deep two-stream CNNs for achieving better recognition results. Considering many frames of a video sequence could be irrelevant or useless for HAR, Kar et al. [63] recursively predicted the discriminative importance of each frame for feature pooling. To perform HAR on low spatial resolution videos, Zhang et al. [64] proposed two video super-resolution methods producing high resolution videos, which were fed to the spatial and temporal streams to predict the action class. In the work of [65], several fusion strategies were studied, showing that it is effective to fuse the spatial and temporal networks at the last convolution layer hence reducing the number of parameters, yet keeping the accuracy.

The two-stream 2D CNN architectures, that learn different types of information (e.g., spatial and temporal) from the input videos through separate networks and then perform fusion to get the final result, enable the traditional 2D CNNs to effectively handle the video data and achieve high HAR accuracy. However, this type of architectures is still not powerful enough for long-term dependency modeling, i.e., it has limitations in effectively modeling the video-level temporal information, while temporal sequence modeling networks, such as LSTM, can make up for it.

2.1.2   RNN-Based Methods

RNNs can be used to analyze temporal data due to the recurrent connections in their hidden layers. However, the traditional vanilla RNN suffers from the vanishing gradient problem, making it incapable of effectively modeling the long-term temporal dependency. Thus, most of the existing methods have adopted gated RNN architectures, such as Long-Short Term Memory (LSTM) [67], [68], [69], [70], to model the long-term temporal dynamics in video sequences.

(a) Schema of two-stream 2D CNN-based methods.

(b) Schema of LSTM-based methods.

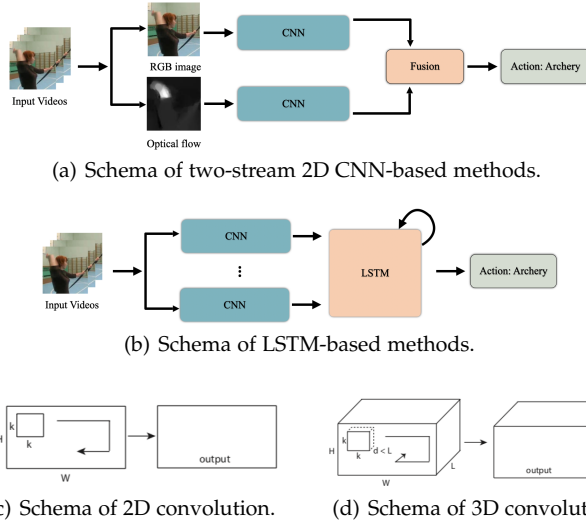(c) Schema of 2D convolution. (d) Schema of 3D convolution.

Fig. 1. Illustration of RGB-based deep learning methods for HAR. (c) and (d) are originally shown in [66].

As shown in Figure 1(b), RNN-based methods usually employ 2D CNNs, which serve as feature extractors, followed by an LSTM model for HAR. Donahue et al. [7] introduced the Long-term Recurrent Convolutional Network (LRCN), which consists of a 2D CNN to extract frame-level RGB features followed by LSTMs to generate a single action label. Ng et al. [71] extracted frame-level RGB and "flow image" features from pre-trained 2D CNNs, and then fed these features to a stacked LSTM framework for HAR. In the work of [72], an encoder LSTM was used to map an input video into a fixed-length representation, which was then decoded through a decoder LSTM to perform the tasks of video reconstruction and prediction in an unsupervised manner. Wu et al. [73] leveraged two LSTMs, which operate on coarse-scale and fine-scale CNN features cooperatively for efficient HAR. Majd and Safabakhsh [74] proposed a C^2LSTM which incorporates convolution and cross-correlation operators to learn motion and spatial features while modeling temporal dependencies. Some other works [75], [76], [77] adopted the Bi-directional LSTM, which consists of two independent LSTMs to learn both the forward and backward temporal information, for HAR.

The introduction of attention mechanisms, in terms of spatial attention [78], [79], [80], [81], temporal attention [82], [83], and both spatial and temporal attention [84], [85], have also benefited the LSTM-based frameworks to achieve better HAR performance. Sharma et al. [78] designed a multi-layer LSTM model, which recursively outputs attention maps weighting the input features of the next frame to focus on the important spatial features, leading to better recognition performance. Sudhakaran et al. [80] introduced a recurrent unit with built-in spatial attention to spatially localize the discriminative information across a video sequence. LSTM has also been used to learn temporal attention to weight features of all frames [82]. Li et al. [85] proposed a Video-LSTM, which incorporates convolutions and motion-based attention into the soft-attention LSTM [86], to better capture both spatial and motion information.

Besides using LSTM, some works [87], [88], [89], [90] have used Gated Recurrent Units (GRUs) [91] for HAR,

which can also alleviate the vanishing gradient problem of the vanilla RNN. Compared to LSTM, GRU has fewer gates, leading to fewer model parameters, yet it can usually provide a similar performance of LSTM for HAR [92].

Several other works [76], [93], [94] focused on the hybrid architectures combining two-stream 2D CNN and RNN models for HAR. For example, Wu et al. [93] utilized two-stream 2D CNN models to extract spatial and short-term motion features, which were fed to two LSTMs respectively to model longer-term temporal information for HAR.

Generally, most of the RNN-based methods operate on the CNN features of frames instead of the high-dimensional frame images for HAR. Instead of using 2D CNN, some other works [95], [96], [97], [98] used 3D CNNs as their feature extractors. In the following section, we review the 3D CNN-based methods.

2.1.3 3D CNN-Based Methods

Plenty of researches [66], [99], [100], [101] have extended 2D CNNs (Figure 1(c)) to 3D structures (Figure 1(d)), to simultaneously model the spatial and temporal context information in videos that is crucial for HAR. As one of the earliest works, Ji et al. [99] segmented human subjects in videos by utilizing a human detector method, and then fed the segmented videos to a novel 3D CNN model to extract spatio-temporal features from videos. Unlike [99], Tran et al. [66] introduced a 3D CNN model, dubbed C3D, to learn the spatio-temporal features from raw videos in an end-to-end learning framework. However, these networks were mainly used for clip-level learning (e.g., 16 frames in each clip) instead of learning from full videos, which thus ignored the long-range spatio-temporal dependencies in videos. Hence, several approaches focused on modeling the long-range spatio-temporal dependencies in videos. For example, Diba et al. [102] extended DenseNet [103] with 3D filters and pooling kernels, and designed a Temporal 3D CNN (T3D), where the temporal transition layer can model variable temporal convolution kernel depths. T3D can densely and efficiently capture the appearance and temporal information at short, middle, and also long terms. In their subsequent study [104], they introduced a new block embedded in some architectures such as ResNext and ResNet, which can model the inter-channel correlations of a 3D CNN with respect to the temporal and spatial features. Varol et al. [105] proposed a Long-term Temporal Convolution (LTC) framework, which increases the temporal extents of 3D convolutional layers at the cost of reducing the spatial resolution, to model the long-term temporal structure. Hussein et al. [106] proposed multi-scale temporal-only convolutions, dubbed Timeception, to account for large variations and tolerate a variety of temporal extents in complex and long actions. Wang et al. [107] proposed a non-local operation, which models the correlations between any two positions in the feature maps to capture long-range dependencies. Besides, Li et al. [101] proposed a Channel Independent Directional Convolution (CIDC) which can be attached to I3D [108] to better capture the long-term temporal dynamics of the full input video.

To enhance the HAR performance, several other works [108], [109], [110], [111], [112] have investigated 3D CNN

models with two-stream or multi-stream designs. For example, Carreira and Zisserman [108] introduced the two-stream Inflated 3D CNN (I3D) inflating the convolutional and pooling kernels of a 2D CNN with an additional temporal dimension. Wang et al. [98] integrated a two-stream 3D CNN with an LSTM model to capture the long-range temporal dependencies. Feichtenhofer et al. [112] designed a two-stream 3D CNN framework containing a slow pathway and a fast pathway that operate on RGB frames at low and high frame rates to capture semantic and motion, respectively. Inspired by [113], the features of the fast pathway were fused (by summation or concatenation) with the features of the slow pathway at each layer. Li et al. [114] introduced a two-stream spatio-temporal deformable 3D CNN with attention mechanisms to capture the long-range temporal and long-distance spatial dependencies. Besides, deep learning frameworks combining 2D CNNs and 3D CNNs have also been investigated for HAR [115], [116], [117]. For example, the ECO architecture [115] uses 2D CNNs to extract spatial features, which are stacked and then fed to 3D CNNs to model long-term dependencies for HAR.

Some works focused on addressing certain other problems in 3D CNNs. For example, Zhou et al. [118] analyzed the spatio-temporal fusion in 3D CNN from a probabilistic perspective. Yang et al. [119] proposed a generic feature-level Temporal Pyramid Network (TPN) to model speed variations in performing actions. Kim et al. [120] proposed a Random Mean Scaling (RMS) regularization method to address the problem of overfitting. The viewpoint variation problem has also been investigated [121], [122]. Inspired by [123], Varol et al. [122] proposed to train 3D CNN on synthetic action videos to address the problem of variations in viewpoints. Piergiovanni and Ryoo [121] proposed a geometric convolutional layer to learn view-invariant representations of actions by imposing the latent action representations to follow 3D geometric transformations and projections. Knowledge distillation has also been investigated to improve motion representations of 3D CNN frameworks. For example, Stroud et al. [124] introduced a Distilled 3D Network (D3D) consisting of a student network and a teacher network, where the student network was trained on RGB videos and it also distilled knowledge from the teacher network that was trained on optical flow sequences. Crasto et al. [125] transferred the knowledge of a teacher model trained on the optical flows to a 3D CNN student network operating on RGB videos, by minimizing the mean squared error between the feature maps of the two streams. Following the research line of handling the large computational cost of obtaining optical flow, Shou et al. [126] proposed an adversarial framework with a lightweight generator to approximate flow information by refining the noisy and coarse motion vector available in the compressed video. Differently, Wang et al. [127] adopted an efficient learnable correlation operator to better learn motion information from 3D appearance features. Fayyaz et al. [128] addressed the problem of dynamically adapting the temporal feature resolution within the 3D CNNs to reduce their computational cost. A Similarity Guided Sampling (SGS) module was proposed to enable 3D CNNs to dynamically adapt their computational resources by selecting the most informative and distinctive temporal features.

The 3D CNN-based methods are very powerful in modeling discriminative features from both the spatial and temporal dimensions for HAR. However, many 3D CNN-based frameworks contain a large number of parameters, and thus require a large amount of training data. Therefore, some studies [109], [129], [130], [131], [132] aimed to factorize 3D convolutions. Sun et al. [129] proposed a Factorized spatio-temporal CNN (F_{st}CN), factorizing the 3D convolution to 2D spatial convolutional layers followed by 1D temporal convolutional layers. Similarly, Qiu et al. [130] decomposed the 3D convolution into a 2D convolution for the spatial domain, followed by a 1D convolution for the temporal domain, to economically and effectively simulate 3D convolutions. Xie et al. [131] explored the performance of several variants of I3D [108] by utilizing a combination of 3D and 2D convolutional filters in the I3D network, and introduced temporally separable convolutions and spatio-temporal feature gating to enhance HAR. Yang et al. [133] proposed efficient asymmetric one-directional 3D convolutions to approximate the traditional 3D convolution. Besides, some other works [134], [135], [136], [137] also utilized 2D CNNs with powerful temporal modules to decrease the computational cost of 3D CNNs. Lin et al. [134] proposed a Temporal Shift Module (TSM), which shifts a part of the channels along the temporal dimension to perform temporal interaction between the features from adjacent frames. Unlike parameter-free temporal shift operations in [134], Sudhakaran et al. [135] introduced a lightweight Gate-Shift Module (GSM), which uses learnable spatial gating blocks for spatial-temporal decomposition of 3D convolutions. Wang et al. [136] designed a two-level Temporal Difference Module (TDM) to capture both finer local and long-range global motion information. Wang et al. [137] introduced an ACTION module leveraging multipath excitation to respectively model spatial-temporal, channel-wise, and motion patterns.

Apart from the above-mentioned architectures, i.e., two-stream 2D CNN, RNN, and 3D CNN, there are also some other frameworks designed for HAR using RGB videos, such as Convolutional Gated Restricted Boltzmann Machines [138], Graph-based Modeling [139], [140], Transformers [141], and 4D CNNs [142].

In general, RGB data is the most commonly used modality for HAR in real-world application scenarios, since it is easy to collect and contains rich information. Besides, RGB-based HAR methods can leverage large-scale web videos to pre-train models for better recognition performance [143], [144]. However, it often requires complex computations for feature extraction from RGB videos. Moreover, HAR methods based on the RGB modality are often sensitive to viewpoint variations and background clutters, etc. Hence, recently, HAR with other modalities, such as 3D skeleton data, has also received great attention, and is therefore discussed in the subsequent sections.

2.2 SKELETON MODALITY

Skeleton sequences encode the trajectories of human body joints, which characterize informative human motions. Therefore, skeleton data is also a suitable modality for HAR. The skeleton data can be acquired by applying pose

TABLE 2
Performance comparison of RGB video-based deep learning methods for HAR on the UCF101, HMDB51, and Kinectis-400 datasets. Note in RGB video-based HAR methods, some other information, e.g., optical flow or motion vector, may also be used along with the RGB data as input for HAR. For simplicity, '%' after the value is omitted. '-' indicates the result is unavailable. 'Flow' denotes optical flow.

Method	Year	Input	Dataset		
			UCF101	HMDB51	Kinetics-400
Multiresolution CNN [45]	2014	High-,Low-Resolution RGB	65.4	-	-
Two-stream CNN+SVM [44]	2014	RGB,Flow	88.0	59.4	-
P-CNN [49]	2015	RGB,Flow	-	-	-
Very deep two-stream [62]	2015	RGB,Flow	91.4	-	-
TDD+iDT [47]	2015	RGB,Flow	91.5	65.9	-
EMV+RGB-CNN [60]	2016	RGB,Motion Vector	86.4	-	-
Siamese Network [59]	2016	RGB,Flow	92.4	62.0	-
Feichtenhofer et al. [65]	2016	RGB,Flow	93.5	69.2	-
TSN [50]	2016	RGB,Flow	94.2	69.4	-
AdaScan+iDT [63]	2017	RGB,Flow	91.3	61.0	-
ActionVLAD+iDT [51]	2017	RGB,Flow	93.6	69.8	-
Feichtenhofer et al. [53]	2017	RGB,Flow	94.9	72.2	-
TLE: Bilinear [52]	2017	RGB,Flow	95.6	71.1	-
Bilen et al. [46]	2017	RGB,Flow,Dynamic Images	96.0	74.9	-
TSN+SoSR+ToSR [64]	2019	RGB,Flow	92.1	68.3	-
MSM-ResNets [54]	2021	RGB,Flow,Saliency Map	93.5	66.7	-
Soft Attention Model [78]	2015	RGB	-	41.3	-
LRCNs [7]	2015	RGB,Flow	82.7	-	-
Composite LSTM Model [72]	2015	RGB,Flow	84.3	-	-
Ng et al. [71]	2015	RGB,Flow	88.6	-	-
Wu et al. [93]	2015	RGB,Flow	91.3	-	-
RNN-FV+iDT [145]	2016	RGB	94.1	67.7	-
L^2STM [68]	2017	RGB,Flow	93.6	66.2	-
shuttleNet [88]	2017	RGB,Flow	95.4	71.7	-
iDT(FV)+Objects+VideoLSTM [85]	2018	RGB,Flow	92.2	64.9	-
CNN+TR-LSTM [146]	2019	RGB	-	63.8	-
Ge et al. [79]	2019	RGB,Flow	92.8	67.1	-
AR-Net [70]	2020	RGB	-	-	-
STS-ALSTM [84]	2020	RGB,Flow,Saliency Map	92.7	64.4	-
C^2LSTM [74]	2020	RGB	92.8	61.3	-
DB-LSTM [76]	2021	RGB,Flow	97.3	81.2	-
Ji et al. [99]	2012	RGB,Gradient,Flow	-	-	-
F_{st}CN [129]	2015	RGB	88.1	59.1	-
C3D+iDT+SVM [66]	2015	RGB	90.4	-	-
scLSTM [96]	2016	RGB	84.0	55.1	-
ST-ResNet*+iDT [113]	2016	RGB,Flow	94.6	70.3	-
STPP [98]	2017	RGB,Flow	92.6	70.5	-
$LTC_{Flow+RGB}$+iDT [105]	2017	RGB,Flow	92.7	67.2	-
T3D+TSN [102]	2017	RGB	93.2	63.5	62.2 (T3D)
P3D ResNet+iDT [130]	2017	RGB	93.7	-	-
Two-Stream I3D [108]	2017	RGB,Flow	97.9	80.2	-
NL I3D [107]	2018	RGB	-	-	77.7
ARTNet with TSN [116]	2018	RGB	94.3	70.9	80.0
Two-stream MiCT-Net [110]	2018	RGB,Flow	94.7	70.5	-
ECO_En [115]	2018	RGB	94.8	72.4	70.0
STC-ResNext 101 (64 frames) [104]	2018	RGB	96.5	74.9	68.7 (32 frames)
S3D-G [131]	2018	RGB	96.8	75.9	74.7
Hidden Two-stream CNN+I3D [111]	2018	RGB	97.1	78.7	-
R(2+1)D-TwoStream [109]	2018	RGB,Flow	97.3	78.7	75.4
Martinez et al. [117]	2019	RGB	-	-	78.8
SlowFast 16×8, R101+NL [112]	2019	RGB	-	-	79.8
irCSN-152 [144]	2019	RGB	-	-	82.8
Asymmetric 3D-CNN+iDT [133]	2019	RGB,RGBF	92.6	65.4	-
I3D RGB+DMC-Net (I3D) [126]	2019	RGB,Residual,Motion Vector	96.5	77.8	-
MARS [125]	2019	RGB,Flow	98.1	80.9	74.9
SlowFast+RMS [120]	2020	RGB	-	-	76.3
TPN-R101 [119]	2020	RGB	-	-	78.9
X3D-XL [147]	2020	RGB	-	-	79.1
CorrNet-101 [127]	2020	RGB	-	-	81.0
STDA-ResNeXt-101 [114]	2020	RGB	95.5	72.7	-
Zhou et al. [118]	2020	RGB	96.5	-	72.5
D3D [124]	2020	RGB	97.0	78.7	75.9
CIDC (I3D-NL) [101]	2020	RGB	97.9	75.2	75.6
OmniSource (RGB & Flow) [143]	2020	RGB,Flow	98.6	83.8	-
OmniSource (RGB) [143]	2020	RGB	-	-	83.6
R(2+1)D+BERT (64f) [148]	2020	RGB	98.7	85.1	-
X3D XL+AFTR$^\beta$ [128]	2021	RGB	-	-	79.3
CT-Net_EN [132]	2021	RGB	-	-	79.8
Yang et al. [149]	2021	RGB	-	-	81.4
3DResNet+ATFR [128]	2021	RGB	97.9	76.7	-

estimation algorithms on RGB videos [150] or depth maps [5]. It can also be collected with motion capture systems. Generally, human pose estimation is sensitive to viewpoint variations. Meanwhile, motion capture systems that are insensitive to view and lighting can provide reliable skeleton data. However, in many application scenarios, it is not convenient to deploy motion capture systems. Thus many recent works on skeleton-based HAR used skeleton data obtained from depth maps [151] or RGB videos [152].

There are many advantages of using skeleton data for HAR, due to its provided body structure and pose information, its essentially simple and informative representation, its scale invariance, and its robustness against variations of clothing textures and backgrounds. Due to these advantages and also the availability of accurate and low-cost depth sensors, skeleton-based HAR has attracted much attention in the research community recently.

Early works focused on extracting hand-crafted spatial and temporal features from skeleton sequences for HAR. The hand-crafted feature-based methods can be roughly divided into joint-based [153] and body part-based [154] methods, depending on the used feature extraction techniques. Due to the strong feature learning capability, deep learning has been widely used for skeleton-based HAR, and has become the mainstream research in this field. Consequently, we review, in the following, the deep learning methods, which can be mainly divided into three categories: RNN, CNN, and Graph Neural Network (GNN) or Graph Convolutional Network (GCN)-based methods, as shown in Figure 2. Table 3 shows the results achieved by different skeleton-based HAR methods on two benchmark datasets.

2.2.1 RNN-Based Methods

As mentioned in Section 2.1.2, RNNs and their gated variants (e.g., LSTMs) are capable of learning the dynamic dependencies in sequential data. Hence, various methods [155], [156], [157], [158] have applied and adapted RNNs and LSTMs to effectively model the temporal context information within the skeleton sequences for HAR.

In one of the classical methods, Du et al. [155] proposed an end-to-end hierarchical RNN, dividing the human skeleton into five body parts instead of inputting the skeleton in each frame as a whole. These five body parts were then separately fed to multiple bidirectional RNNs, whose output representations were hierarchically fused to generate high-level representations of the action. Differential RNN (dRNN) [159] learns salient spatio-temporal information by quantifying the change of the information gain caused by the salient motions between frames. The Derivative of States (DoS) was proposed inside the LSTM unit to act as a signal controlling the information flow into and out of the internal state over time. Zhu et al. [160] introduced a novel mechanism for the LSTM network to achieve automatic co-occurrence mining, since co-occurrence intrinsically characterizes the actions. Sharoudy et al. [161] proposed a Part-aware LSTM (P-LSTM) by introducing a mechanism to simulate the relations among different body parts inside the LSTM unit. Liu et al. [8], [162] extended the RNN design to both the temporal and spatial domains. Specifically, they utilized the tree structure-based skeleton traversal method to further exploit the spatial information, and trust gates to deal with noise and occlusions. In their subsequent study [163], an attention-based LSTM network named Global Context-Aware Attention LSTM (GCA-LSTM) was proposed to selectively focus on the informative joints using the global context information. The GCA-LSTM network contains two LSTM layers, where the first layer encodes the skeleton sequence and outputs a global context memory, while the second layer outputs attention representations to refine the global context. Finally, a softmax classifier outputs the HAR result. In the work of [164], a two-stream RNN structure was proposed to model both the temporal dynamics and spatial configurations. Deep LSTM with spatio-temporal attention was proposed in [165], where a spatial attention sub-network and a temporal attention sub-network work jointly under the main LSTM network. To model variable temporal dynamics of skeleton sequences, Lee et al. [166] proposed an ensemble Temporal Sliding LSTM (TS-LSTM) framework composed of multiple parts, containing short-term, medium-term, and long-term TS-LSTM networks. IndRNN [167] not only addresses the gra-

dient vanishing and exploding issues, but it is also faster than the original LSTM.

2.2.2 CNN-Based Methods

CNNs have achieved great success in 2D image analysis due to their superior capability in learning features in the spatial domain. However, when facing skeleton-based HAR, the modeling of spatio-temporal information becomes a challenge. Nevertheless, plenty of advanced approaches have been proposed, which apply temporal convolution on skeleton data [168], or represent skeleton sequences as pseudo-images that are then fed to standard CNNs for HAR [169], [170]. These pseudo-images were designed to simultaneously encode spatial structure information in each frame and temporal dynamic information between frames.

Hou et al. [169] and Wang et al. [170] respectively proposed the skeleton optical spectra and the joint trajectory maps. They both encoded the spatio-temporal information of skeleton sequences into color-texture images, and then adopted CNNs for HAR. As an extension of these works, the Joint Distance Map (JDM) [171] produces view-invariant color-texture images encoding pair-wise distances of skeleton joints. Ke et al. [172] transformed each skeleton sequence into three "video clips", which were then fed to a pre-trained CNN to produce a compact representation, followed by a multi-task learning network for HAR. Kim and Reiter [168] utilized the Temporal CNN (TCN) [173] explicitly providing a type of interpretable spatio-temporal representations. An end-to-end framework to learn co-occurrence features with a hierarchical methodology was proposed in [174]. Specifically, they learned the point-level features of each joint independently, and then utilized these features as a channel of the convolutional layer to learn hierarchical co-occurrence features, and a two-stream framework was adopted to fuse the motion features. Caetano et al. introduced SkeleMotion [175] and the Tree Structure Reference Joints Image (TSRJI) [176] as representations of skeleton sequences for HAR. In [177], Skepxels were utilized as basic building blocks to construct skeletal images mainly encoding the spatio-temporal information of human joint locations and velocities. Besides, Skepxels were used to hierarchically capture the micro-temporal relations between the joints in the frames, and Fourier Temporal Pyramids [9] were utilized to exploit the macro-temporal relations.

Plenty of researches focused on addressing certain specific problems. For example, the works of [171], [178], [179] focused on handling the viewpoint variation issue. Since features learned from skeleton data are not always translation, scale, and rotation invariant, several methods [180], [181] have also been designed to handle these issues. CNN architectures are often complex, resulting in high computational costs. Thus Yang et al. [182] proposed a Double-feature Double-motion Network (DD-Net) to make the CNN-based recognition models run faster.

2.2.3 GNN or GCN-Based Methods

Due to the expressive power of graph structures, analyzing graphs with learning models have received great attention recently [183], [184]. As shown in Figure 2(c), skeleton data is naturally in the form of graphs. Hence, simply representing skeleton data as a vector sequence processed by RNNs,

(a) RNN-based method

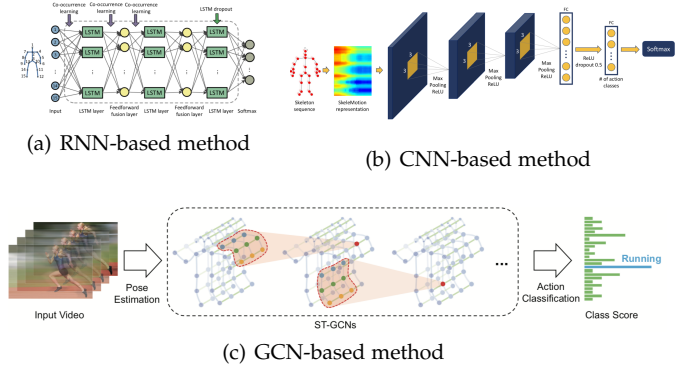(b) CNN-based method

(c) GCN-based method

Fig. 2. Illustration of deep learning frameworks for skeleton-based HAR. (a) Skeleton sequences can be processed by RNN and LSTM as time series. (b) Skeleton sequences can be converted to 2D pseudo-images and then be fed to 2D CNNs for feature learning. (c) The joint dependency structure can naturally be represented via a graph structure, and thus GCN models are also suitable for this task. (a)-(c) are originally shown in [152], [160], [175].

or 2D/3D maps processed by CNNs, cannot fully model the complex spatio-temporal configurations and correlations of the body joints. This indicates that topological graph representations can be more suitable for representing the skeleton data. As a result, many GNN and GCN-based HAR methods [152], [185] have been proposed to treat the skeleton data as graph structures of edges and nodes.

GNN is a connection model capturing the dependence within a graph through message passing between nodes. Si et al. [185] proposed a spatial reasoning network, followed by RNNs, to capture the high-level spatial structural and temporal dynamics of skeleton data. Shi et al. [186] represented the skeleton as a directed acyclic graph to effectively incorporate both the joint and bone information, and utilized a GNN to perform HAR.

More recently, GCN-based HAR has become a hot research direction [187], [188], [189], [190], [191]. Yan et al. [152] exploited GCNs for skeleton-based HAR by introducing Spatial-Temporal GCNs (ST-GCNs) that can automatically learn both the spatial and temporal patterns from skeleton data, as shown in Figure 2(c). More specifically, the pose information was estimated from the input videos and then passed through the spatio-temporal graphs to achieve action representations with strong generalization capabilities for HAR. Since implicit joint correlations have been ignored by previous works [152], Li et al. [192] further proposed an Actional-Structural GCN (AS-GCN), combining actional links and structural links into a generalized skeleton graph. The actional links were used to capture action-specific latent dependencies and the structural links were used to represent higher-order dependencies. To better explore the implicit joint correlations, Peng et al. [193] determined their GCN architecture via a neural architecture search scheme. Specifically, they enriched the search space to implicitly capture the joint correlations based on multiple dynamic graph sub-structures and higher-order connections with a Chebyshev polynomial approximation. Besides, context information integration was used to effectively model long-range dependencies in the work of [194].

Shi et al. [195] proposed a two-stream Adaptive GCN (2s-AGCN), where the topology of graphs can be either

uniformly or individually learned by the back-propagation algorithm, instead of setting it manually. The 2s-AGCN explicitly combined the second-order information (lengths and directions of human bones) of the skeleton with the first-order information (coordinates of joints). Wu et al. [196] introduced a cross-domain spatial residual layer to capture the spatio-temporal information, and a dense connection block to learn global information based on ST-GCN. Li et al. [197] fed frame-wise skeleton and node trajectories of a skeleton sequence to a spatial graph router and a temporal graph router to generate new skeleton-joint-connectivity graphs, followed by a ST-GCN for classification. Liu et al. [198] integrated a disentangled multi-scale aggregation scheme and a spatio-temporal graph convolutional operator named G3D to achieve a powerful feature extractor. High-level semantics of joints were introduced for HAR in [199]. Attention mechanisms were applied for extracting discriminative information and global dependencies in [200], [201]. Besides, to reduce computational costs of GCNs, a Shift-GCN was designed by Cheng et al. [202], which adopts shift graph operations and lightweight point-wise convolutions, instead of using heavy regular graph convolutions. Following this research line, Song et al. [203] proposed a multi-stream GCN model, which fuses the input branches including joint positions, motion velocities, and bone features at early stage, and utilized separable convolutional layers and a compound scaling strategy to extremely reduce the redundant trainable parameters while increasing the capacity of model. Different from the above mentioned methods, Li et al. [204] proposed symbiotic GCNs to handle both action recognition and motion prediction tasks simultaneously. The proposed Sym-GNN consists of a multi-branch multi-scale GCN followed by an action recognition and a motion prediction heads, to jointly conduct action recognition and motion prediction tasks. This allows the two tasks to enhance each other.

In summary, the skeleton modality provides the body structure information, which is simple, efficient, and informative for representing human behaviors. Nevertheless, HAR using skeleton data still faces challenges, due to its very sparse representation, the noisy skeleton information, and the lack of shape information that can be important when handling human-object interactions. Hence, some of the existing works on HAR also focused on using depth maps, as discussed in the following section, since depth maps not only provide the 3D geometric structural information but also conserve the shape information.

2.3 DEPTH MODALITY

Depth maps refer to images where the pixel values represent the distance information from a given viewpoint to the points in the scene. The depth modality, which is often robust to variations of color and texture, provides reliable 3D structural and geometric shape information of human subjects, and thus can be used for HAR. The essence of constructing a depth map is to convert the 3D data into a 2D image, and different types of devices have been developed to obtain depth images, which include active sensors (e.g., Time-of-Flight and structured-light-based cameras) and passive sensors (e.g., stereo cameras) [19]. Active sensors emit

radiation in the direction towards the objects in the scene, and then measure the reflected energy from the objects to acquire the depth information. In contrast, passive sensors measure the natural energy that is emitted or reflected by the objects in the scene. For example, as a type of passive sensor, stereo cameras usually have two or more lenses to simulate the binocular vision of human eyes to acquire depth information, and depth maps are recovered by seeking image point correspondences between stereo pairs [213]. Compared to active sensors (e.g., Kinect and RealSense3D), passive depth map generation is usually computationally expensive, and moreover, depth maps obtained with passive sensors can be ineffective in texture-less regions or highly textured regions with repetitive patterns.

In this section, we review methods that used depth maps for HAR. Note that only a few works [214], [215] used depth maps captured by stereo cameras for HAR, while most of the other methods [9], [56], [216] focused on using depth videos captured by active sensors to recognize actions, due to the availability of low-cost and reliable active sensors like Kinect. Consequently, here we focus on reviewing the methods using depth videos captured with active sensors for HAR. Table 4 shows the results of some depth-based HAR methods on several benchmark datasets.

TABLE 3
Performance of skeleton-based deep learning HAR methods on NTU RGB+D and NTU RGB+D 120 datasets. 'CS', 'CV', and 'CP' denote Cross-Subject, Cross-View, and Cross-Setup evaluation criteria.

	Methods	Year	NTU RGB+D		NTU RGB+D 120	
			CS	CV	CS	CP
RNN-Based	dRNN [159]	2015	-	-	-	-
	HBRNN-L [155]	2015	59.1	64.0	-	-
	Co-occurrence LSTM [160]	2016	-	-	-	-
	2 Layer P-LSTM [161]	2016	62.9	70.3	25.5	26.3
	Trust Gate ST-LSTM [8]	2016	69.2	77.7	58.2	60.9
	Zhang et al. [156]	2017	70.3	82.4	-	-
	Two-stream RNN [164]	2017	71.3	79.5	-	-
	STA-LSTM [165]	2017	73.4	81.2	-	-
	GCA-LSTM [163]	2017	74.4	82.8	58.3	59.2
	Ensemble TS-LSTM [166]	2017	74.6	81.3	-	-
	VA-LSTM [205]	2017	79.4	87.6	-	-
	Zhang et al. [157]	2018	76.4	87.7	-	-
	MANs [206]	2018	82.7	93.2	-	-
	VA-RNN (aug.) [179]	2019	79.8	88.9	-	-
	dense-IndRNN-aug [167]	2019	86.7	93.7	-	-
CNN-Based	Du et al. [207]	2015	-	-	-	-
	Hou et al. [169]	2016	-	-	-	-
	JTM [170]	2016	73.4	75.2	-	-
	Res-TCN [168]	2017	74.3	83.1	-	-
	SkeletonNet [180]	2017	75.9	81.2	-	-
	JDM [171]	2017	76.2	82.3	-	-
	Clips+CNN+MTLN [172]	2017	79.6	84.8	58.4	57.9
	Liu et al. [178]	2017	80.0	87.2	60.3	63.2
	Li et al. [181]	2017	85.0	92.3	-	-
	RotClips+MTCNN [208]	2018	81.1	87.4	62.2	61.8
	Xu et al. [209]	2018	84.8	91.2	-	-
	HCN [174]	2018	86.5	91.1	-	-
	DD-Net [182]	2019	-	-	-	-
	TSRJI [176]	2019	73.3	80.3	67.9	62.8
	SkeleMotion [175]	2019	76.5	84.7	67.7	66.9
	Skepxel [177]	2019	81.3	89.2	-	-
	Li et al. [210]	2019	82.9	90.0	-	-
	VA-CNN (aug.) [179]	2019	88.7	94.3	-	-
	Banerjee et al. [211]	2020	84.2	89.7	74.8	76.9
GNN or GCN-Based	ST-GCN [152]	2018	81.5	88.3	-	-
	DPRL [187]	2018	83.5	89.8	-	-
	SR-TSL [185]	2018	84.8	92.4	-	-
	motif-GCNs [201]	2019	84.2	90.2	-	-
	BPLHM [188]	2019	85.4	91.1	-	-
	AS-GCN [192]	2019	86.8	94.2	-	-
	STGR-GCN [197]	2019	86.9	92.3	-	-
	2s-AGCN [195]	2019	88.5	95.1	-	-
	AGC-LSTM [200]	2019	89.2	95.0	-	-
	2s-SDGCN [196]	2019	89.6	95.7	-	-
	DGNN [186]	2019	89.9	96.1	-	-
	Advanced CA-GCN [194]	2020	83.5	91.4	-	-
	DC-GCN+ADG [190]	2020	88.2	95.2	90.8	96.6
	SGN [199]	2020	89.0	94.5	79.2	81.5
	GCN-NAS [193]	2020	89.4	95.7	-	-
	4s Shift-GCN [202]	2020	90.7	96.5	85.9	87.6
	DDGCN [189]	2020	91.1	97.1	-	-
	MS-G3D Net [198]	2020	91.5	96.2	86.9	88.4
	3s-CrosSCLR (FT) [212]	2021	86.2	92.5	80.5	80.4
	Sym-GNN [204]	2021	90.1	96.4	-	-
	EfficientGCN-B4 [203]	2021	91.7	95.7	88.3	89.1

Compared to hand-crafted feature-based methods [217], [218], deep learning models [9], [56] have shown to be more powerful and achieved better performance for HAR from depth maps. Inspired by the hand-crafted Depth Motion Map (DMM) features [219], a deep learning framework utilizing weighted hierarchical DMMs was proposed in [220]. As DMMs are not able to capture detailed temporal information, Wang et al. [56] proposed to represent depth sequences with three pairs of structured dynamic images [55] at the body, body part, and joint levels, which were then fed to CNNs followed by a score fusion module for fine-grained HAR.

The performance of this method was further improved in [216] by introducing three representations of depth maps, including dynamic depth images, dynamic depth normal images, and dynamic depth motion normal images. Rahmani et al. [9] transferred the human data obtained from different views to a view-invariant high-level space to address the view-invariant HAR problem. Specifically, they utilized a CNN model to learn the view-invariant human pose model and Fourier Temporal Pyramids to model the temporal action variations. To obtain more multi-view training data, synthetic data was generated by fitting synthetic 3D human models to real motion capture data, and rendering the human data from various viewpoints. In [57], multi-view dynamic images were extracted through multi-view projections from depth videos for action recognition. In order to effectively capture spatio-temporal information in depth videos, Sanchez et al. [221] proposed a 3D fully CNN architecture for HAR. In their subsequent work [222], a variant of LSTM unit was introduced to address the problem of memory limitation during video processing, which can be used to perform HAR from long and complex videos.

Note that besides using depth maps obtained with active sensors or stereo cameras for HAR, there has been another approach designed for depth-based HAR from RGB videos. Specifically, Zhu and Newsam [223] estimated the depth maps from RGB videos using existing depth estimation techniques [224], [225], which were then passed through a deep learning architecture for action classification.

In general, the depth modality provides geometric shape information that is useful for HAR. However, the depth data is often not used alone, due to the lack of appearance information that can also be helpful for HAR in some scenarios. Thus, many works focused on fusing depth information with other data modalities for enhanced HAR, and more details can be found in Section 3.

2.4 INFRARED MODALITY

Generally, infrared sensors do not need to rely on external ambient light, and thus are particularly suitable for HAR at night. Infrared sensing technologies can be divided into active and passive ones. Some infrared sensors, such as Kinect, rely on active infrared technology, which emit infrared rays and utilize target reflection rays to perceive objects in the scene. In contrast, thermal sensors relying on passive infrared technology do not emit infrared rays. Instead, they work by detecting rays (i.e., heat energy) emitted from targets.

Some deep learning methods [30], [226] have been proposed for HAR from infrared data. In [227], the extremely low-resolution thermal images were first cropped by the gravity center of human regions. The cropped sequences and the frame differences were then passed through a CNN followed by an LSTM layer to model spatio-temporal information for HAR. To simultaneously learn both the spatial and temporal features from thermal videos, Shah et al. [228] utilized a 3D CNN and achieved real-time HAR. Instead of using raw thermal images, Meglouli et al. [229] passed the optical flow information computed from thermal sequences into a 3D CNN for HAR.

Inspired by the two-stream CNN model [44], several multi-stream architectures [10], [30], [230] have also been proposed. Gao et al. [30] passed optical flow and optical flow motion history images (OF-MHIs) [231] through a two-stream CNN for HAR. In [10], both infrared data and optical flow volumes were fed to a two-stream 3D CNN framework to effectively capture the complementary information on appearance from still thermal frames and motion between frames. The two-stream model was trained using a combination of cross-entropy and label consistency regularization loss in [232]. To better represent the global temporal information, Optical flow motion history images (OF-MHIs) [231], optical flow and a stacked difference of optical flow images, were fed to a three-stream CNN for HAR [230]. Imran and Raman [233] proposed a four-stream architecture, where each stream consists of a CNN followed by an LSTM. The local/global stacked dense flow difference images [234] and local/global Stacked saliency difference images were fed to these four streams to capture both the local and global spatio-temporal information in videos. Mehta et al. [235] presented an adversarial framework consisting of a two-stream 3D convolutional auto-encoder as the generator and two 3D CNNs as the joint discriminator. The generator network took thermal data and optical flow as inputs, and the joint discriminator tried to discriminate the real thermal data and optical flow from the reconstructed ones.

In all, infrared data has been used for HAR, especially for night HAR. However, infrared images may suffer from a relatively low contrast and a low signal-to-noise ratio, making it challenging for robust HAR in some scenarios.

2.5 POINT CLOUD MODALITY

Point cloud data is composed of a numerous collection of points that represent the spatial distribution and surface characteristics of the target under a spatial reference system. There are two main ways to obtain 3D point cloud data, namely, (1) using 3D sensors, such as LiDAR and Kinect, or (2) using image-based 3D reconstruction. As a 3D data modality, point cloud has great power to represent the spatial silhouettes and 3D geometric shapes of the subjects, and hence can be used for HAR.

Early methods [238], [251] focused on extracting hand-crafted spatio-temporal descriptors from the point cloud sequences for HAR, while the current mainstream research focuses on deep learning architectures [245], [246] that generally achieve better performance. Wang et al. [11] transformed the raw point cloud sequence into regular voxel sets. Then temporal rank pooling [55] was applied on all the voxel sets to encode 3D action information into one single voxel set. Finally, the voxel representation was abstracted

TABLE 4
Performance of depth, infrared, point cloud, and event stream-based deep learning HAR methods. The datasets are MSRDailyActivity3D (M) [236], Northwestern-UCLA (N-UCLA) [237], UWA3D Multiview II (UWA3D II) [238], NTU RGB+D (N) [161], InfAR [30], MSR-Action3D [239], DvsGesture [240], and DHP19 [32].

| | Method | Year | Dataset | | | N | |
			M	N–UCLA	UWA3D II	CS	CV
Depth	Wang et al. [220]	2015	85.0	-	-	-	-
	Rahmani et al. [9]	2016	80.0	92.0	76.9	-	-
	S²DDI [56]	2017	100	-	-	-	-
	Wang et al. [216]	2018	-	-	-	87.1	84.2
	MVDI [57]	2019	-	84.2	68.1	84.6	87.3
	3DFCNN [221]	2020	-	83.6	66.6	78.1	80.4
	Stateful ConvLSTM [222]	2020	-	-	-	80.4	79.9
			InfAR				
Infrared	Gao et al. [30]	2016	76.7				
	Jiang et al. [10]	2017	77.5				
	Kawashima et al. [227]	2017	-				
	Shah et al. [228]	2018	-				
	TSTDDs [230]	2018	79.3				
	Akula et al. [226]	2018	-				
	Imran et al. [233]	2019	83.5				
	Meglouli et al. [229]	2019	88.2				
	Mehta et al. [235]	2020	-				
			MSR-Action3D			N CS	CV
Point Cloud	MeteorNet [241]	2019	88.5			-	-
	PointLSTM [242]	2020	-			-	-
	3DV-PointNet++ [11]	2020	-			88.8	96.3
	MeteorNet (24 frames) [243]	2020	93.0			-	-
	4D MinkNet [244]	2021	86.3			-	-
	P4Transformer [245]	2021	90.9			90.2	96.4
	PSTNet [246]	2021	91.2			90.5	96.5
			DvsGesture			**DHP19**	
Event Stream	Amir et al. [240]	2017	96.5			-	
	Ghosh et al. [12]	2019	94.9			-	
	Wang et al. [247]	2019	97.1			-	
	Huang et al. [248]	2020	97.4			-	
	Chen et al. [249]	2020	98.6			95.9	
	Innocenti et al. [250]	2021	99.6			-	

and passed through the PointNet++ model [252] for 3D HAR. However, converting point clouds into voxel representations causes quantization errors and inefficient processing performance. Instead of quantizing the point clouds into voxels, Liu et al. [241] proposed MeteorNet, which directly stacks multi-frame point clouds and calculates local features by aggregating information from spatio-temporal neighboring points. Unlike MeteorNet [241], which learns spatio-temporal information by appending 1D temporal dimension to 3D points, PSTNet [246] disentangles the space and time to reduce the impacts of the spatial irregularity of points on temporal modeling. Fan et al. [245] introduced a point 4D convolution, followed by a transformer to capture the global appearance and motion information across the entire point cloud video. In the work of [244], a self-supervised learning framework was introduced to learn 4D spatio-temporal information from the point cloud sequence by predicting the temporal orders of 4D clips within the sequence. A 4D CNN model [241], [253] followed by an LSTM was utilized to predict the temporal order. Finally, the 4D CNN+LSTM network was fine-tuned on the existing HAR dataset to evaluate its performance. Wang et al. [243] proposed an anchor-based spatio-temporal attention convolution model to capture the dynamics of 3D point cloud sequences. However, these methods are not able to sufficiently capture the long-term relationships within point cloud sequences. To address this issue, Min et al. [242] introduced a modified version of LSTM unit named PointLSTM to update state information for neighbor point pairs to perform HAR.

In all, point clouds can effectively capture the 3D shapes and silhouettes of subjects, and can be used for HAR. Gener-

ally, 3D point cloud-based HAR methods can be insensitive to viewpoint variations, since viewpoint normalization can be conveniently performed by rotating the point cloud in the 3D space. However, point clouds often suffer from the presence of noise and high non-uniformity distribution of points, making it challenging for robust HAR. In addition, processing all the points within the point cloud sequence is often computationally expensive.

2.6 EVENT STREAM MODALITY

Event cameras, also known as neuromorphic cameras or dynamic-vision sensors, which can capture illumination changes and produce asynchronous events independently for each pixel [250], have received lots of attention recently. Different from conventional video cameras capturing the entire image arrays, event cameras only respond to changes in the visual scene. Taking a high-speed object as an example, traditional RGB cameras may not be able to capture enough information of the object, due to the low frame rate and motion blur. However, this issue can be significantly mitigated when using event cameras that operate at extremely high frequencies, generating events at a μs temporal scale [250]. Besides, event cameras have some other characteristics, such as high dynamic range, low latency, low power consumption, and no motion blur, which make them suitable for HAR. Particularly, event cameras are able to effectively filter out background information and keep foreground movement only, avoiding considerable redundancies in the visual information. However, the information obtained with event cameras is generally spatio-temporally sparse, and asynchronous. Common event cameras include the Dynamic Vision Sensor (DVS) [254] and the Dynamic and Active-pixel Vision Sensor (DAVIS) [255].

The output data of event cameras is highly different from that of conventional RGB cameras, as shown in Table 1. Thus, some of the existing methods [12], [250] mainly focused on designing event aggregation strategies converting the asynchronous output of the event camera into synchronous visual frames, which can then be processed with conventional computer vision techniques. While hand-crafted feature-based methods [256], [257] have been proposed, deep learning methods have been more popularly used recently. Given a fixed time interval Δt, Innocenti et al. [250] first built a sequence of binary representations from the raw event data by checking the presence or absence of an event for each pixel during Δt. These intermediate binary representations were then stacked together to form a single frame via a binary to decimal conversion. The sequence of such frames extracted from the whole event stream was finally fed to a CNN+LSTM for HAR. Huang et al. [248] utilized time-stamp image encoding to transform the event data sequence into frame-based representations, which were then fed to a CNN for HAR. More precisely, the value of each pixel in the generated time-stamp image encodes the amount of event taking place within a time-window. However, since the size of the frame to be processed is basically larger than the original Neuromorphic Vision Stream (NVS), the advantages of event cameras are diluted.

Therefore, different from the aforementioned methods, some other deep learning approaches [12], [247], [249], [258],

[259] have been proposed, which directly take the event data as input of deep neural networks. Ghosh et al. [12] learned a set of 3D spatio-temporal convolutional filters in an unsupervised manner to generate a structured matrix form of the raw event data, which was then fed to a 3D CNN for HAR. George et al. [258] utilized Spiking Neural Networks (SNNs) for event stream-based HAR. The work in [247] treated the event stream as a 3D point cloud, which was then fed to a PointNet [252] for gesture recognition. Recently, Bi et al. [259] represented events as graphs and used a GCN network for an end-to-end feature learning directly from the raw event data.

In general, the event stream modality is an emerging modality for HAR, and has received great attention in the past few years. Processing event data is computationally cheap, and the captured frames do not usually contain background information, which can be helpful for action understanding. However, the event stream data cannot generally be effectively and directly processable using conventional video analysis techniques, and thus effectively and directly utilizing the asynchronous event stream for HAR is still a challenging research problem.

2.7 AUDIO MODALITY

Audio signal is often available with videos for HAR. Due to the synchronization between the visual and audio streams, the audio data can be used to locate actions to reduce human labeling efforts and decrease computational costs. Some deep learning-based methods [13], [260] have been proposed to perform general activity recognition from audio signals. However, in recent years, only a few deep learning methods were proposed for HAR from audio signal alone. For example, Liang and Thomaz [13] used a pre-trained VGGish model [261] as the feature extractor followed by a deep classification network to perform HAR.

Using audio modality alone is not a very popular scheme for HAR, compared to other data modalities, since audio signal does not contain enough information for accurate HAR. Nevertheless, the audio modality can serve as complementary information for more reliable and efficient HAR, and thus most of the audio-based HAR methods [25], [262], [263], [264] focused on taking advantage of multi-modal deep learning techniques, which are discussed in Section 3.

2.8 ACCELERATION MODALITY

Acceleration signals obtained from accelerometers have been used for HAR [265], due to their robustness against occlusion, viewpoint, lighting, and background variations, etc. Specifically, a tri-axial accelerometer can return an estimation of acceleration along the x, y, and z axes, that can be used to perform human activity analysis [266]. As for the feasibility of using acceleration signal for HAR, although the size and proportion of the human body vary from person to person, people generally have similar qualitative ways to perform an action, so the acceleration signal usually does not have obvious intra-class variations for the same action. HAR using acceleration signal can generally achieve a high accuracy, and thus has been adopted for remote monitoring systems [267], [268] while taking care of privacy issues.

Recently, deep learning networks have been widely used for acceleration-based HAR. In [14], [269], [270], [271], the tri-axial acceleration data was fed to different CNN architectures for HAR. Wang et al. [272] proposed a framework consisting of CNN and Bi-LSTM networks to extract spatial and temporal features from the raw acceleration data. Unlike the above-mentioned works, Lu et al. [273] utilized a modified Recurrence Plot (RP) [274] to transform the raw tri-axial acceleration data into color images, which were then fed to a ResNet for HAR. Besides, some acceleration-based methods [275] focused on the fall detection task.

In general, the acceleration modality can be utilized for fine-grained HAR, and has been used for action monitoring, especially for elderly care due to its privacy-protecting characteristic. However, the subject needs to carry wearable sensors that are often cumbersome and disturbing. In addition, the position of the sensors on the human body can also affect the HAR performance.

2.9 RADAR MODALITY

Radar is an active sensing technology that transmits electromagnetic waves and receives returned waves from targets. Continuous-wave radars, such as Doppler [276] and Frequency Modulated Continuous Wave (FMCW) radars [277], are most often chosen for HAR. Specifically, Doppler radars detect the radial velocity of the body parts, and the frequency changes according to the distance, which is known as the Doppler shift. The micro-Doppler signatures generated by the micro-motion of the radar contain the motion and structure information of the target, which thus can be used for HAR. As for the FMCW radars, they can measure the distances of the targets as well. There are some advantages of using spectrograms obtained from radars for HAR, which include the robustness to variations in illumination and weather conditions, privacy protection, and the capability of through-wall HAR [22].

Several deep learning architectures have been proposed recently. The works in [15], [278], [279], [280], [281], [282] fed the micro-Doppler spectrogram images into CNNs for predicting action classes in different scenarios, such as aquatic activities [278], [280], different geometrical locations [282], and simulated environments [279]. Unlike the aforementioned methods, the work in [283] directly took the raw radar range data as input. The raw data was passed through an auto-correlation function, followed by a CNN to extract action-related features. Finally, a Random Forest classifier was used to predict the action class.

The two-stream architecture has also been investigated. Hernangómez et al. [284] designed a two-stream CNN taking both micro-Doppler and range spectrograms representing the structural signatures of the target as inputs. In the work of [285], micro-Doppler spectrograms and echos of the radar were fed to a two-stream CNN model for HAR.

By interpreting micro-Doppler spectrograms as temporal sequences rather than images, several RNN-based architectures [286], [287] have also been proposed recently. For example, Yang et al. [287] and Wang et al. [286] utilized an LSTM model and a stacked RNN model, respectively to predict the action classes.

In general, the characteristics and advantages of the radar modality make it suitable to be used for HAR in some

scenarios, but radars are relatively expensive. Though HAR using radar data has achieved satisfactory results on some datasets, there is still plenty of room for the development of radar-based methods, and the work of [22] also pointed out some future directions in this area, such as handling more complex actions in real-world scenarios with radar data.

2.10 WIFI MODALITY

WiFi is considered to be one of the most common indoor wireless signal types nowadays [288]. Since human bodies are good reflectors of wireless signal, WiFi signal can be utilized for HAR, and sometimes even for through-wall HAR [289]. There are some advantages of using the WiFi modality for action analysis, mainly due to the convenience, simplicity, and privacy protection of WiFi signal, and also the low cost of WiFi devices. Specifically, most of the existing WiFi-based HAR methods [290], [291] focused on using the Channel State Information (CSI) to conduct the HAR task. CSI is the fine-grained information computed from the raw WiFi signal, and the WiFi signal reflected by a person who performs an action, usually generates unique variations in the CSI on the WiFi receiver.

Deep learning-based HAR from the CSI signal has received attention recently. Wang et al. [292] proposed a deep sparse auto-encoder to learn discriminative features from the CSI streams. In the work of [16], the WiFi-based sample-level HAR model, named Temporal Unet, was proposed, which consists of several temporal convolutional, deconvolutional, and max pooling layers. This method classified every WiFi distortion sample in the series into one action for fine-grained HAR. LSTM networks have also been used for HAR using CSI signal [293], [294], [295]. Huang et al. [293] passed raw CSI measurements through a noise and redundancy removal module followed by a clip reconstruction module to segment the processed CSI signal into multiple clips. These clips were then fed to a multi-stream CNN+LSTM model to perform HAR. In the work of [294], the spatial features of the CSI signal were first extracted from a fully connected layer of a pre-trained CNN. These features were then fed to a Bi-LSTM to capture the temporal information for HAR. Chen et al. [295] directly passed the raw CSI signal through an attention-based Bi-LSTM to predict the action class. Different from the above-mentioned works, Gao et al. [296] transformed the CSI signal into radio images, which were then fed to a deep sparse auto-encoder to learn discriminative features for HAR.

In general, because of the advantages (e.g., convenience), the WiFi modality can be used for HAR in some scenarios. However, there are still some challenges which need to be further addressed, such as how to more effectively use the CSI phase and amplitude information, and to improve the robustness when handling dynamic environments.

Apart from the various modalities mentioned above, some other modalities, such as pressure [297], orientation [297], gyroscope [297], [298], Radio-Frequency [299], [300], Piezoelectric Energy Harvester [301], [302], and Electromyography [303], have also been used for HAR.

3 MULTI-MODALITY

In real life, humans often perceive the environment in a multi-modal cognitive way. Similarly, multi-modal machine

TABLE 5
Performance comparison of multi-modality fusion-based HAR methods on MSRDailyActivity3D (M), UTD-MHAD (U), and NTU RGB+D (N) datasets. S: Skeleton, D: Depth, IR: Infrared, PC: Point Cloud, Ac: Acceleration, Gyr: Gyroscope.

Method	Year	Modality	Fusion	Dataset			
				M	U	N	
						CS	CV
Imran et al. [305]	2016		Score	-	91.2	-	-
SFAM [306]	2017		Feature,Score	-	-	-	-
c-ConvNet [58]	2018		Feature,Score	-	-	86.4	89.1
GMVAR [307]	2019	RGB,D	Score	-	-	-	-
Dhiman et al. [308]	2020		Score	-	-	79.4	84.1
Wang et al. [309]	2020		Feature	-	-	89.5	91.7
Ren et al. [310]	2021		Score	-	-	89.7	93.0
ST-LSTM [8]	2017		Feature	-	-	73.2	80.6
Chain-MS [311]	2017		Feature	-	-	80.8	-
GRU+STA-Hands [312]	2017		Score	-	-	82.5	88.6
Zhao et al. [313]	2017		Feature	-	-	83.7	93.7
Baradel et al. [314]	2017		Score	90.0	-	84.8	90.6
SI-MM [315]	2018	RGB,S	Feature	91.9	-	92.6	97.9
Separable STA [316]	2019		Feature	-	-	92.2	94.6
SGM-Net [317]	2020		Score	-	-	89.1	95.9
VPN (RNX3D101) [318]	2020		Feature	-	-	95.5	98.0
PoseC3D [319]	2021		Feature,Score	-	-	97.0	-
JOLO-GCN [320]	2021		Score	-	-	93.8	98.1
Rahmani et al. [321]	2017		Feature	-	-	75.2	83.1
Kamel et al. [322]	2018	S,D	Score	88.1	-	-	-
3DSTCNN [323]	2019		Score	-	95.3	-	-
Rani et al. [324]	2021		Score	-	88.5	-	-
DSSCA-SSLM [325]	2017		Feature	97.5	-	74.9	-
Deep Bilinear [326]	2018		Feature	-	-	85.4	90.7
Cardenas et al. [23]	2018	RGB,S,D	Feature	94.6	-	-	-
Khaire et al. [327]	2018		Score	-	95.1	-	-
Khaire et al. [328]	2018		Score	-	95.4	-	-
Ye et al. [329]	2015	RGB,S,D,PC	Score	-	-	-	-
Ardianto et al. [330]	2018	RGB,D,IR	Score	-	-	-	-
FUSION-CPA [331]	2020	S,IR	Feature	-	-	91.6	94.5
Wang et al. [332]	2016		Feature,Score	-	-	-	-
Owens et al. [262]	2018		Feature	-	-	-	-
TBN [25]	2019		Feature	-	-	-	-
TSN+audio stream [25]	2019	RGB, Au	Score	-	-	-	-
AVSlowFast [264]	2020		Feature	-	-	-	-
Gao et al. [263]	2020		Feature	-	-	-	-
MAFnet [333]	2021		Feature	-	-	-	-
Dawar et al. [334]	2018	D,Ac,Gyr	Score	-	89.2	-	-
Dawar et al. [335]	2018	D,Ac,Gyr	Score	-	-	-	-
Wei et al. [336]	2019	RGB,Ac,Gyr	Score	-	95.6	-	-
Ahmad et al. [337]	2019	D,Ac,Gyr	Feature,Score	-	99.3	-	-
MGAF [338]	2020	D,Ac,Gyr	Feature	-	96.8	-	-
Pham et al. [339]	2021	S,Ac	Feature	-	97.0	-	-
DCNN [340]	2015	Ac,Gyr	Feature	-	-	-	-
DeepConvLSTM [341]	2016	Ac,Gyr	Feature	-	-	-	-
WiVi [288]	2019	RGB,WiFi	Score	-	-	-	-
Zou et al. [342]	2020	Ac,Gyr	Score	-	-	-	-
Memmesheimer et al [343]	2020	S,WiFi,etc.	Feature	-	93.3	-	-
Imran et al [234]	2020	RGB,S,Gyr	Feature	-	97.9	-	-

learning is a modeling approach aiming to process and relate the sensory information from multiple modalities [304]. By aggregating the advantages and capabilities of various data modalities, multi-modal machine learning can often provide more robust and accurate HAR. There are two main types of multi-modality learning methods, namely, fusion and co-learning. Fusion refers to the integration of information from two or more modalities for training and inference, while co-learning refers to the transfer of knowledge between different data modalities.

3.1 FUSION

As discussed in Section 2, different modalities can have different strengths. Thus it becomes a natural choice to take advantage of the complementary strengths of different data modalities via fusion, so as to achieve enhanced HAR performance. There are two widely used multi-modality fusion schemes in HAR, namely, score fusion and feature fusion. Generally, the score fusion [308] integrates the decisions that are separately made based on different modalities (e.g., by weighted averaging [324] or by learning a score fusion model [307]) to produce the final classification results. On

the other hand, the feature fusion [321] generally combines the features from different modalities to yield aggregated features that are often very discriminative and powerful for HAR. Note that data fusion, i.e., fusing the multi-modality input data before feature extraction [340], has also been exploited. Since the input data can be treated as the original raw features, here we simply categorize the data fusion approaches under feature fusion. Table 5 gives the results of multi-modality fusion-based HAR methods on the MSRDailyActivity3D [236], UTD-MHAD [298], and NTU RGB+D [161] benchmark datasets.

3.1.1 Fusion of Visual Modalities

With the emergence of low-cost RGB-D cameras, many multi-modality datasets [161], [236] have been created by the community, and consequently, some multi-modality fusion-based HAR methods have been proposed. Most of these methods focused on the fusion of visual modalities which are reviewed below.

Fusion of RGB and Depth Modalities. The RGB and depth videos respectively capture rich appearance and 3D shape information, that are complementary and can be used for HAR. Early methods [344], [345] focused on extracting hand-crafted features which capture spatio-temporal structural relationships from the RGB and depth modalities for more reliable HAR. Since the current mainstream of research focuses on deep learning architectures, we review, in the following, deep learning methods which fuse RGB and depth data modalities. In [305], a four-stream deep CNN was introduced to extract features from different representations of depth data (i.e., three Depth Motion Maps [219] from three different viewpoints) and RGB data (i.e., a Motion History Image [346]). The output scores of these four streams were fused to perform action classification. By considering the depth and RGB modalities as a single entity, Wang et al. [306] extracted scene flow features from the spatially aligned and temporally synchronized RGB and depth frames. The bidirectional rank pooling [55] was utilized to generate two dynamic images from the sequence of scene flow features. The dynamic images were then fed to two different CNNs, and finally, their classification scores were fused to perform HAR. Wang et al. [58] represented the RGB and depth data as two pairs of RGB and depth dynamic images, which were then passed through a cooperatively trained CNN (c-ConvNet). The c-ConvNet consists of two streams that exploit features for action classification by jointly optimizing a ranking loss and a softmax loss. In [309], a hybrid network that consists of multi-stream CNNs and 3D ConvLSTMs [347] was introduced to extract features from RGB and depth videos. These features were then fused via canonical correlation analysis to perform action classification. Wang et al. [307] proposed a generative framework to explore the feature distribution across the RGB and depth modalities. The fusion was performed by constructing a cross-modality discovery matrix, which was then fed to a modality correlation discovery network for the final prediction. Dhiman et al. [308] designed a two-stream network composed of a motion stream and a Shape Temporal Dynamic (STD) stream to encode features from RGB and depth videos, respectively. Particularly, the motion stream takes a dynamic image from the RGB data as input and outputs classification scores. The STD network consists of the Human Pose Model in [9] as its backbone, followed by several LSTMs and a softmax layer. The final classification scores were obtained by aggregating the scores of these two streams.

Fusion of RGB and Skeleton Modalities. The appearance information provided by the RGB data, and the body posture and joint motion information provided by the skeleton sequences, are complementary and useful for activity analysis. Thus, several works have investigated deep learning architectures to fuse RGB and skeleton data for HAR. Zhao et al. [313] introduced a two-stream deep network, which consists of an RNN and a CNN to process skeleton and RGB data, respectively. Both the feature fusion and the score fusion were evaluated and the former achieved better performance. The two-stream architecture was also investigated in [312], [314], where the classification scores from a CNN model [314] or an RNN model [312] trained on skeleton data and a skeleton conditioned spatio-temporal attention network trained on RGB data were fused for final classification. Zolfaghari [311] designed a three-stream 3D-CNN to process pose, motion, and the raw RGB images. The three streams were fused via a Markov chain model for action classification. Liu et al. [8] proposed a spatio-temporal LSTM network, which is able to effectively fuse the RGB and skeleton features within the LSTM unit. In order to effectively capture the subtle variations in action performing, Song et al. [315] proposed a learning framework containing two streams of skeleton-guided deep CNN to extract features from RGB and optical flow. Particularly, the local image patches around human body parts were first extracted from both RGB and flow sequences, and then fed into two individual CNNs followed by a part-aggregated pooling layer to generate two fixed-length feature vectors corresponding to RGB and flow frames. The skeleton data and the two sequences of RGB and flow feature vectors were then passed through a three-stream LSTM model followed by a fusion layer to obtain the final classification scores. Das et al. [316] introduced a pose guided spatio-temporal attention network on top of a 3D CNN model taking RGB videos as input to perform HAR. In their subsequent work [318], they extended their work in [316] by paying attention to the topology of the human body while computing the spatio-temporal attention maps. Li et al. [317] proposed a two-stream network, which consists of three main components namely the ST-GCN network [152] to extract the skeleton features, R(2+1)D network [109] to extract RGB features, and a guided block which takes these features to enhance the action-related information in RGB videos. Finally, the score fusion approach was utilized to perform classification. Cai et al. [320] introduced a two-stream GCN network, respectively taking skeleton data and joint-aligned flow patches obtained from RGB videos as the inputs for HAR. PoseC3D [319] represents skeleton data as a 3D heatmap volume and passes the RGB video and the heatmap volume through a two-stream 3D CNN framework whose architecture is inspired by [112].

Fusion of Skeleton and Depth Modalities. The skeleton data has been shown to be a succinct yet informative representation for human behavior analysis [348], which, however, is a very sparse representation without the encoding of the shape information of the human body and

the interacted objects. Besides, skeleton data is often noisy, limiting the performance of HAR when it is used alone [8]. Meanwhile, depth maps provide discriminative 3D shape and silhouette information that can be helpful for HAR. Therefore, early methods [349], [350] have attempted to fuse hand-crafted features extracted from both the skeleton and the depth sequences for more accurate HAR. Recently, deep learning-based approaches have become the mainstream for fusing the skeleton and depth modalities. In order to learn the relationships between the human body and the objects, as well as the relations between different human body parts, Rahmani et al. [321] proposed an end-to-end learning framework, which consists of a CNN followed by the bilinear compact pooling and fully-connected layers for action classification. In particular, the model takes the relative geometry between every body part and others as the skeleton features and depth image patches around different body parts as the appearance features to encode body part-object and body part-body part relations for reliable HAR. Kamel et al. [322] utilized a three-stream CNN to extract action-related features from Depth Motion Images (DMIs), Moving Joint Descriptors (MJDs), and their combination. The DMI and MJD encode the depth and skeleton sequences as images. The output scores of these three CNNs were fused to obtain the final classification scores. Zhao et al. [323] utilized two 3D CNN streams taking raw depth data and Depth Motion Maps [219] as inputs, and a manifold representation stream taking 3D skeleton as input, for feature extraction from depth and skeleton sequences. The classification scores from these three networks were fused via score multiplications. Rani et al. [324] proposed a three-stream 2D CNN to perform classification on three different hand-crafted features extracted from the depth and skeleton sequences, followed by a score fusion module to obtain the final classification result.

Fusion of RGB, Skeleton, and Depth Modalities. Several hand-crafted feature-based methods [345], [351] have explored the fusion of RGB, skeleton, and depth modalities to further enhance the robustness of HAR. Meanwhile, deep learning-based methods [23], [325], [326], [327], [328] have received greater attention due to their superior performance. Shahroudy et al. [325] focused on studying the correlation between modalities, and factorizing them into their correlated and independent components. Then a structured sparsity-based classifier was utilized for HAR. Hu et al. [326] learned the time-varying information across RGB, skeleton, and depth modalities by extracting temporal feature maps from each modality, and then concatenating them along the modality dimension. These multi-modal temporal features were then fed to a stack of bilinear blocks to exploit the mutual information from both modality and temporal directions. Khaire et al. [327] proposed a five-stream CNN network, which takes Motion History Images [346], Depth Motion Maps [219] (i.e., from three different viewpoints), and skeleton images generated respectively from RGB, depth, and skeleton sequences as inputs. Each CNN was trained individually, and the output scores of these five streams were fused with a weighted product model [352] to obtain the final classification scores. In their subsequent work [328], three fusion methods were explored for combining skeletal, RGB, and depth modalities. Car-

denas and Chavez [23] fed three different optical spectra channels from skeleton data [169] and dynamic images from both RGB and depth videos to a pre-trained CNN to extract multi-modal features. Finally, a feature aggregation module was used to perform classification.

Other Fusion Methods. Besides, some other multi-modality fusion methods were proposed [330], [331]. In the work of [330], a multi-modal fusion model was built on top of the TSN framework [50]. The RGB, depth, infrared, and optical flow sequences were passed through TSNs to perform initial classification, followed by a fusion network applied on the initial classification scores to obtain the final classification scores. Main and Noumeir [331] utilized a 3D CNN and a 2D CNN to extract features from infrared and skeleton data, respectively. The feature vectors were then concatenated and exploited by a multi-layer perceptron for HAR.

3.1.2 Fusion of Visual and Non-visual Modalities

Visual and non-visual modalities can also be fused to leverage their complementary discriminative capabilities for more accurate and robust HAR models.

Fusion of Audio and Visual Modalities. Audio data provides complementary information to the appearance and motion information in the visual data. Several deep learning-based methods have been proposed to fuse these two types of modalities for HAR. Wang et al. [332] introduced a three-stream CNN to extract multi-modal features from audio signal, RGB frames, and optical flows. Both feature fusion and score fusion were evaluated, and the former achieved better performance. Owens and Efros [262] trained a two-stream CNN in a self-supervised manner to identify any misalignment between the audio and visual sequences. The learned model was then fine-tuned on the HAR datasets for audio-visual HAR. Inspired by TSN [50], Kazakos et al. [25] introduced a Temporal Binding Network (TBN), which takes audio, RGB, and optical flow as inputs for egocentric HAR. The TBN network contains a three-stream CNN to fuse the multi-modal input sequences within each Temporal Binding Window, followed by a temporal aggregation module for classification. Their results showed that the TBN outperformed the Temporal Segment Network (TSN) [50] for the audio-visual HAR task. Gao et al. [263] utilized audio signal to reduce temporal redundancies in videos. Particularly, this method distills the knowledge from a teacher network trained on video clips to a student network trained on image-audio pairs for efficient HAR. The student network is a two-stream deep model which takes the starting frame and audio signal of the video clip as input followed by a feature fusion, fully-connected, and softmax layers for audio-visual HAR. Inspired by the work of [112], Xiao et al. [264] introduced a hierarchically integrated audio-visual representation framework containing slow and fast visual pathways that are deeply integrated with a faster audio pathway at multiple layers. Two training strategies, including randomly dropping the audio pathway and hierarchical audio-visual synchronization, were also introduced to enable joint training of audio and video modalities.

Fusion of Acceleration and Visual Modalities. Some hand-crafted feature-based methods [353], [354] have exploited the fusion of acceleration and visual modalities for HAR. Besides, many deep learning methods [334], [335], [336],

[337], [338], [339] have also been proposed recently. Dawar et al. [335] represented inertial signal as an image, and utilized two CNNs to fuse the depth images and inertial signals using score fusion. Due to the limited depth-inertial training data, Dawar et al. [334] proposed a data augmentation framework based on depth and inertial modalities, which were separately fed to a CNN and a CNN+LSTM respectively. The scores of the two models were fused during testing for better classification. Wei et al. [336] respectively fed the 3D video frames and 2D inertial images to a 3D CNN and a 2D CNN for HAR, and the score fusion achieved better performance than feature fusion. Several depth-inertial fusion techniques have also been investigated in [337], [338] by designing different two-stream CNN architectures, where the inertial signal was transformed into images using the technique in [340].

Other Fusion Methods. Memmesheimer et al. [343] transformed the skeleton, inertial, and WiFi data to a color image, which was fed to a CNN to perform HAR. Imran and Raman [234] designed a three-stream architecture, where a 1D CNN, a 2D CNN, and an RNN were used for gyroscopic, RGB, and skeleton data, respectively. Several fusion methods were adopted to predict the final class label, while fusing features using canonical correlation analysis achieved the best performance. Zou et al. [288] designed a two-stream architecture, named WiVi, which takes the RGB and WiFi frames as inputs of the C3D and CNN streams. A multimodal score fusion module was built on top of the network to perform HAR.

Although the aforementioned multi-modality fusion-based HAR methods have achieved promising results on some benchmark datasets, the task of effective modality fusion is still largely open. Specifically, most of the existing multi-modality methods have complicated architectures which require high computational costs. Thus efficient multi-modality HAR also needs to be addressed.

3.2 CO-LEARNING

Co-learning explores how knowledge learned from auxiliary modalities can be used to assist the learning of a model on another modality [304]. Transferring knowledge among different modalities can overcome the shortcomings of a single data modality and enhance its performance. Unlike fusion methods, in co-learning methods, the data of the auxiliary modalities is only required during training rather than testing. This is particularly beneficial in cases where some modalities are missing during testing. Co-learning can also benefit the learning of a certain modality with fewer samples, by leveraging other correlated modalities with richer samples for assisting model training.

3.2.1 Co-Learning with Visual Modalities

Most of the co-learning-based HAR methods focused on co-learning with visual modalities, such as RGB with depth modalities, and RGB with skeleton modalities.

Co-Learning with RGB and Depth Modalities. Knowledge transfer [304] between the RGB modality which captures appearance information, and depth data which encodes 3D shape information, has been shown to be useful for improving the representation capability of each modality for HAR, especially when one of these modalities has a limited amount of annotated data for training. There have been some methods that used hand-crafted features for co-learning [344], [345]. Recently, Garcia et al. [355], [356], [357] proposed several deep learning-based HAR methods based on cross-modality knowledge distillation. In [355], a knowledge distillation framework was proposed to distill knowledge from a teacher network taking depth videos as input to a RGB-based student network. The knowledge distillation was achieved by forcing the feature maps and prediction scores of the student network to be similar to the teacher network. While in [357], an adversarial learning-based knowledge distillation strategy was proposed to learn the student stream. In their subsequent work [356], a three-stream network taking RGB, depth, and optical flow data as inputs, was trained using a cooperative learning strategy, i.e., the predicted labels generated by the modality stream with the minimum classification loss were used as an additional supervision for the training of other streams.

Co-Learning with RGB and Skeleton Modalities. Co-learning between the RGB and skeleton modalities has also been investigated in existing works [24], [26], [358]. Mahasseni and Todorovic [24] utilized a CNN+LSTM network to perform classification based on RGB videos, and an LSTM model trained on skeleton data to act as a regularizer by forcing the output features of the two models (i.e., CNN+LSTM and LSTM) to be similar. Thoker and Gall [358] utilized TSN [50] as the teacher network taking RGB videos, and an ensemble of student networks (e.g., ST-GCN [152] and HCN [174]) handling skeleton data, to perform knowledge distillation for cross-modality HAR. Specifically, each student network was trained from the supervisory information provided by the teacher network as well as other student networks. In [26], a supervision signal from an auxiliary LSTM network taking the skeleton data as input was fed to a two-stream CNN+LSTM architecture taking RGB and optical flow sequences as inputs. The auxiliary supervision signal forces the model to align the distributions of the auxiliary (skeleton) and source (RGB and optical flow) modalities, which thus enhances the representation of the RGB and optical flow modalities.

Other Co-Learning Methods. Besides using RGB with depth or RGB with skeleton data for co-learning, other visual modalities have also been investigated. For example, Wang et al. [359] learned a transferable generative model which takes an infrared video as input and generates a fake feature representation of its corresponding RGB video. The discriminator consists of two sub-networks, i.e., a deep binary classifier which attempts to differentiate between the generated fake features and the real RGB feature representation, and a predictor which takes both the infrared video representation and the generated features as inputs to perform action classification. Chadha et al. [360] embedded the PIX2NVS emulator [361] (i.e., converting pixel domain video frames to neuromorphic events) into a teacher-student framework, which transfers knowledge from a pre-trained optical flow teacher network to a neuromorphic event student network.

3.2.2 Co-Learning with Visual and Non-visual Modalities

There are also several works on co-learning between visual and non-visual modalities [297], [362], [363]. Kong et al. [297] and Liu et al. [362] trained multiple teacher networks on non-visual modalities, such as acceleration, gyroscope, and orientation signal, to teach an RGB video-based student network through knowledge distillation with an attention mechanism. In [297], the knowledge was transferred by first integrating the classification scores of the teachers with attention weights learned by feeding the concatenated classification scores of the teachers to a feed-forward neural network. Then the attended-fusion scores were used as an additional supervision to train the student network. In [362], knowledge distillation was achieved by forcing the visual explanations (i.e., attention maps that highlight the important regions for the classification scores) of both the student and teacher networks to be similar to bridge the modality gap. In another work, Perez et al. [363] proposed a knowledge distillation framework to transfer knowledge from a teacher network trained on RGB videos to a student network taking the raw sound data as input. Specifically, an objective function [364], which encourages the student model to predict the true labels as well as matching the soft labels provided by the teacher network, was utilized.

Besides fusing multiple modalities and transferring knowledge between different modalities, there are also a few works that leverage the correlation between different modalities for self-supervised learning. For example, Al-wassel et al. [365] leveraged unsupervised clustering in the audio/video modality as the supervisory signal for the video/audio modality, respectively. The audio-visual correspondence and temporal synchronization have also been used as self-supervised signals to learn discriminative audio and visual features [366], [367]. Several other works [368], [369] leverage narrations of videos as a weak supervisory signal to jointly learn video and text representations for action recognition and detection. Miech et al. [368] introduced a text-video embedding model, which is trained from the caption-clip pairs using the max-margin ranking loss for learning joint representations of video and language automatically. In their further study [369], a novel training loss derived from Multiple Instance Learning and Noise Contrastive Estimation, was introduced to address misalignments in narrated videos. Besides, Sun et al. [370] introduced a joint visual-linguistic model by extending BERT [371] to videos for zero-shot HAR.

4 DATASETS

A large number of datasets have been created to train and evaluate the HAR methods. Table 6 lists a series of benchmark datasets. In particular, the attributes of these datasets are also summarized.

For RGB-based HAR, the UCF101 [372], HMDB51 [27], and Kinectis-400 [373] are widely used as benchmark datasets. Besides, Kinetics-600 [374], Kinetics-700 [375], EPIC-KITCHENS-55 [376], THUMOS Challenge 15 [377], ActivityNet [378], and Something-Something-v1 [379] datasets are also popularly used. For 3D skeleton, depth, infrared, and point cloud-based HAR, the large-scale NTU RGB+D [161] and NTU RGB+D 120 [151] are widely

TABLE 6
Some representative benchmark datasets with various data modalities for HAR. S: Skeleton, D: Depth, IR: Infrared, PC: Point Cloud, ES: Event Stream, Au: Audio, Ac: Acceleration, Gyr: Gyroscope, EMG: Electromyography.

Dataset	Year	Modality	#Class	#Subject	#Sample	#Viewpoint
KTH [380]	2004	RGB	6	25	2,391	1
Weizmann [41]	2005	RGB	10	9	90	1
IXMAS [381]	2006	RGB	11	10	330	5
HDM05 [382]	2007	RGB,S	130	5	2,337	1
Hollywood [383]	2008	RGB	8	-	430	-
Hollywood2 [384]	2009	RGB	12	-	3,669	-
MSR-Action3D [239]	2010	S,D	20	10	567	1
Olympic [385]	2010	RGB	16	-	783	-
CAD-60 [386]	2011	RGB,S,D	12	4	60	-
HMDB51 [27]	2011	RGB	51	-	6,766	-
RGB-HuDaAct [29]	2011	RGB,D	13	30	1,189	1
ACT4^2 [387]	2012	RGB,D	14	24	6,844	4
DHA [388]	2012	RGB,D	17	21	357	1
MSRDailyActivity3D [236]	2012	RGB,S,D	16	10	320	1
UCF101 [372]	2012	RGB	101	-	13,320	1
UTKinect [389]	2012	RGB,S,D	10	10	200	1
Berkeley MHAD [33]	2013	RGB,S,D,Au,Ac	12	12	660	4
CAD-120 [390]	2013	RGB,S,D	10	4	120	-
IAS-lab [391]	2013	RGB,S,D,PC	15	12	540	1
J-HMDB [392]	2013	RGB,S	21	-	31,838	1
MSRAction-Pair [218]	2013	RGB,S,D	12	10	360	1
UCFKinect [393]	2013	S	16	16	1,280	1
Multi-View TJU [394]	2014	RGB,S,D	20	22	7,040	2
Northwestern-UCLA [237]	2014	RGB,S,D	10	10	1,475	3
Sports-1M [45]	2014	RGB	487	-	1,113,158	-
UPCV [395]	2014	S	10	20	400	1
UWA3D Multiview [251], [396]	2014	RGB,S,D	30	10	∼900	4
ActivityNet [378]	2015	RGB	203	-	27,801	-
SYSU 3D HOI [351]	2015	RGB,S,D	12	40	480	1
THUMOS Challenge 15 [377]	2015	RGB	101	-	24,017	-
TJU [397]	2015	RGB,S,D	15	20	1,200	1
UTD-MHAD [298]	2015	RGB,S,D,Ac,Gyr	27	8	861	1
UWA3D Multiview II [238]	2015	RGB,S,D	30	10	1,075	4
Charades [398]	2016	RGB	157	267	9,848	-
InfAR [30]	2016	IR	12	40	600	2
NTU RGB+D [161]	2016	RGB,S,D,IR	60	40	56,880	80
YouTube-8M [399]	2016	RGB	4,800	-	8, 264, 650	-
Something-Something-v1 [379]	2017	RGB	174	-	108,499	-
AVA [400]	2017	RGB	80	-	437	-
DvsGesture [240]	2017	ES	17	29	-	-
FCVID [401]	2017	RGB	239	-	91,233	-
Kinetics-400 [373]	2017	RGB	400	-	306,245	-
NEU-UB [345]	2017	RGB,D	6	20	600	-
PKU-MMD [28]	2017	RGB,S,D,IR	51	66	1,076	3
UniMiB SHAR [402]	2017	Ac	17	30	11,771	-
EPIC-KITCHENS-55 [376]	2018	RGB,Au	-	32	39,594	Egocentric
Kinetics-600 [374]	2018	RGB	600	-	495,547	-
RGB-D Varying-view [403]	2018	RGB,S,D	40	118	25,600	8+1(360°)
DHP19 [32]	2019	ES, S	33	17	-	4
Drive&Act [404]	2019	RGB,S,D,IR	83	15	-	6
EV-Action [303]	2019	RGB,S,D,EMG	20	70	7,000	9
Hernangómez et al. [284]	2019	Radar	8	11	1,056	-
Kinetics-700 [375]	2019	RGB	700	-	650,317	-
Kitchen20 [405]	2019	Au	20	-	800	-
MMAct [297]	2019	RGB,S,Ac,Gyr,etc.	37	20	36,764	4+Egocentric
Moments in Time [406]	2019	RGB	339	-	∼1,000,000	-
Wang et al. [407]	2019	WiFi CSI	6	1	1,394	-
NTU RGB+D 120 [151]	2019	RGB,S,D,IR	120	106	114,480	155
IKEA ASM [408]	2020	RGB,S,D	33	48	16,764	3
EPIC-KITCHENS-100 [409]	2020	RGB,Au,Ac	-	45	89,979	Egocentric
ETRI-Activity3D [410]	2020	RGB,S,D	55	100	112,620	-
RareAct [411]	2020	RGB	122	-	905	-
BABEL [412]	2021	Mocap	252	-	13,220	-
HOMAGE [413]	2021	RGB,,IR,Ac,Gyr,etc.	75	27	1,752	2∼5
UAV-Human [414]	2021	RGB,S,D,IR,etc.	155	119	67,428	-

used benchmark datasets. Besides, the MSRDailyActivity3D [236], Northwestern-UCLA [237], and UWA3D Multiview II [238] datasets are also widely used for depth-based HAR, while the InfAR [30] dataset is often used for infrared-based HAR. Some point cloud-based methods [241], [244] have also been evaluated on point cloud data obtained from depth images. DvsGesture [240] and DHP19 [32] datasets are popular datasets for event stream-based HAR. However, there are no widely used benchmark datasets for non-visual modality-based HAR.

In addition, for multi-modality-based HAR, the NTU RGB+D [161], NTU RGB+D 120 [151], and MMAct [297] are large benchmark datasets suitable for fusion and co-learning of different modalities. The MSRDailyActivity3D [236], UTD-MHAD [298], and PKU-MMD [28] datasets are also popularly used.

5 DISCUSSION

In the previous sections, we review the methods and datasets for HAR with various data modalities. Below we discuss some of the potential and important directions that could need further investigation in this domain.

Datasets. Large and comprehensive datasets generally have a vital importance for the development of HAR, especially for the deep learning-based HAR methods. There are many aspects which indicate the quality of a dataset, such as its size, diversity, applicability, and type of modality. Despite the large number of existing datasets that have advanced the HAR area greatly, to further facilitate the research on HAR, new benchmark datasets are still required. For example, most of the existing multi-modality datasets were collected in controlled environments, where actions were usually performed by volunteers. Thus collecting multi-modality data from uncontrolled environments to achieve large and challenging benchmarks, for further promoting multi-modality HAR in practical applications, can be important. Besides, the construction of large and challenging datasets for every person's action recognition in crowded environments can also be further investigated.

Multi-modality Learning. As discussed in Section 3, several multi-modality learning methods, including multi-modality fusion and cross-modality transfer learning, have been proposed for HAR. The fusion of multi-modality data which can often complement each other, results in improvements of HAR performance, while co-learning can be used to handle the issue of the lack of data of some modalities. However, as pointed out by [415], many existing multi-modality methods are not as effective as expected owing to a series of challenges, such as over-fitting. This implies there remain opportunities to design more effective fusion and co-learning strategies for multi-modality HAR.

Efficient Action Analysis. The superior performance of many HAR methods is built on high computational complexity, while efficient HAR is also crucial for many real-life practical applications. Hence how to reduce the computational costs and resource consumption (e.g., CPU, GPU, and energy consumption), and achieve efficient and fast HAR, deserve further studies [134], [202].

Early Action Recognition. Early action recognition (action prediction) enables recognition when only a part of the action has been performed, i.e., recognizing an action before it has been fully performed [416], [417], [418]. This is also an important problem due to its relevance in some applications, such as online human-robot interaction and early alarm in some real-life scenarios.

Few-shot Action Analysis. It can be difficult to collect a large amount of training data (especially multi-modality data) for all action classes. To handle this issue, one of the possible solutions is to take advantage of few-shot learning techniques [419]. Though there have been some attempts for few-shot HAR [100], [151], considering the significance of handling the issues of scarcity of data in many practical scenarios, more advanced few-shot action analysis can still be further explored.

Unsupervised and Semi-supervised Learning. Supervised learning methods, especially deep learning-based ones, often require a large amount of data with expensive labels for model training. Meanwhile, unsupervised and semi-supervised learning techniques [420], [421], [422] often enable to leverage the availability of unlabelled data to train the models, which significantly reduces the requirement of large labeled datasets. Since unlabelled action samples are often easier to be collected than labeled ones, unsupervised

and semi-supervised HAR are also an important research direction that is worthy of further development.

6 CONCLUSION

HAR is an important task that has attracted significant research attention in the past decades, and various data modalities with different characteristics have been used for this task. In this paper, we have given a comprehensive review of HAR methods using different data modalities. Besides, multi-modality recognition methods, including fusion and co-learning methods, have been also surveyed. Benchmark datasets have also been reviewed, and some potential research directions have been discussed.

REFERENCES

[1] W. Lin, M.-T. Sun, R. Poovandran, and Z. Zhang, "Human activity recognition for video surveillance," in *ISCAS*, 2008.

[2] M. Lu, Y. Hu, and X. Lu, "Driver action recognition using deformable and dilated faster r-cnn with optimized region proposals," *Appl. Intell.*, vol. 50, no. 4, 2020.

[3] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank, "Semantic-based surveillance video retrieval," *TIP*, vol. 16, no. 4, 2007.

[4] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, and P. Maragos, "Multimodal human action recognition in assistive human-robot interaction," in *ICASSP*, 2016.

[5] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR*, 2011.

[6] R. Poppe, "A survey on vision-based human action recognition," *Image Vis Comput*, 2010.

[7] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015.

[8] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal lstm network with trust gates," *TPAMI*, vol. 40, no. 12, 2018.

[9] H. Rahmani and A. Mian, "3d action recognition from novel viewpoints," in *CVPR*, 2016.

[10] Z. Jiang, V. Rozgic, and S. Adali, "Learning spatiotemporal features for infrared action recognition with 3d convolutional neural networks," in *CVPRW*, 2017.

[11] Y. Wang, Y. Xiao, F. Xiong, W. Jiang, Z. Cao, J. T. Zhou, and J. Yuan, "3dv: 3d dynamic voxel for action recognition in depth video," in *CVPR*, 2020.

[12] R. Ghosh, A. Gupta, A. Nakagawa, A. Soares, and N. Thakor, "Spatiotemporal filtering for event-based action recognition," *arXiv preprint arXiv:1903.07067*, 2019.

[13] D. Liang and E. Thomaz, "Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos," *IMWUT*, vol. 3, no. 1, 2019.

[14] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," in *MobiCASE*, 2014.

[15] Y. Kim and T. Moon, "Human detection and activity classification based on micro-doppler signatures using deep convolutional neural networks," *GRSL*, vol. 13, no. 1, 2015.

[16] F. Wang, Y. Song, J. Zhang, J. Han, and D. Huang, "Temporal unet: Sample level human action recognition using wifi," *arXiv preprint arXiv:1904.11953*, 2019.

[17] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *CSUR*, vol. 43, no. 3, 2011.

[18] B. Ren, M. Liu, R. Ding, and H. Liu, "A survey on 3d skeleton-based action recognition using learning method," *arXiv preprint arXiv:2002.05907*, 2019.

[19] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recognit. Lett.*, vol. 34, no. 15, 2013.

[20] S. Slim, A. Atia, M. Elfattah, and M. Mostafa, "Survey on human activity recognition based on acceleration data," *Intl. J. Adv. Comput. Sci. Appl*, vol. 10, 2019.

[21] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A survey on behavior recognition using wifi channel state information," *IEEE Commum. Mag.*, vol. 55, no. 10, 2017.

[22] X. Li, Y. He, and X. Jing, "A survey of deep learning-based human activity recognition in radar," *Remote Sensing*, vol. 9, 2019.

[23] E. J. E. Cardenas and G. C. Chavez, "Multimodal human action recognition based on a fusion of dynamic images using cnn descriptors," in *SIBGRAPI*, 2018.

[24] B. Mahasseni and S. Todorovic, "Regularizing long short term memory with 3d human-skeleton sequences for action recognition," in *CVPR*, 2016.

[25] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *ICCV*, 2019.

[26] S. Song, J. Liu, Y. Li, and Z. Guo, "Modality compensation network: Cross-modal adaptation for action recognition," *TIP*, vol. 29, 2020.

[27] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *ICCV*, 2011.

[28] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding," *arXiv preprint arXiv:1703.07475*, 2017.

[29] B. Ni, G. Wang, and P. Moulin, "Rgbd-hudaact: A color-depth video database for human daily activity recognition," in *ICCVW*, 2011.

[30] C. Gao, Y. Du, J. Liu, J. Lv, L. Yang, D. Meng, and A. G. Hauptmann, "Infar dataset: Infrared action recognition at different times," *Neurocomputing*, vol. 212, 2016.

[31] H. Cheng and S. M. Chung, "Orthogonal moment-based descriptors for pose shape query on 3d point cloud patches," *Pattern Recognition*, vol. 52, 2016.

[32] E. Calabrese, G. Taverni, C. Awai Easthope, S. Skriabine, F. Corradi, L. Longinotti, K. Eng, and T. Delbruck, "Dhp19: Dynamic vision sensor 3d human pose dataset," in *CVPRW*, 2019.

[33] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *WACV*, 2013.

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

[35] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of wifi signal based human activity recognition," in *MobiCom*, 2015.

[36] K. Soomro and A. R. Zamir, "Action recognition in realistic sports videos," in *Computer vision in sports*, 2014.

[37] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *CVIU*, vol. 117, no. 6, 2013.

[38] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: a study of bag-of-features and part-based representations," in *BMVC*, 2010.

[39] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *CVPR*, 2010.

[40] G. Sharma, F. Jurie, and C. Schmid, "Discriminative spatial saliency for image classification," in *CVPR*, 2012.

[41] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *TPAMI*, vol. 29, no. 12, 2007.

[42] I. Laptev, "On space-time interest points," *IJCV*, vol. 64, no. 2-3, 2005.

[43] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *CVPR*, pp. 3169–3176, 2011.

[44] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NeurIPS*, 2014.

[45] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[46] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, "Action recognition with dynamic image networks," *TPAMI*, vol. 40, no. 12, 2017.

[47] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *CVPR*, 2015.

[48] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *IJCV*, vol. 105, no. 3, 2013.

[49] G. Chéron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," in *ICCV*, 2015.

[50] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *ECCV*, 2016.

[51] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "Actionvlad: Learning spatio-temporal aggregation for action classification," in *CVPR*, 2017.

[52] A. Diba, V. Sharma, and L. Van Gool, "Deep temporal linear encoding networks," in *CVPR*, 2017.

[53] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *CVPR*, 2017.

[54] M. Zong, R. Wang, X. Chen, Z. Chen, and Y. Gong, "Motion saliency based multi-stream multiplier resnets for action recognition," *Image Vis Comput*, vol. 107, 2021.

[55] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *TPAMI*, vol. 39, no. 4, 2016.

[56] P. Wang, S. Wang, Z. Gao, Y. Hou, and W. Li, "Structured images for rgb-d action recognition," in *ICCVW*, 2017.

[57] Y. Xiao, J. Chen, Y. Wang, Z. Cao, J. T. Zhou, and X. Bai, "Action recognition for depth video using multi-view dynamic images," *Inf. Sci.*, vol. 480, 2019.

[58] P. Wang, W. Li, J. Wan, P. Ogunbona, and X. Liu, "Cooperative training of deep aggregation networks for rgb-d action recognition," in *AAAI*, 2017.

[59] X. Wang, A. Farhadi, and A. Gupta, "Actions ~ transformations," in *CVPR*, 2016.

[60] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, "Real-time action recognition with enhanced motion vector cnns," in *CVPR*, 2016.

[61] A. Piergiovanni and M. S. Ryoo, "Representation flow for action recognition," in *CVPR*, 2019.

[62] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream convnets," *arXiv preprint arXiv:1507.02159*, 2015.

[63] A. Kar, N. Rai, K. Sikka, and G. Sharma, "Adascan: Adaptive scan pooling in deep convolutional neural networks for action recognition in videos," in *CVPR*, 2017.

[64] H. Zhang, D. Liu, and Z. Xiong, "Two-stream action recognition-oriented video super-resolution," in *ICCV*, 2019.

[65] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *CVPR*, 2016.

[66] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.

[67] W. Du, Y. Wang, and Y. Qiao, "Rpan: An end-to-end recurrent pose-attention network for action recognition in videos," in *ICCV*, 2017.

[68] L. Sun, K. Jia, K. Chen, D.-Y. Yeung, B. E. Shi, and S. Savarese, "Lattice long short-term memory for human action recognition," in *ICCV*, 2017.

[69] T. Perrett and D. Damen, "Ddlstm: Dual-domain lstm for cross-dataset action recognition," in *CVPR*, 2019.

[70] Y. Meng, C.-C. Lin, R. Panda, P. Sattigeri, L. Karlinsky, A. Oliva, K. Saenko, and R. Feris, "Ar-net: Adaptive frame resolution for efficient action recognition," in *ECCV*, 2020.

[71] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *CVPR*, 2015.

[72] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *ICML*, 2015.

[73] Z. Wu, C. Xiong, Y.-G. Jiang, and L. S. Davis, "Liteeval: A coarse-to-fine framework for resource efficient video recognition," in *NeurIPS*, 2019.

[74] M. Majd and R. Safabakhsh, "Correlational convolutional lstm for human action recognition," *Neurocomputing*, vol. 396, 2020.

[75] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional lstm with cnn features," *IEEE Access*, vol. 6, 2017.

[76] J.-Y. He, X. Wu, Z.-Q. Cheng, Z. Yuan, and Y.-G. Jiang, "Db-lstm: Densely-connected bi-directional lstm for human action recognition," *Neurocomputing*, vol. 444, 2021.

[77] H. Zhao and X. Jin, "Human action recognition based on improved fusion attention cnn and rnn," in *ICCIA*, 2020.

[78] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv:1511.04119*, 2015.

[79] H. Ge, Z. Yan, W. Yu, and L. Sun, "An attention mechanism based convolutional lstm network for video action recognition," *Multimed. Tools. Appl.*, vol. 78, no. 14, 2019.

[80] S. Sudhakaran, S. Escalera, and O. Lanz, "Lsta: Long short-term attention for egocentric action recognition," in *CVPR*, 2019.

[81] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," in *NeurIPS*, 2017.

[82] L. Meng, B. Zhao, B. Chang, G. Huang, W. Sun, F. Tung, and L. Sigal, "Interpretable spatio-temporal attention for video action recognition," in *ICCVW*, 2019.

[83] Z. Wu, C. Xiong, C.-Y. Ma, R. Socher, and L. S. Davis, "Adaframe: Adaptive frame selection for fast video recognition," in *CVPR*, 2019.

[84] Z. Liu, Z. Li, R. Wang, M. Zong, and W. Ji, "Spatiotemporal saliency-based multi-stream networks with attention-aware lstm for action recognition," *Neural. Comput. Appl.*, 2020.

[85] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. Snoek, "Videolstm convolves, attends and flows for action recognition," *CVIU*, vol. 166, 2018.

[86] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.

[87] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," in *ICLR*, 2016.

[88] Y. Shi, Y. Tian, Y. Wang, W. Zeng, and T. Huang, "Learning long-term dependencies for action recognition with a biologically-inspired deep network," in *ICCV*, 2017.

[89] D. Dwibedi, P. Sermanet, and J. Tompson, "Temporal reasoning in videos using convolutional gated recurrent units," in *CVPRW*, 2018.

[90] P.-S. Kim, D.-G. Lee, and S.-W. Lee, "Discriminative context learning with gated recurrent unit for group activity recognition," *Pattern Recognition*, vol. 76, 2018.

[91] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *EMNLP*, 2014.

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

[93] Z. Wu, X. Wang, Y-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *ACM Multimedia*, 2015.

[94] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two stream lstm: A deep fusion framework for human action recognition," in *WACV*, 2017.

[95] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *HBU*, 2011.

[96] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level cnn: saliency-aware 3-d cnn with lstm for video action recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, 2016.

[97] S. Vyas, Y. S. Rawat, and M. Shah, "Multi-view action recognition using cross-view video prediction," in *ECCV*, 2020.

[98] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, "Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length," *TMM*, vol. 20, no. 3, 2017.

[99] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *TPAMI*, vol. 35, no. 1, 2012.

[100] H. Zhang, L. Zhang, X. Qui, H. Li, P. H. Torr, and P. Koniusz, "Few-shot action recognition with permutation-invariant attention," in *ECCV*, 2020.

[101] X. Li, B. Shuai, and J. Tighe, "Directional temporal modeling for action recognition," in *ECCV*, 2020.

[102] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. Van Gool, "Temporal 3d convnets: New architecture and transfer learning for video classification," *arXiv preprint arXiv:1711.08200*, 2017.

[103] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[104] A. Diba, M. Fayyaz, V. Sharma, M. Mahdi Arzani, R. Yousefzadeh, J. Gall, and L. Van Gool, "Spatio-temporal channel correlation networks for action classification," in *ECCV*, 2018.

[105] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *TPAMI*, 2017.

[106] N. Hussein, E. Gavves, and A. W. Smeulders, "Timeception for complex action recognition," in *CVPR*, 2019.

[107] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018.

[108] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017.

[109] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *CVPR*, 2018.

[110] Y. Zhou, X. Sun, Z.-J. Zha, and W. Zeng, "Mict: Mixed 3d/2d convolutional tube for human action recognition," in *CVPR*, 2018.

[111] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann, "Hidden two-stream convolutional networks for action recognition," in *ACCV*, 2018.

[112] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *ICCV*, 2019.

[113] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal residual networks for video action recognition. corr abs/1611.02155 (2016)," *arXiv preprint arXiv:1611.02155*, 2016.

[114] J. Li, X. Liu, M. Zhang, and D. Wang, "Spatio-temporal deformable 3d convnets with attention for action recognition," *Pattern Recognition*, vol. 98, 2020.

[115] M. Zolfaghari, K. Singh, and T. Brox, "Eco: Efficient convolutional network for online video understanding," in *CVPR*, 2018.

[116] L. Wang, W. Li, W. Li, and L. Van Gool, "Appearance-and-relation networks for video classification," in *CVPR*, 2018.

[117] B. Martinez, D. Modolo, Y. Xiong, and J. Tighe, "Action recognition with spatial-temporal discriminative filter banks," in *ICCV*, 2019.

[118] Y. Zhou, X. Sun, C. Luo, Z.-J. Zha, and W. Zeng, "Spatiotemporal fusion in 3d cnns: A probabilistic view," in *CVPR*, 2020.

[119] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *CVPR*, 2020.

[120] J. Kim, S. Cha, D. Wee, S. Bae, and J. Kim, "Regularization on spatio-temporally smoothed feature for action recognition," in *CVPR*, 2020.

[121] A. Piergiovanni and M. S. Ryoo, "Recognizing actions in videos from unseen viewpoints," in *CVPR*, 2021.

[122] G. Varol, I. Laptev, C. Schmid, and A. Zisserman, "Synthetic humans for action recognition from unseen viewpoints," *IJCV*, 2021.

[123] H. Rahmani, A. Mian, and M. Shah, "Learning a deep model for human action recognition from novel viewpoints," *TPAMI*, vol. 40, no. 3, 2018.

[124] J. Stroud, D. Ross, C. Sun, J. Deng, and R. Sukthankar, "D3d: Distilled 3d networks for video action recognition," in *WACV*, 2020.

[125] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "Mars: Motion-augmented rgb stream for action recognition," in *CVPR*, 2019.

[126] Z. Shou, X. Lin, Y. Kalantidis, L. Sevilla-Lara, M. Rohrbach, S.-F. Chang, and Z. Yan, "Dmc-net: Generating discriminative motion cues for fast compressed video action recognition," in *CVPR*, 2019.

[127] H. Wang, D. Tran, L. Torresani, and M. Feiszli, "Video modeling with correlation networks," in *CVPR*, 2020.

[128] M. Fayyaz, E. Bahrami, A. Diba, M. Noroozi, E. Adeli, L. Van Gool, and J. Gall, "3d cnns with adaptive temporal feature resolutions," in *CVPR*, 2021.

[129] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *ICCV*, 2015.

[130] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *ICCV*, 2017.

[131] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *ICLR*, 2018.

[132] K. Li, X. Li, Y. Wang, J. Wang, and Y. Qiao, "Ct-net: Channel tensorization network for video classification," in *ICLR*, 2021.

[133] H. Yang, C. Yuan, B. Li, Y. Du, J. Xing, W. Hu, and S. J. Maybank, "Asymmetric 3d convolutional neural networks for action recognition," *Pattern Recognition*, vol. 85, 2019.

[134] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *ICCV*, 2019.

[135] S. Sudhakaran, S. Escalera, and O. Lanz, "Gate-shift networks for video action recognition," in *CVPR*, 2020.

[136] L. Wang, Z. Tong, B. Ji, and G. Wu, "Tdn: Temporal difference networks for efficient action recognition," in *CVPR*, 2021.

[137] Z. Wang, Q. She, and A. Smolic, "Action-net: Multipath excitation for action recognition," in *CVPR*, 2021.

[138] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *ECCV*, 2010.

[139] D. Li, Z. Qiu, Y. Pan, T. Yao, H. Li, and T. Mei, "Representing videos as discriminative sub-graphs for action recognition," in *CVPR*, 2021.

[140] J. Zhou, K.-Y. Lin, H. Li, and W.-S. Zheng, "Graph-based high-order relation modeling for long-term action recognition," in *CVPR*, 2021.

[141] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *CVPR*, 2019.

[142] S. Zhang, S. Guo, W. Huang, M. R. Scott, and L. Wang, "V4d: 4d convolutional neural networks for video-level representation learning," in *ICLR*, 2020.

[143] H. Duan, Y. Zhao, Y. Xiong, W. Liu, and D. Lin, "Omni-sourced webly-supervised learning for video recognition," in *ECCV*, 2020.

[144] D. Ghadiyaram, D. Tran, and D. Mahajan, "Large-scale weakly-supervised pre-training for video action recognition," in *CVPR*, 2019.

[145] G. Lev, G. Sadeh, B. Klein, and L. Wolf, "Rnn fisher vectors for action recognition and image annotation," in *ECCV*, 2016.

[146] Y. Pan, J. Xu, M. Wang, J. Ye, F. Wang, K. Bai, and Z. Xu, "Compressing recurrent neural networks with tensor ring for action recognition," in *AAAI*, vol. 33, 2019.

[147] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *CVPR*, 2020.

[148] M. E. Kalfaoglu, S. Kalkan, and A. A. Alatan, "Late temporal modeling in 3d cnn architectures with bert for action recognition," in *ECCV*, 2020.

[149] X. Yang, H. Fan, L. Torresani, L. S. Davis, and H. Wang, "Beyond short clips: End-to-end video-level learning with collaborative memories," in *CVPR*, 2021.

[150] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019.

[151] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding," *TPAMI*, 2020.

[152] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.

[153] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3d human action recognition," *TPAMI*, vol. 36, no. 5, 2013.

[154] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *CVPR*, 2014.

[155] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *CVPR*, 2015.

[156] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer lstm networks," in *WACV*, 2017.

[157] S. Zhang, Y. Yang, J. Xiao, X. Liu, Y. Yang, D. Xie, and Y. Zhuang, "Fusing geometric features for skeleton-based action recognition using multilayer lstm networks," *TMM*, vol. 20, no. 9, 2018.

[158] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention lstm networks," *TIP*, vol. 27, no. 4, 2018.

[159] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *ICCV*, 2015.

[160] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," in *AAAI*, 2016.

[161] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *CVPR*, 2016.

[162] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *ECCV*, 2016.

[163] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *CVPR*, 2017.

[164] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *CVPR*, 2017.

[165] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *AAAI*, 2017.

[166] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks," in *ICCV*, 2017.

[167] S. Li, W. Li, C. Cook, Y. Gao, and C. Zhu, "Deep independently recurrent neural network (indrnn)," *arXiv preprint arXiv:1910.06251*, 2019.

[168] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *CVPRW*, 2017.

[169] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *TCSVT*, vol. 28, no. 3, 2016.

[170] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *ACM Multimedia*, 2016.

[171] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 24, no. 5, 2017.

[172] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *CVPR*, 2017.

[173] C. Lea, M. Flynn, R. Vidal, A. Reiter, and G. Hager, "Temporal convolutional networks for action segmentation and detection," *CVPR*, 2017.

[174] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," *arXiv preprint arXiv:1804.06055*, 2018.

[175] C. Caetano, J. Sena, F. Brémond, J. A. Dos Santos, and W. R. Schwartz, "Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition," in *AVSS*, 2019.

[176] C. Caetano, F. Brémond, and W. R. Schwartz, "Skeleton image representation for 3d action recognition based on tree structure and reference joints," in *SIBGRAPI*, 2019.

[177] J. Liu, N. Akhtar, and A. Mian, "Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition.," in *CVPRW*, 2019.

[178] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, 2017.

[179] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *TPAMI*, vol. 41, no. 8, 2019.

[180] Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, "Skeletonnet: Mining deep part features for 3-d action recognition," *IEEE signal process. lett.*, vol. 24, no. 6, 2017.

[181] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, and M. He, "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn," in *ICMEW*, 2017.

[182] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *ACM Multimedia Asia*, 2019.

[183] L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, and H. Li, "T-gcn: A temporal graph convolutional network for traffic prediction," *IEEE trans. Intell. Transp. Syst.*, 2019.

[184] F. Monti, K. Otness, and M. M. Bronstein, "Motifnet: a motif-based graph convolutional network for directed graphs," in *DSW*, 2018.

[185] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *ECCV*, 2018.

[186] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *CVPR*, 2019.

[187] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *CVPR*, 2018.

[188] X. Zhang, C. Xu, X. Tian, and D. Tao, "Graph edge convolutional neural networks for skeleton-based action recognition," *TNNLS*, 2019.

[189] M. Korban and X. Li, "Ddgcn: A dynamic directed graph convolutional network for action recognition," in *ECCV*, 2020.

[190] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu, "Decoupling gcn with dropgraph module for skeleton-based action recognition," 2020.

[191] P. Yu, Y. Zhao, C. Li, J. Yuan, and C. Chen, "Structure-aware human-action generation," in *ECCV*, 2020.

[192] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019.

[193] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning graph convolutional network for skeleton-based human action recognition by neural searching," in *AAAI*, 2020.

[194] X. Zhang, C. Xu, and D. Tao, "Context aware graph convolution for skeleton-based action recognition," in *CVPR*, 2020.

[195] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019.

[196] C. Wu, X.-J. Wu, and J. Kittler, "Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition," in *ICCVW*, 2019.

[197] B. Li, X. Li, Z. Zhang, and F. Wu, "Spatio-temporal graph routing for skeleton-based action recognition," in *AAAI*, vol. 33, 2019.

[198] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *CVPR*, 2020.

[199] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based action recognition," in *CVPR*, 2020.

[200] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *CVPR*, 2019.

[201] Y.-H. Wen, L. Gao, H. Fu, F.-L. Zhang, and S. Xia, "Graph cnns with motif and variable temporal block for skeleton-based action recognition," in *AAAI*, vol. 33, 2019.

[202] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *CVPR*, 2020.

[203] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *arXiv preprint arXiv:2106.15125*, 2021.

[204] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction," *TPAMI*, 2021.

[205] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *ICCV*, 2017.

[206] C. Xie, C. Li, B. Zhang, C. Chen, J. Han, C. Zou, and J. Liu, "Memory attention networks for skeleton-based action recognition," *arXiv preprint arXiv:1804.08254*, 2018.

[207] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *ACPR*, 2015.

[208] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning clip representations for skeleton-based 3d action recognition," *TIP*, vol. 27, no. 6, 2018.

[209] Y. Xu, J. Cheng, L. Wang, H. Xia, F. Liu, and D. Tao, "Ensemble one-dimensional convolution neural networks for skeleton-based action recognition," *IEEE Signal Process. Lett.*, 2018.

[210] Y. Li, R. Xia, X. Liu, and Q. Huang, "Learning shape-motion representations from geometric algebra spatio-temporal model for skeleton-based action recognition," in *ICME*, 2019.

[211] A. Banerjee, P. K. Singh, and R. Sarkar, "Fuzzy integral based cnn classifier fusion for 3d skeleton action recognition," *TCSVT*, 2019.

[212] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang, "3d human action representation learning via cross-view consistency pursuit," in *CVPR*, 2021.

[213] H. Mohaghegh, N. Karimi, R. Soroushmehr, S. Samavi, and K. Najarian, "Aggregation of rich depth aware features in a modified stacked generalization model for single image depth estimation," *TCSVT*, vol. 29, 2019.

[214] M. Harville and D. Li, "Fast, integrated person tracking and activity recognition with plan-view templates from a single stereo camera," in *CVPR*, vol. 2, 2004.

[215] M.-C. Roh, H.-K. Shin, and S.-W. Lee, "View-independent human action recognition with volume motion template on single stereo camera," *Pattern Recognition Letters*, vol. 31, no. 7, 2010.

[216] P. Wang, W. Li, Z. Gao, C. Tang, and P. O. Ogunbona, "Depth pooling based large-scale 3-d action recognition with convolutional neural networks," *TMM*, vol. 20, no. 5, 2018.

[217] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *CVPR*, 2014.

[218] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *CVPR*, 2013.

[219] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *ACM Multimedia*, 2012.

[220] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Trans. Hum. Mach. Syst.*, 2015.

[221] A. Sanchez-Caballero, S. de López-Diz, D. Fuentes-Jimenez, C. Losada-Gutiérrez, M. Marrón-Romera, D. Casillas-Perez, and M. I. Sarker, "3dfcnn: Real-time action recognition using 3d deep neural networks with raw depth information," *arXiv preprint arXiv:2006.07743*, 2020.

[222] A. Sanchez-Caballero, D. Fuentes-Jimenez, and C. Losada-Gutiérrez, "Exploiting the convlstm: Human action recognition using raw depth video-based recurrent neural networks," *arXiv preprint arXiv:2006.07744*, 2020.

[223] Y. Zhu and S. Newsam, "Depth2action: Exploring embedded depth for large-scale action recognition," in *ECCV*, 2016.

[224] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *CVPR*, 2015.

[225] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *ICCV*, 2015.

[226] A. Akula, A. K. Shah, and R. Ghosh, "Deep learning approach for human action recognition in infrared images," *Cogn. Syst. Res.*, vol. 50, 2018.

[227] T. Kawashima, Y. Kawanishi, I. Ide, H. Murase, D. Deguchi, T. Aizawa, and M. Kawade, "Action recognition from extremely low-resolution thermal image sequence," in *AVSS*, 2017.

[228] A. K. Shah, R. Ghosh, and A. Akula, "A spatio-temporal deep learning approach for human action recognition in infrared videos," in *Optics and Photonics for Information Processing XII*, vol. 10751, 2018.

[229] H. Meglouli, L. Bentabet, M. Airouche, *et al.*, "A new technique based on 3d convolutional neural networks and filtering optical flow maps for action classification in infrared video," *Control Eng. Appl. Inf.*, 2019.

[230] Y. Liu, Z. Lu, J. Li, T. Yang, and C. Yao, "Global temporal representation based cnns for infrared action recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 6, 2018.

[231] D.-M. Tsai, W.-Y. Chiu, and M.-H. Lee, "Optical flow-motion history image (of-mhi) for action recognition.," *Signal Image Video Process.*, no. 8, 2015.

[232] Z. Jiang, Y. Wang, L. Davis, W. Andrews, and V. Rozgic, "Learning discriminative features via label consistent neural network," *WACV*, 2017.

[233] J. Imran and B. Raman, "Deep residual infrared action recognition by integrating local and global spatio-temporal cues," *Infrared Phys. Technol.*, vol. 102, 2019.

[234] J. Imran and B. Raman, "Evaluating fusion of rgb-d and inertial sensors for multimodal human action recognition," *JAIHC*, vol. 11, no. 1, 2020.

[235] V. Mehta, A. Dhall, S. Pal, and S. Khan, "Motion and region aware adversarial learning for fall detection with thermal imaging," *arXiv*, 2020.

[236] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*, 2012.

[237] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *CVPR*, 2014.

[238] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Histogram of oriented principal components for cross-view action recognition," *TPAMI*, vol. 38, no. 12, 2016.

[239] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *CVPRW*, 2010.

[240] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, *et al.*, "A low power, fully event-based gesture recognition system," in *CVPR*, 2017.

[241] X. Liu, M. Yan, and J. Bohg, "Meteornet: Deep learning on dynamic 3d point cloud sequences," in *ICCV*, 2019.

[242] Y. Min, Y. Zhang, X. Chai, and X. Chen, "An efficient pointlstm for point clouds based gesture recognition," in *CVPR*, 2020.

[243] G. Wang, H. Liu, M. Chen, Y. Yang, Z. Liu, and H. Wang, "Anchor-based spatial-temporal attention convolutional networks for dynamic 3d point cloud sequences," *arXiv preprint arXiv:2012.10860*, 2020.

[244] H. Wang, L. Yang, X. Rong, J. Feng, and Y. Tian, "Self-supervised 4d spatio-temporal feature learning via order prediction of sequential point cloud clips," *WACV*, 2021.

[245] H. Fan, Y. Yang, and M. Kankanhalli, "Point 4d transformer networks for spatio-temporal modeling in point cloud videos," in *CVPR*, 2021.

[246] H. Fan, X. Yu, Y. Ding, Y. Yang, and M. Kankanhalli, "Pstnet: Point spatio-temporal convolution on point cloud sequences," in *ICLR*, 2021.

[247] Q. Wang, Y. Zhang, J. Yuan, and Y. Lu, "Space-time event clouds for gesture recognition: from rgb cameras to event cameras," in *WACV*, 2019.

[248] C. Huang, "Event-based action recognition using timestamp image encoding network," *arXiv preprint arXiv:2009.13049*, 2020.

[249] J. Chen, J. Meng, X. Wang, and J. Yuan, "Dynamic graph cnn for event-camera based gesture recognition," in *ISCAS*, 2020.

[250] S. U. Innocenti, F. Becattini, F. Pernici, and A. Del Bimbo, "Temporal binary representation for event-based action recognition," in *ICPR*, 2021.

[251] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition," in *ECCV*, 2014.

[252] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *NeurIPS*, 2017.

[253] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *CVPR*, 2019.

[254] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 × 128120 db 15 μs latency asynchronous temporal contrast vision sensor," *IEEE journal of solid-state circuits*, vol. 43, no. 2, 2008.

[255] R. Berner, C. Brandli, M. Yang, S.-C. Liu, and T. Delbruck, "A 240× 180 10mw 12us latency sparse-output vision sensor for mobile applications," in *Symposium on VLSI Circuits*, 2013.

[256] S. A. Baby, B. Vinod, C. Chinni, and K. Mitra, "Dynamic vision sensors for human activity recognition," in *ACPR*, 2017.

[257] X. Clady, J.-M. Maro, S. Barré, and R. B. Benosman, "A motion-based feature for event-based pattern recognition," *Frontiers in neuroscience*, vol. 10, 2017.

[258] A. M. George, D. Banerjee, S. Dey, A. Mukherjee, and P. Balamurali, "A reservoir-based convolutional spiking neural network for gesture recognition from dvs input," in *IJCNN*, 2020.

[259] Y. Bi, A. Chadha, A. Abbas, E. Bourtsoulatze, and Y. Andreopoulos, "Graph-based spatiotemporal feature learning for neuromorphic vision sensing," *TIP*, vol. 29, 2020.

[260] G. Laput, K. Ahuja, M. Goel, and C. Harrison, "Ubicoustics: Plug-and-play acoustic activity recognition," in *UIST*, 2018.

[261] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "Cnn architectures for large-scale audio classification," in *ICASSP*, 2017.

[262] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *ECCV*, 2018.

[263] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani, "Listen to look: Action recognition by previewing audio," in *CVPR*, 2020.

[264] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer, "Audiovisual slowfast networks for video recognition," *arXiv preprint arXiv:2001.08740*, 2020.

[265] P. Casale, O. Pujol, and P. Radeva, "Human activity recognition from accelerometer data using a wearable device," in *ibPRIA*, 2011.

[266] A. Bayat, M. Pomplun, and D. A. Tran, "A study on human activity recognition using accelerometer data from smartphones," *Procedia Comput. Sci.*, vol. 34, 2014.

[267] A. M. Khan, Y.-K. Lee, S.-Y. Lee, and T.-S. Kim, "Human activity recognition via an accelerometer-enabled-smartphone using kernel discriminant analysis," in *FutureTech*, 2010.

[268] U. Maurer, A. Smailagic, D. P. Siewiorek, and M. Deisher, "Activity recognition and monitoring using multiple sensors on different body positions," in *BSN*, 2006.

[269] Y. Chen and Y. Xue, "A deep learning approach to human activity recognition based on single accelerometer," in *SMC*, 2015.

[270] M. Panwar, S. R. Dyuthi, K. C. Prakash, D. Biswas, A. Acharyya, K. Maharatna, A. Gautam, and G. R. Naik, "Cnn based approach for activity recognition using a wrist-worn accelerometer," in *EMBC*, 2017.

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

[272] J. Wang, Q. Long, K. Liu, Y. Xie, *et al.*, "Human action recognition on cellphone using compositional bidir-lstm-cnn networks," in *CNCI*, 2019.

[273] J. Lu and K.-Y. Tong, "Robust single accelerometer-based activity recognition using modified recurrence plot," *IEEE Sens. J.*, vol. 19, no. 15, 2019.

[274] J. Eckmann, S. O. Kamphorst, D. Ruelle, *et al.*, "Recurrence plots of dynamical systems," *World Scientific Series on Nonlinear Science Series A*, vol. 16, 1995.

[275] G. L. Santos, P. T. Endo, K. H. d. C. Monteiro, E. d. S. Rocha, I. Silva, and T. Lynn, "Accelerometer-based human fall detection using convolutional neural networks," *Sensors*, vol. 19, no. 7, 2019.

[276] A. Lin and H. Ling, "Doppler and direction-of-arrival (ddoa) radar for multiple-mover sensing," *IEEE trans. Aero. Elec. Sys.*, vol. 43, no. 4, 2007.

[277] A. Stove, "Modern fmcw radar-techniques and applications," in *EURAD*, 2004.

[278] J. Park, R. J. Javier, T. Moon, and Y. Kim, "Micro-doppler based classification of human aquatic activities via transfer learning of convolutional neural networks," *Sensors*, vol. 16, no. 12, 2016.

[279] R. Trommel, R. Harmanny, L. Cifola, and J. Driessen, "Multi-target human gait classification using deep convolutional neural networks on micro-doppler spectrograms," in *EuRAD*, 2016.

[280] Y. Kim, J. Park, and 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, 2017.

[281] H. Du, Y. He, and T. Jin, "Transfer learning for human activities classification using micro-doppler spectrograms," in *ICCEM*, 2018.

[282] S. A. Shah and F. Fioranelli, "Human activity recognition: preliminary results for dataset portability using fmcw radar," in *RADAR*, 2019.

[283] Y. Lin, J. Le Kernec, S. Yang, F. Fioranelli, O. Romain, and Z. Zhao, "Human activity classification with radar: Optimization and noise robustness with iterative convolutional neural networks followed with random forests," *IEEE Sens. J.*, vol. 18, no. 23, 2018.

[284] R. Hernangómez, A. Santra, and S. Stańczak, "Human activity classification with frequency modulated continuous wave radar using deep convolutional neural networks," in *RADAR*, 2019.

[285] H. Zhou, Y. He, H. Huang, and X. Jing, "Two-stream convolutional neural network for action recognition in radar," in *ICSINC*, 2020.

[286] M. Wang, Y. D. Zhang, and G. Cui, "Human motion recognition exploiting radar with stacked recurrent neural network," *Digit. Signal Process.*, vol. 87, 2019.

[287] S. Yang, J. Le Kernec, and F. Fioranelli, "Action recognition using indoor radar systems," *IET Human Motion Analysis for Healthcare Applications*, 2019.

[288] H. Zou, J. Yang, H. Prasanna Das, H. Liu, Y. Zhou, and C. J. Spanos, "Wifi and vision multimodal learning for accurate and robust device-free human activity recognition," in *CVPRW*, 2019.

[289] X. Wu, Z. Chu, P. Yang, C. Xiang, X. Zheng, and W. Huang, "Tw-see: Human activity recognition through the wall with commodity wi-fi devices," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 1, 2018.

[290] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, and H. Liu, "E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures," in *MobiCom*, 2014.

[291] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Device-free human activity recognition using commercial wifi devices," *J-SAC*, vol. 35, no. 5, 2017.

[292] J. Wang, X. Zhang, Q. Gao, H. Yue, and H. Wang, "Device-free wireless localization and activity recognition: A deep learning approach," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, 2017.

[293] S. Huang, D. Wang, R. Zhao, and Q. Zhang, "Wiga: A wifi-based contactless activity sequence recognition system based on deep learning," in *MSN*, 2019.

[294] B. Sheng, F. Xiao, L. Sha, and L. Sun, "Deep spatial–temporal model based cross-scene action recognition using commodity wifi," *IEEE Internet of Things Journal*, vol. 7, no. 4, 2020.

[295] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, "Wifi csi based passive human activity recognition using attention based blstm," *IEEE Trans. Mob. Comput.*, vol. 18, no. 11, 2019.

[296] Q. Gao, J. Wang, X. Ma, X. Feng, and H. Wang, "Csi-based device-free wireless localization and activity recognition using radio image features," *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, 2017.

[297] Q. Kong, Z. Wu, Z. Deng, M. Klinkigt, B. Tong, and T. Murakami, "Mmact: A large-scale dataset for cross modal human action understanding," in *ICCV*, 2019.

[298] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *ICIP*, 2015.

[299] J. Yang, J. Lee, and J. Choi, "Activity recognition based on rfid object usage for smart mobile devices," *JCST*, vol. 26, no. 2, 2011.

[300] T. Li, L. Fan, M. Zhao, Y. Liu, and D. Katabi, "Making the invisible visible: Action recognition through walls and occlusions," in *ICCV*, 2019.

[301] W. Xu, G. Lan, Q. Lin, S. Khalifa, N. Bergmann, M. Hassan, and W. Hu, "Keh-gait: Towards a mobile healthcare user authentication system by kinetic energy harvesting.," in *NDSS*, 2017.

[302] D. Ma, G. Lan, W. Xu, M. Hassan, and W. Hu, "Simultaneous energy harvesting and gait recognition using piezoelectric energy harvester," *arXiv preprint arXiv:2009.02752*, 2020.

[303] Y. Wang, B. Du, Y. Shen, K. Wu, G. Zhao, J. Sun, and H. Wen, "Ev-gait: Event-based robust gait recognition using dynamic vision sensors," in *CVPR*, 2019.

[304] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *TPAMI*, vol. 41, no. 2, 2018.

[305] J. Imran and P. Kumar, "Human action recognition using rgb-d sensor and deep convolutional neural networks," in *ICACCI*, 2016.

[306] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona, "Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks," in *CVPR*, 2017.

[307] L. Wang, Z. Ding, Z. Tao, Y. Liu, and Y. Fu, "Generative multi-view human action recognition," in *ICCV*, 2019.

[308] C. Dhiman and D. K. Vishwakarma, "View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics," *TIP*, vol. 29, 2020.

[309] H. Wang, Z. Song, W. Li, and P. Wang, "A hybrid network for large-scale action recognition from rgb and depth modalities," *Sensors*, vol. 20, no. 11, 2020.

[310] Z. Ren, Q. Zhang, X. Gao, P. Hao, and J. Cheng, "Multi-modality learning for human action recognition," *Multimedia Tools and Applications*, vol. 80, no. 11, 2021.

[311] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *ICCV*, 2017.

[312] F. Baradel, C. Wolf, and J. Mille, "Human action recognition: Pose-based attention draws focus to hands," in *ICCVW*, 2017.

[313] R. Zhao, H. Ali, and P. Van der Smagt, "Two-stream rnn/cnn for action recognition in 3d videos," in *IROS*, 2017.

[314] F. Baradel, C. Wolf, and J. Mille, "Pose-conditioned spatio-temporal attention for human action recognition," *arXiv preprint arXiv:1703.10106*, 2017.

[315] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "Skeleton-indexed deep multi-modal feature learning for high performance human action recognition," in *ICME*, 2018.

[316] S. Das, R. Dai, M. Koperski, L. Minciullo, L. Garattoni, F. Bremond, and G. Francesca, "Toyota smarthome: Real-world activities of daily living," in *ICCV*, 2019.

[317] J. Li, X. Xie, Q. Pan, Y. Cao, Z. Zhao, and G. Shi, "Sgm-net: Skeleton-guided multimodal network for action recognition," *Pattern Recognition*, vol. 104, 2020.

[318] S. Das, S. Sharma, R. Dai, F. Bremond, and M. Thonnat, "Vpn: Learning video-pose embedding for activities of daily living," in *ECCV*, 2020.

[319] H. Duan, Y. Zhao, K. Chen, D. Shao, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," *arXiv preprint arXiv:2104.13586*, 2021.

[320] J. Cai, N. Jiang, X. Han, K. Jia, and J. Lu, "Jolo-gcn: mining joint-centered light-weight information for skeleton-based action recognition," in *WACV*, 2021.

[321] H. Rahmani and M. Bennamoun, "Learning action recognition model from depth and skeleton videos," in *ICCV*, 2017.

[322] A. Kamel, B. Sheng, P. Yang, P. Li, R. Shen, and D. D. Feng, "Deep convolutional neural networks for human action recognition using depth maps and postures," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 49, no. 9, 2018.

[323] C. Zhao, M. Chen, J. Zhao, Q. Wang, and Y. Shen, "3d behavior recognition based on multi-modal deep space-time learning," *Applied Sciences*, vol. 9, no. 4, 2019.

[324] S. S. Rani, G. A. Naidu, and V. U. Shree, "Kinematic joint descriptor and depth motion descriptor with convolutional neural networks for human action recognition," *MATPR*, vol. 37, 2021.

[325] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang, "Deep multimodal feature analysis for action recognition in rgb+ d videos," *TPAMI*, vol. 40, no. 5, 2017.

[326] J.-F. Hu, W.-S. Zheng, J. Pan, J. Lai, and J. Zhang, "Deep bilinear learning for rgb-d action recognition," in *ECCV*, 2018.

[327] P. Khaire, P. Kumar, and J. Imran, "Combining cnn streams of rgb-d and skeletal data for human activity recognition," *Pattern Recognition Letters*, vol. 115, 2018.

[328] P. Khaire, J. Imran, and P. Kumar, "Human activity recognition by fusion of rgb, depth, and skeletal data," in *ICCVIP*, 2018.

[329] J. Ye, K. Li, G.-J. Qi, and K. A. Hua, "Temporal order-preserving dynamic quantization for human action recognition from multimodal sensor streams," in *ACM Multimedia*, 2015.

[330] S. Ardianto and H.-M. Hang, "Multi-view and multi-modal action recognition with learned fusion," in *APSIPA ASC*, 2018.

[331] A. Main de Boissiere and R. Noumeir, "Infrared and 3d skeleton feature fusion for rgb-d action recognition," *arXiv*, 2020.

[332] C. Wang, H. Yang, and C. Meinel, "Exploring multimodal video representation for action recognition," in *IJCNN*, 2016.

[333] M. Brousmiche, J. Rouat, and S. Dupont, "Multi-level attention fusion network for audio-visual event recognition," *arXiv preprint arXiv:2106.06736*, 2021.

[334] N. Dawar, S. Ostadabbas, and N. Kehtarnavaz, "Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition," *IEEE Sensors Letters*, 2018.

[335] N. Dawar and N. Kehtarnavaz, "A convolutional neural network-based sensor fusion system for monitoring transition movements in healthcare applications," in *ICCA*, 2018.

[336] H. Wei, R. Jafari, and N. Kehtarnavaz, "Fusion of video and inertial sensing for deep learning–based human action recognition," *Sensors*, vol. 19, no. 17, 2019.

[337] Z. Ahmad and N. Khan, "Human action recognition using deep multilevel multimodal (M^2) fusion of depth and inertial sensors," *IEEE Sens. J.*, vol. 20, no. 3, 2019.

[338] Z. Ahmad and N. Khan, "Cnn based multistage gated average fusion (mgaf) for human action recognition using depth and inertial sensors," *IEEE Sens. J.*, 2020.

[339] C. Pham, L. Nguyen, A. Nguyen, N. Nguyen, and V.-T. Nguyen, "Combining skeleton and accelerometer data for human fine-grained activity recognition and abnormal behaviour detection with deep temporal convolutional networks," *Multimedia Tools and Applications*, 2021.

[340] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *ACM Multimedia*, 2015.

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

[342] Q. Zou, Y. Wang, Q. Wang, Y. Zhao, and Q. Li, "Deep learning-based gait recognition using smartphones in the wild," *IEEE Trans. Inf. Forens. Sec.*, vol. 15, 2020.

[343] R. Memmesheimer, N. Theisen, and D. Paulus, "Gimme signals: Discriminative signal encoding for multimodal activity recognition," *arXiv preprint arXiv:2003.06156*, 2020.

[344] Y. Kong and Y. Fu, "Bilinear heterogeneous information machine for rgb-d action recognition," in *CVPR*, 2015.

[345] Y. Kong and Y. Fu, "Max-margin heterogeneous information machine for rgb-d action recognition," *IJCV*, vol. 123, no. 3, 2017.

[346] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *TPAMI*, vol. 23, no. 3, 2001.

[347] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *NeurIPS*, 2015.

[348] J. Liu, A. Shahroudy, G. Wang, L.-Y. Duan, and A. C. Kot, "Skeleton-based online action prediction using scale selection network," *TPAMI*, vol. 42, no. 6, 2019.

[349] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "Real time action recognition using histograms of depth gradients and random decision forests," in *WACV*, 2014.

[350] A. Shahroudy, T.-T. Ng, Q. Yang, and G. Wang, "Multimodal multipart learning for action recognition in depth videos," *TPAMI*, vol. 38, no. 10, 2015.

[351] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," in *CVPR*, 2015.

[352] E. Triantaphyllou, B. Shu, S. N. Sanchez, and T. Ray, "Multi-criteria decision making: an operations research approach," *Encyclopedia of electrical and electronics engineering*, 1998.

[353] Q. Zou, L. Ni, Q. Wang, Q. Li, and S. Wang, "Robust gait recognition by integrating inertial and rgbd sensors," *IEEE Trans. Cybern.*, vol. 48, no. 4, 2017.

[354] N. E. D. Elmadany, Y. He, and L. Guan, "Multimodal learning for human action recognition via bimodal/multimodal hybrid centroid canonical correlation analysis," *TMM*, vol. 21, no. 5, 2018.

[355] N. C. Garcia, P. Morerio, and V. Murino, "Modality distillation with multiple stream networks for action recognition," in *ECCV*, 2018.

[356] N. C. Garcia, S. A. Bargal, V. Ablavsky, P. Morerio, V. Murino, and S. Sclaroff, "Dmcl: Distillation multiple choice learning for multimodal action recognition," *arXiv preprint arXiv:1912.10982*, 2019.

[357] N. C. Garcia, P. Morerio, and V. Murino, "Learning with privileged information via adversarial discriminative modality distillation," *TPAMI*, 2019.

[358] F. M. Thoker and J. Gall, "Cross-modal knowledge distillation for action recognition," in *ICIP*, 2019.

[359] L. Wang, C. Gao, L. Yang, Y. Zhao, W. Zuo, and D. Meng, "Pm-gans: Discriminative representation learning for action recognition using partial-modalities," in *ECCV*, 2018.

[360] A. Chadha, Y. Bi, A. Abbas, and Y. Andreopoulos, "Neuromorphic vision sensing for cnn-based action recognition," in *ICASSP*, 2019.

[361] Y. Bi and Y. Andreopoulos, "Pix2nvs: Parameterized conversion of pixel-domain video frames to neuromorphic vision streams," in *ICIP*, 2017.

[362] Y. Liu, G. Li, and L. Lin, "Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition," *arXiv preprint arXiv:2009.00210*, 2020.

[363] A. Perez, V. Sanguineti, P. Morerio, and V. Murino, "Audio-visual model distillation using acoustic images," in *WACV*, 2019.

[364] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

[365] H. Alwassel, D. Mahajan, L. Torresani, B. Ghanem, and D. Tran, "Self-supervised learning by cross-modal audio-video clustering," *arXiv preprint arXiv:1911.12667*, 2019.

[366] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *ICCV*, 2017.

[367] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," in *NeurIPS*, 2018.

[368] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *ICCV*, 2019.

[369] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *CVPR*, 2020.

[370] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *ICCV*, 2019.

[371] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[372] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[373] 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," *arXiv preprint arXiv:1705.06950*, 2017.

[374] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *arXiv preprint arXiv:1808.01340*, 2018.

[375] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019.

[376] D. Damen, H. Doughty, G. Maria Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *ECCV*, 2018.

[377] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes." http://www.thumos.info/, 2015.

[378] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *CVPR*, 2015.

[379] R. Goyal, S. E. 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 *ICCV*, vol. 1, 2017.

[380] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *ICPR*, vol. 3, 2004.

[381] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *CVIU*, vol. 104, no. 2-3, 2006.

[382] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation mocap database hdm05," 2007.

[383] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008.

[384] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," *CVPR*, 2009.

[385] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *ECCV*, 2010.

[386] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from rgbd images," in *AAAI*, 2011.

[387] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian, "Human daily action analysis with multi-view and color-depth data," in *ECCV*, 2012.

[388] Y.-C. Lin, M.-C. Hu, W.-H. Cheng, Y.-H. Hsieh, and H.-M. Chen, "Human action recognition and retrieval using sole depth information," in *ACM MM*, 2012.

[389] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *CVPRW*, 2012.

[390] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *Int. J. Rob. Res.*, vol. 32, no. 8, 2013.

[391] M. Munaro, G. Ballin, S. Michieletto, and E. Menegatti, "3d flow estimation for human action recognition from colored point clouds," *BICA*, vol. 5, 2013.

[392] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *ICCV*, 2013.

[393] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. LaViola, and R. Sukthankar, "Exploring the trade-off between accuracy and observational latency in action recognition," *IJCV*, vol. 101, no. 3, 2013.

[394] A.-A. Liu, Y.-T. Su, P.-P. Jia, Z. Gao, T. Hao, and Z.-X. Yang, "Multiple/single-view human action recognition via part-induced multitask structural learning," *IEEE Trans. Cybern.*, 2014.

[395] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, "Pose-based human action recognition via sparse representation in dissimilarity space," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, 2014.

[396] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Action classification with locality-constrained linear coding," in *ICPR*, 2014.

[397] A.-A. Liu, W.-Z. Nie, Y.-T. Su, L. Ma, T. Hao, and Z.-X. Yang, "Coupled hidden conditional random fields for rgb-d human action recognition," *Signal Processing*, vol. 112, 2015.

[398] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *ECCV*, 2016.

[399] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[400] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *CVPR*, 2018.

[401] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," *TPAMI*, vol. 40, no. 2, 2017.

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

[403] Y. Ji, F. Xu, Y. Yang, F. Shen, H. T. Shen, and W.-S. Zheng, "A large-scale rgb-d database for arbitrary-view human action recognition," in *ACM Multimedia*, 2018.

[404] M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelhagen, "Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles," in *ICCV*, 2019.

[405] M. Moreaux, M. G. Ortiz, I. Ferrané, and F. Lerasle, "Benchmark for kitchen20, a daily life dataset for audio-based human action recognition," in *CBMI*, 2019.

[406] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick, *et al.*, "Moments in time dataset: one million videos for event understanding," *TPAMI*, vol. 42, no. 2, 2019.

[407] F. Wang, J. Feng, Y. Zhao, X. Zhang, S. Zhang, and J. Han, "Joint activity recognition and indoor localization with wifi fingerprints," *IEEE Access*, vol. 7, 2019.

[408] Y. Ben-Shabat, X. Yu, F. Saleh, D. Campbell, C. Rodriguez-Opazo, H. Li, and S. Gould, "The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose," in *WACV*, 2020.

[409] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, "Rescaling egocentric vision," *arXiv preprint arXiv:2006.13256*, 2020.

[410] J. Jang, D. Kim, C. Park, M. Jang, J. Lee, and J. Kim, "Etri-activity3d: A large-scale rgb-d dataset for robots to recognize daily activities of the elderly," *arXiv*, 2020.

[411] A. Miech, J.-B. Alayrac, I. Laptev, J. Sivic, and A. Zisserman, "Rareact: A video dataset of unusual interactions," *arXiv preprint arXiv:2008.01018*, 2020.

[412] A. R. Punnakkal, A. Chandrasekaran, N. Athanasiou, A. Quiros-Ramirez, and M. J. Black, "Babel: Bodies, action and behavior with english labels," in *CVPR*, 2021.

[413] N. Rai, H. Chen, J. Ji, R. Desai, K. Kozuka, S. Ishizaka, E. Adeli, and J. C. Niebles, "Home action genome: Cooperative compositional action understanding," in *CVPR*, 2021.

[414] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles," in *CVPR*, 2021.

[415] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?," in *CVPR*, 2020.

[416] T. Li, J. Liu, W. Zhang, and L. Duan, "Hard-net: hardness-aware discrimination network for 3d early activity prediction," in *ECCV*, 2020.

[417] J. Liu, A. Shahroudy, G. Wang, L.-Y. Duan, and A. C. Kot, "Ssnet: scale selection network for online 3d action prediction," in *CVPR*, 2018.

[418] B. Fernando and S. Herath, "Anticipating human actions by correlating past with the future with jaccard similarity measures," in *CVPR*, 2021.

[419] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *CSUR*, vol. 53, no. 3, 2020.

[420] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien, "Unsupervised learning from narrated instruction videos," in *CVPR*, 2016.

[421] A. Singh, O. Chakraborty, A. Varshney, R. Panda, R. Feris, K. Saenko, and A. Das, "Semi-supervised action recognition with temporal contrastive learning," in *CVPR*, 2021.

[422] X. Song, S. Zhao, J. Yang, H. Yue, P. Xu, R. Hu, and H. Chai, "Spatio-temporal contrastive domain adaptation for action recognition," in *CVPR*, 2021.