IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 29 April 2023, accepted 28 May 2023, date of publication 2 June 2023, date of current version 7 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3282311

SURVEY

A Comprehensive Survey of RGB-Based and Skeleton-Based Human Action Recognition

CAILING WANG AND JINGJING YAN
School of Computer Science, Xi'an Shiyou University, Xi'an 710065, China
Corresponding author: Cailing Wang (azering@163.com)

ABSTRACT With the advancement of computer vision, human action recognition (HAR) has shown its broad research worth and application prospects in a wide range of fields such as intelligent security, automatic driving and human-machine interaction. Based on the type of data captured by cameras and sensors, e.g., RGB, depth, skeleton, and infrared data, HAR methods can be classified into RGB-based and skeleton-based. RGB data is easy and inexpensive to obtain, but RGB-based methods need to cope with a large amount of irrelevant background information and are easily affected by factors such as lighting and shooting angle. The skeleton-based methods eliminate the impact of background variables and require little computational work due to their skeleton-focused features, but they lack the context data necessary for HAR. This paper gives a thorough survey of these two approaches, covering deep learning methods, handcrafted feature extraction methods, common datasets, challenges, and future research directions. The skeleton-based action recognition methods section specifically presents the most well-liked 2D and 3D pose estimation algorithms. This survey aims to give researchers new to the area or engaged in a long-term study a selection of datasets and algorithms, as well as an overview of the present issues and expected future directions in the field.

INDEX TERMS Action dataset, deep learning, pose estimation, RGB-based action recognition, skeleton-based action recognition, systematic survey.

I. INTRODUCTION

Human action recognition (HAR) aims to develop an automated system that mimics the human visual system to understand and describe human actions in a given scene. HAR refers to detecting static features in the same frame and dynamic features between several adjacent frames from time sequences (video frames, human skeleton sequences, etc.) containing the complete action execution and classifying human actions, as shown in Fig. 1 (a for applying eye makeup and b for pull-ups). With the increasing demands on and dependence on machine intelligence, the application of HAR technology is becoming more widespread and has high commercial value in the fields of intelligent security [1], [2], virtual reality [3], [4], [5], human-computer interaction [6], [7],etc.

The data for HAR is now more diverse than it was in the past, including data from new modalities including depth, skeleton, and infrared, thanks to ongoing research on wearable sensors and depth cameras. RGB data contains rich texture and context information yet includes a complex background environment, while the new modality data is more robust to noise than RGB data. Depending on the type of input data, popular research methods for HAR include the RGB-based method and the skeleton-based method, both of which are hot directions in the field.

The initial research approach focused on feature extraction from RGB static images, which recognizes human actions from a single image without considering temporal information. Guo et al. [8] surveyed HAR based on static RGB images, discussing different methods of machine learning and deep learning for low-level feature extraction and high-level action representation. Vrigkas et al. [9] similarly surveyed HAR based on RGB static image representation, detailing both unimodal and multimodal types of approaches. In terms of feature representation, Vishwakarma et al. [10] summarized the classical HAR methods, dividing them into hierarchical and non-hierarchical methods. Survey [11] shows a

The associate editor coordinating the review of this manuscript and approving it for publication was Amjad Ali.

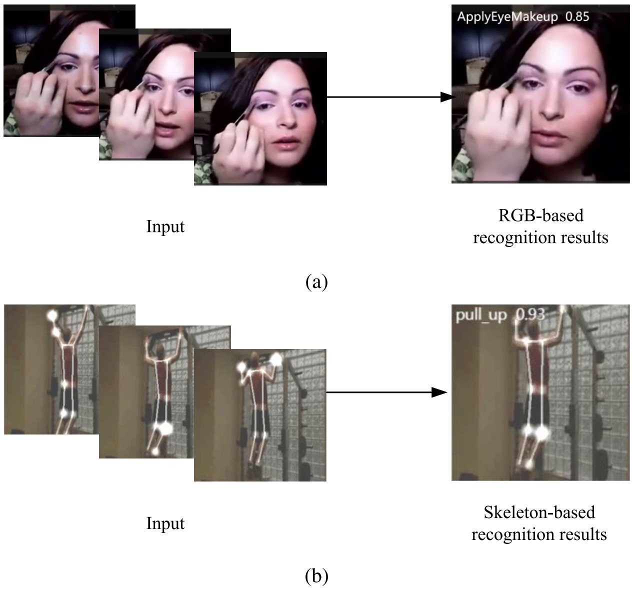

FIGURE 1. (a) shows an example of RGB-based HAR and (b) shows an example of skeleton-based HAR.

comprehensive overview of the handcrafted methods used in HAR. In addition, some surveys [12], [13], [14], [15] discuss the merits and demerits of handcrafted and deep learning in detail and highlight the benefits of deep learning-based methods. More recently, Saleem et al. [16] compared and analyzed various studies based on predefined parameter analysis of 46 state-of-the-art methods proposed since 2011, providing an update on recent trends of HAR research and emphasizing open challenges for future research. However, these surveys do not provide a comprehensive understanding of methods for HAR research based on other data modalities.

In recent years, the advantages of combining skeleton data with deep learning have been gradually demonstrated. Many researchers have gradually focused on the study of skeleton-based HAR, successively proposing many impressive methods, especially GCN-based methods. Xing et al. [17] described the development of HAR based on 3D skeleton data, meanwhile reviewing the existing variants of three mainstream techniques based on deep learning and comparing their performance in three dimensions. The survey [18], [19] not only detailed graph convolutional network structures and data modalities for HAR but also focused on the application of GCNs in HAR. Gupta et al. [20] investigated the current and future frontiers of skeleton-based HAR and introduced a large-scale action dataset, named skeleton-152, which opens up a new field. As human pose is also crucial for HAR, Song et al. [21] review the research progress on human pose estimation and its application in HAR. In addition, [22] focused on data fusion and recognition techniques in a visual context from an RGB-D perspective. [23], [24] reviewed popular approaches using vision and inertial sensors for HAR. However, these surveys lack comparative studies with RGB-based methods and a macroscopic and comprehensive presentation.

Therefore, we perform a comprehensive survey of the two popular methods mentioned above, which are RGB-based and skeleton-based HAR methods. The specifics include four parts: feature representation methods, common datasets, challenges, and prospects. The extraction of significantly distinguishable action features from video data is a crucial step in HAR. Our study details both handcrafted features and deep learning-based feature extraction approaches for RGB and skeleton data, and it discusses the advantages and disadvantages of the milestone algorithms. Our investigation includes a comprehensive public dataset on RGB and skeleton data for common datasets and their importance as algorithms. While many excellent and efficient algorithms have been proposed in succession, factors such as the surrounding environment and the limitations of hardware devices still pose many challenges in this field. This survey also analyzes the challenges of both RGB-based and skeleton-based approaches separately. We also discuss the future direction of the field. Considering that the acquisition of the skeleton data relies on sensors and pose estimation algorithms, the current popular 2D and 3D pose estimation algorithms are presented before discussing the skeleton-based feature representation methods.

The four key contributions are as follows.

1) For RGB and skeleton data, we give a thorough survey of handcrafted features and deep learning-based feature extraction approaches (as shown in Fig. 4), and we discuss the benefits and drawbacks of conventional approaches.
2) We present and compare the current public available common datasets for HAR, including details of the RGB dataset and the skeleton dataset.
3) In the context of skeleton-based HAR, this paper provides a comprehensive review of recent 2D and 3D deep human pose estimation models and their applications in the field of HAR.
4) We address the challenges and open issues facing the field based on the two approaches, respectively, and prospect for future directions to promote HAR.

The rest of this paper is organized as follows: Section II reviews RGB-based approaches, from shallow features to deep architectures. Section III collates the recently popular 2D and 3D deep human pose estimation models and discusses the skeleton-based approach from handcrafted features to deep learning. Section IV presents a comprehensive dataset of both RGB and skeleton data modalities. Section V analyzes the current challenges in the field for each of the two approaches. Section VI prospects the future research directions of HAR. Finally, Section VI concludes the survey. The detailed framework of this paper is shown in Fig. 2.

II. RGB-BASED ACTION RECOGNITION METHOD

Early studies were conducted based on RGB data. Initially, feature extraction relied on manual annotation [25], [26], [27], [28], which tended to rely on more a priori knowledge. Then, deep architectures were gradually adopted to extract

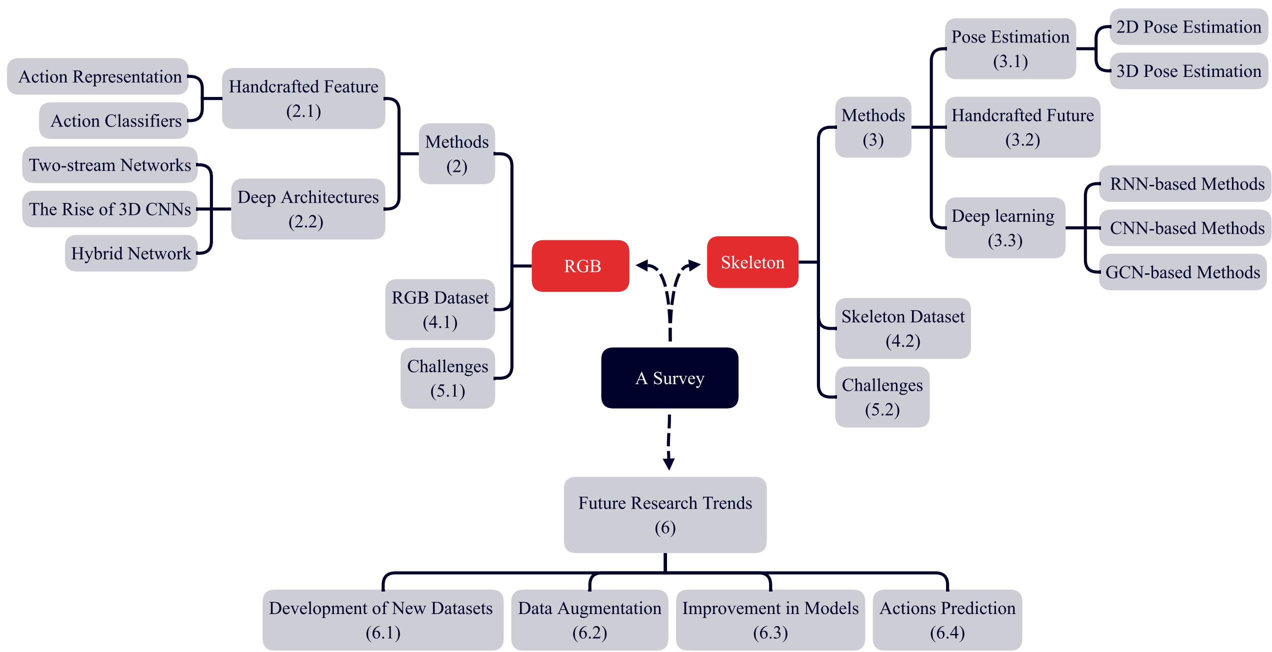

FIGURE 2. The framework of this paper.

features, with remarkable results. The following is a methodological review of RGB-based handcrafted features and deep architectures respectively.

A. RGB-BASED HANDCRAFTED FEATURE METHOD

Action representation and action classification are often the two key steps of handcrafted feature-based HAR methods [29], [30], [31]. In the action representation step, RGB data is transformed into a feature vector [32], [33], [34] or a set of feature vectors [35], [36], [37], and the vectors are then fed to classifiers [38], [39], [40] to get the results in the action classification step.

1) ACTION REPRESENTATION

The extraction of representative and distinct information about human actions is essential for feature representation since it significantly improves recognition precision. There are two types of action representation methods: holistic representation and local representation.

- *Holistic representation:*
  Holistic representation captures the motion information of the whole human subject. Bobick et al. [41] proposed motion energy image (MEI) and motion history image (MHI) to encode dynamic human motion into a single image based on the holistic representation, as shown in Fig. 3. It is sensitive to noise from the background. However, it inevitably introduces irrelevant background information noise besides the foreground for the information capture region, which is a fixed rectangle.
- *Local representation:*
  Local representation identifies just local regions with significant motion information, overcoming the problems of holistic representation. For example, spatio-temporal interest points [32], [34], [42], motion trajectories [31], [43] and other methods are robust to background information noise, camera motion, appearance changes, etc.

2) ACTION CLASSIFIERS

The action classifiers are employed to generate results followed by the action representation. The classification methods, classifiers and their descriptions are shown in Table 1.

B. RGB-BASED DEEP ARCHITECTURES METHODS

While holistic and local features yielded significant results, these handcrafted features require a large amount of prior knowledge to predefine the parameters. Moreover, for sizable datasets, they usually do not generalize well.

Deep neural networks [65], [66], [67] have recently been used with remarkable success in HAR to process large datasets. Convolutional neural networks (CNNs) [68] were initially applied to feature extraction and classification in 2D only. For spatio-temporal feature extraction, researchers have proposed different ideas, which are broadly classified into three genre branches, namely, two-stream network-based, 3D convolutional network-based, and hybrid network-based approaches.

1) TWO-STREAM NETWORKS

The motion of an object or scene can be effectively represented by optical flow [71]. Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBH), which can support optical flow, are examples of traditional handcrafted features that also include optical flow-like features.

TABLE 1. Action classification methods, classifiers and their descriptions.

Classification Methods	Classifiers	Description
Direct Classification	Support Vector Machines(SVM) [29], [44], [45]	These methods input the feature vector directly to the existing classifier for recognition.
	K-Nearest Neighbor (K-NN) [46]–[48]	
Sequential Methods	Conditional Random Fields (CRF) [37], [49], [50]	These methods employ sequential state models for classification.
	Hidden Markov Models (HMM) [51]–[53]	
	Structured Support Vector Machines(SSVM) [40], [54], [55]	
Spatio-temporal Methods	Global Gaussian Mixture Models (GMM) [56]	These methods take into account spatio-temporal correlations between local variables and possible details regarding the holistic spatiotemporal distribution of points of interest.
	Directional Pyramidal Co-occurrence Matrices (DPCM) [57]	
	Context-dependent Graph Kernels (CGK) [58]	
Part-based Methods	Constellation Model [59]	The geometric connections between body components are automatically represented by these methods, which take into account motion data from the complete human body as well as specific parts of the body.
Manifold Learning Methods	Kernel Principal Component Analysis(KPCA) [50]	These methods decrease the contour representation's dimensionality and embed it on a low-dimensional, nonlinear dynamic shape manifold, which is then further decreased via kernel principal component analysis.
Mid-Level Feature Methods	Hierarchical Methods [38], [60], [61]	These methods can learn additional representation layers and abstract low-level features for classification.
Feature Fusion Methods	Maximum Margin Distance Learning(MMDL) [62]	These methods all combine various types of characteristics to improve recognition.
	Multi-task Sparse Learning Model(MTSLM) [63]	
	Multi-feature Max-margin Hierarchical Bayesian Model(M3HBM) [64]	

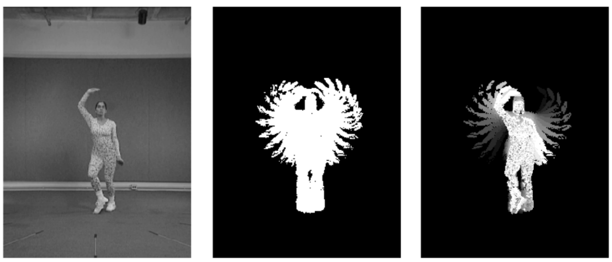

FIGURE 3. Examples in [41] of the input video frame and the comparison of MEI and MHI.

In light of this, Simonyan et al. [69] presented a two-stream network (as shown in Fig. 5) that combines spatial and temporal streams. The spatial stream takes the original video frames as input to capture the visual appearance information. The temporal stream takes the optical flow image information as an input to capture the motion information between video frames. Since the network uses a relatively shallow network architecture [72], Wang et al. [73] introduced cross-modal initialization, batch normalization, and multiscale cropping to prevent overfitting of the network at deeper levels, enabling the network to be trained using VGG16 [74] and to be far superior to [69] on UCF101.

The performance of classification is significantly impacted by feature fusion methods. Late fusion [69], [73], which weighted averages the prediction scores of the two streams, is the easiest and most straightforward method. Feichtenhofer et al. [75] also looked into where and how to fuse the network, and they made the case that fusing interactions early in the model learning process results in richer features and better performance. Feichtenhofer [76] extended ResNet [77] to the spatio-temporal domain by introducing a residual connection between two streams. Based on [76], Feichtenhofer et al. [65] further proposed a multiplicative gating function for the residual network to learn better spatio-temporal features. Wang et al. [78] performed hierarchical early fusion between two streams using a spatio-temporal pyramid. Feichtenhofer et al. also suggested SlowFastNet [70], shown in Fig. 6. The network replicates the characteristics of human visual cells, where slow paths can concentrate more on spatial and semantic information and fast paths can maintain temporal fidelity, while adopting lateral connections to fuse the features extracted by each path. The Fast path's low computational effort and high channel capacity greatly increase the overall effectiveness of SlowFast.

2) THE RISE OF 3D CNNs

Two-stream approaches always divide spatial and temporal information, which makes them unsuitable for real-time deployment. Afterward, other researchers put forth 3D convolutional methods that directly extract information in the three dimensions.

Ji et al. [79] first use a 3D CNN for HAR, which consists of five hardwired kernels that perform 3D convolution on adjacent frames to extract features from the spatial and

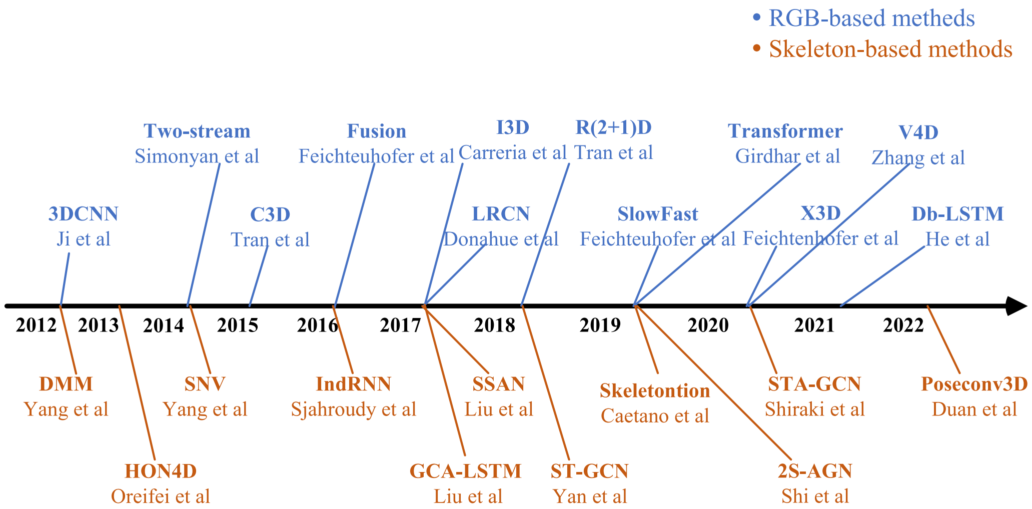

FIGURE 4. Milestone method for HAR. The blue font is the RGB-based milestone algorithm. The red font is the skeleton-based milestone algorithm.

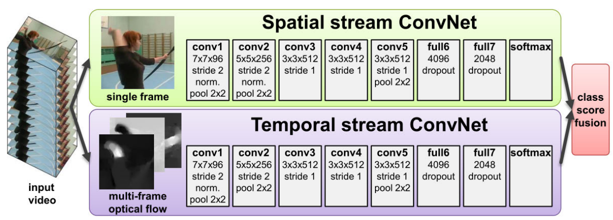

FIGURE 5. Two-stream architecture for video classification in [69].

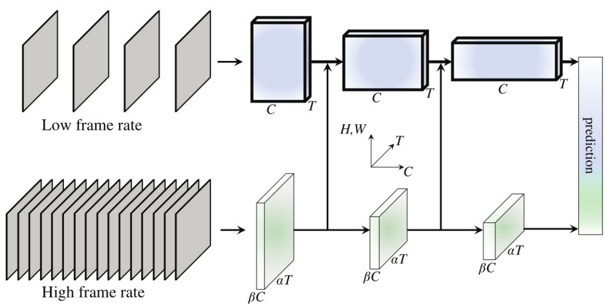

FIGURE 6. The SlowFast network in [70], which has a Slow pathway, a Fast pathway, and Lateral connections.

temporal dimensions. Tran et al. [80] proposed C3D based on an extension of 3DCNN [79]. The network can be seen as a 3D version of the VGG16 [74] network and shows strong generalization ability. However, to better train C3D networks, large-scale datasets with different contents and classes are often required. To improve the generalization capability even further, Carreira et al. [81] proposed I3D, which inflates the network into a spatio-temporal feature extractor along the temporal dimension. It adapts well-established image classification architectures for use in 3D CNNs and inflates the 2D model weights pre-trained by ImageNet to the corresponding weights in the 3D model.

P3D [82] and R(2+1)D [83] employ the concept of factorization to simplify 3D network training by combining a

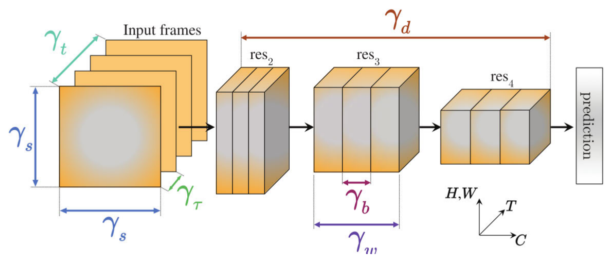

FIGURE 7. The framework of X3D networks.

2D spatial convolution (1×3) and a 1D temporal convolution (3×1×1) in place of the conventional 3D convolution (3×3). To better process motion, the trajectory convolution [84] employs deformable convolution for the temporal component. Combining 2D and 3D convolutions in a single neural network to produce richer and more illuminating feature maps is another method for simplifying 3D CNNs, such as MiCTNet [85], ARTNet [86], S3D [87].

To improve the efficiency of 3DCNN, CSN [88] demonstrated that it is a good idea to discompose 3D convolution by isolating channel interactions from spatio-temporal interactions in order to get cutting-edge performance. It can accelerate two to three times faster than the previous best method. Feichtenhofer et al. proposed the X3D algorithm [89], whose structure is shown in Fig. 7. The X3D network is not only expanded in temporal and spatial dimensions, but also improved in spatial resolution, input resolution, and channel dimension. X3D pushes 3D model decomposition to the extreme, which can meet different target complexity requirements. Yang et al. [90] considered that some morphologically similar actions such as walking, jogging, and running need to rely on visual speed-assisted discrimination, and proposed a Temporal Pyramid Network (TPN) similarity to X3D. With this model, the network can extract features at different rates, reducing the computational effort while improving efficiency.

Wang et al. [91] suggested a temporal segment network (TSN) in response to the network's inability to capture

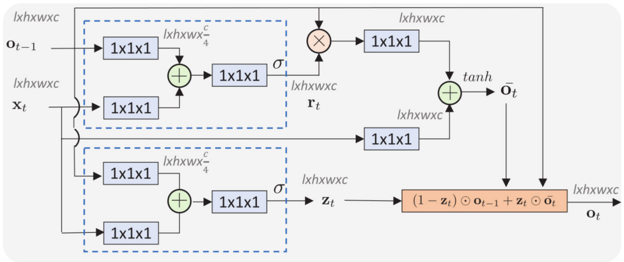

FIGURE 8. The FAST-GRU architecture.

long-time information and resulting feature loss. By utilizing a sparse sampling strategy, the TSN is able to create long-term dependencies while lowering the cost of training. The temporal relationship network [92] is also capable of learning and analyzing the temporal relationships between video frames on various time scales. Later, a new building block known as the non-local block was developed by Wang et al [93]. Like self-attention [94], non-local is a plug-and-play technique. A 4D CNN with 4D convolution was recently offered by V4D [95] to model the evolution of distant spatio-temporal representations.

In general, 3DCNNs create the relationship between temporal and spatial features in different ways, rather than replacing two-stream networks or being mutually exclusive.

3) HYBRID NETWORK

Adding more recurrent layers to CNN to create hybrid networks [96], [97], like LSTM and RNN, is another well-liked method for HAR. This hybrid network exhibits outstanding superiority in extracting spatial dimensional features and long-term feature dependence because it incorporates the benefits of both CNN and LSTM [67], [98], [99].

Donahue et al. investigated LSTM and proposed LRCN [96] for modeling CNN-generated spatial features over temporal sequences. Ng et al. [97] used CNN and LSTM to evaluate six different time-dimension pooling operations, including Slow pooling and Conv pooling, among others. Next, He et al. [100] suggested a deep bidirectional LSTM that similarly combined the benefits of temporal information extraction with bi-LSTM and spatial features extraction with CNN. The method can process long videos by analyzing features at predetermined intervals, producing better results. A lightweight motion-based attention mechanism and a correlation-based spatial attention mechanism are both included in the suggested VideoLSTM [101]. By learning separate hidden state transitions of storage units at separate spatial locations, the Lattice LSTM [102] extends the LSTM and can precisely describe long-term and complex motions.

Due to the LSTM module's construction, parallel computing is not feasible. The most widely used deep learning architecture nowadays, Transformer [94], is capable of resolving this issue. Girdhar Rohit et al. [103] combined context features using Transformer's architecture and added an attention mechanism. Using mutual attention fusion and inter-frame attention encoder blocks, Li et al. [104] introduced the Transformer-based RGB-D egocentric action recognition framework (Trear). Moreover, ShuttleNet [105] emphasizes parallel work while taking into account feedforward and feedback connections in RNNs, learning long-term relationships, and parallel computation. FAST-GRU is a strategy created by FASTER [106] that expedites training by lowering the cost of redundant frame processing, as shown in Fig. 8.

III. SKELETON-BASED ACTION RECOGNITION METHOD

It has become simpler to obtain joint position data as a result of the advancement of depth cameras like Kinect, Asus Xtion, and Intel RealSense and the maturing of joint coordinate estimation algorithms like OpenPose and SDK [107]. Skeleton data also has better robustness to illumination, view angle, and backdrop occlusion compared to RGB data, and it can better prevent noise influence. Researchers prefer the HAR based on skeleton data because it has more focused information and significantly lowers the calculation of redundant information.

By feature extraction method, HAR based on skeleton data can be divided into deep learning methods based on deep features and machine learning methods based on handcrafted features. Additionally, as skeleton data is dependent on pose estimation algorithms, this section methodically covers well-known posture estimation algorithms and offers work on skeleton-based action recognition from the perspective of features.

A. POSE ESTIMATION

In order to reconstruct the human limb trunk, the human pose is estimated by detecting the position information of the joints in the human skeleton and determining the connection between the joints. Traditional methods for estimating human pose [108] rely on manually labeling features and regression to obtain the joint coordinates, but the accuracy is low. Deep learning-based human pose estimation, which can be separated into 2D and 3D pose estimation, has emerged as a key research area.

1) 2D HUMAN POSE ESTIMATION

The goal of a 2D human pose estimate is to locate the important human body parts in an image and connect them in a sequential manner to create a human skeleton graph. The classification of single and multiple human targets is generally used in research.

There is only one target to be discovered in the single-person pose estimate image. All the joints in the target body are first recognized, followed by the bounding box image of the target. In general, there are two categories of single-person pose estimation models. First is the direct regression-based approach, which involves regressing key points directly from features, as shown in Fig. 9. Examples are DeepPose [109], Deconstructive Key Point Regression (DEKR) [110], Self-Correction Model [111], and the Structure-Aware Regression Method [112]. The alternative,

FIGURE 9. An example of regressing the key-points in [110].

known as a heat map-based framework [113], [114], [115], [116], involves first creating a heat map first and determining the locations of the critical points from the heat map.

Multi-person pose estimation necessitates the concurrent processing of detection and localization operations, unlike single-person pose estimation. Depending on the detecting step, top-down and bottom-up approaches for estimating human pose can be distinguished. Top-down based methods execute pose estimation on a single human target after using a target detection algorithm to detect multiple people in the image. G-RMI [117], Mask R-CNN [118], AlphaPose [119], HRNet [120], and DNAnet [121] are a few examples. The bottom-up approach includes joint detection and clustering, which first detects every joint in the image and then clusters the joints into a person using the appropriate algorithm to estimate pose. DeepCut [122], OpenPose [123], Lightweight OpenPose [124], PiPaf [125], and HigherHRNet [126] are examples of bottom-up approaches that do away with the notion of first detecting people.

2) 3D HUMAN POSE ESTIMATION

By estimating information such as the 3D coordinate positions and angles of body joints, 3D human pose estimation attempts to construct a body representation. The three major categories of deep learning-based 3D human pose estimation are listed below.

These methods directly forecast 3D pose coordinates from a single image using a large network structure. Deep learning was first applied to a 3D human pose estimation study by Li et al. [127]. Based on this, Park et al. [128]and Tekin et al. [129] conducted more research. Heatmap regression can preserve more image data, and it is generally accepted to use the heatmap of key human skeleton points to estimate 3D human poses [114], [130], [131], [132], [133].

Researchers have tried combining 2D and 3D pose networks [134], [135], [136], [137] or using 2D skeleton sequences as input [116], [138], [139], [140] in an effort to overcome the limitations of the direct regression method and networks in model optimization and their usefulness in a real-world setting.

These methods require 2D pose information with complementary data on human joint points and motion characteristics to develop a network model for a 3D human pose estimate [112], [141], [142]. They are based on 2D information with additional image information, geometric constraints, and other requirements.

B. SKELETON-BASED HANDCRAFTED ACTION RECOGNITION

Handcrafted features are specified by the researcher based on prior knowledge or statistical features retrieved from action data, which can be used to describe the dynamics or statistical characteristics of the action.

Depth motion map (DMM) [143] was proposed in an effort to represent actions by calculating motion data from depth information. DMM created three motion history maps by projecting and compressing the spatiotemporal depth structure from the top, side, and front viewpoints, and then represented them with HOG features. Lastly, actions were described by concatenating the extracted features. Yang et al. [144] constructed a super normal vector feature(SNV) to represent actions based on the depth map sequence. Local binary-valued pattern features were employed by Chen et al. [145] to describe the DMM-based actions instead of HOG.

Numerous academics suggested various skeleton representation methods to boost the algorithm's effectiveness and efficiency. Vemulapalli et al. [146] employed curves in the Lie group to mimic the motion after modeling the geometric connections between various body components using three-dimensional rotation and translation operations. The low-latency oriented model proposed by Cai et al [147]. is robust in computing joint position-related features. A new approach that enables real-time tracking was proposed by Papadopoulos et al. [148] and is based on the determination of the spherical angle between the joints. Su et al. [149] recently extracted features of statistical attributes, such as mean and variance, as well as features of physical attributes, such as relative location of joints, to conduct research.

Handcrafted features are highly interpretable and straightforward. Yet, they fall short of fully describing the overall state of the motion because they depend on the researcher's a priori knowledge, which is more individualized and difficult to generalize.

C. DEEP LEARNING-BASED ACTION RECOGNITION WITH SKELETON

Recently, the benefits of merging skeleton data with deep learning have been gradually demonstrated, and a number of outstanding approaches, primarily based on RNN, CNN, and GCN, have been developed.

1) RNN-BASED METHODS

Recurrent neural networks (RNNs) are used in natural language processing (NLP) [150], video analysis [151], [152],

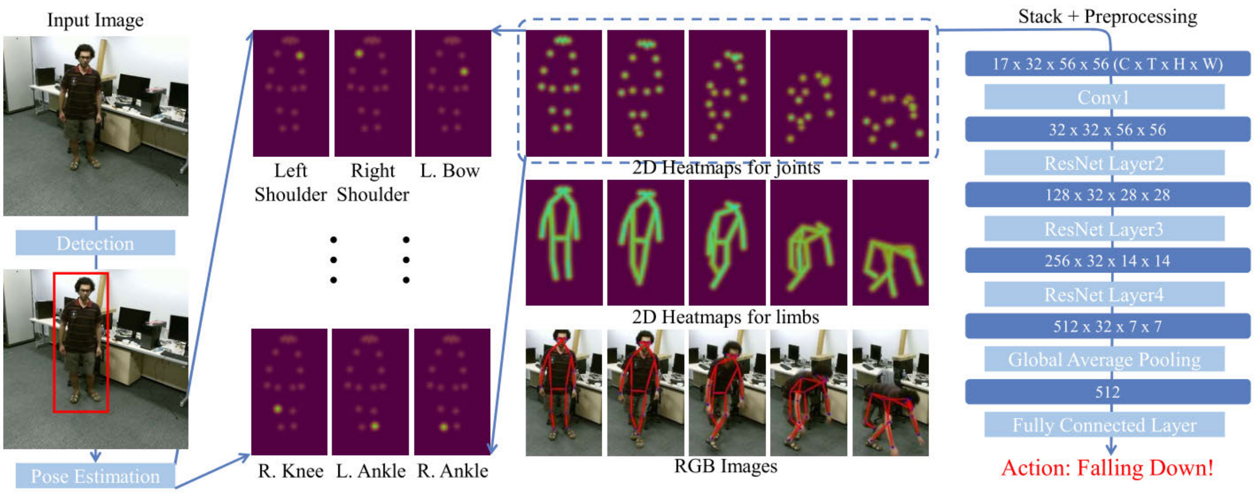

FIGURE 10. The framework of PoseConv3D.

[153], and RGB-based action recognition [154] and offer considerable advantages for processing variable-length sequence data [155], [156].

If the sequence is too long during actual training, gradient explosion and disappearance may occur during optimization. The independent recurrent neural network (IndRNN) [157] has been suggested as a solution to this issue. Gradient backpropagation is regulated by IndRNN over time, enabling the network to acquire long-term dependencies. The function of neurons in each layer can also be explained by the fact that neurons in the same layer are independent of one another and linked across layers.

By developing "recurrent bodies," long short-term memory (LSTM) networks improve upon RNNs' drawbacks and have significant benefits in the extraction of temporal sequence features. Lee et al. [158] propose that LSTM networks with varying time steps may "remember" distinct attributes. They suggested an integrated temporal sliding long short term memory (TS-LSTM) network that takes into account both short- and medium-term features in addition to long-term ones.

When all joints are used as inputs, irrelevant joints degrade the network's performance as noise, so more attention should be given to joints with important information. Considering the interference of noisy data, Liu et al. [159] suggested a global context-aware attention LSTM (GCA-LSTM) with a circular attention mechanism. With the aid of global context memory units, GCA-LSTM is better able to selectively pay attention to the joints of varying importance. To increase the network's expressiveness, they integrate coarse- and fine-grained attention simultaneously.

Co-occurrence features improve the expressiveness of network features by combining features from different dimensions. Zhu et al. [160] proposed a regularization technique for investigating skeleton co-occurrence features. Si et al. [161] introduced the attention-enhanced graph convolution LSTM network (AGC-LSTM), which can extract the co-occurrence feature of the spatio-temporal dimension, and incorporate an attention method to improve the information of key joints. Additionally, they suggested a temporal hierarchy to expand the AGC-LSTM layer's temporal perceptual domain, which improves the high-level semantic representation and greatly lowers the computing cost. The attention recurrent relational networks (ARRN-LSTM) that Zheng et al. [162] suggested can modularize both spatial layout and temporal motion features.

2) CNN-BASED METHODS

The CNN model, which is frequently employed in skeleton-based action recognition, has a great ability to extract high-level semantic information fast and readily.

To meet the criteria of CNN input, it is crucial to convert 3D skeleton data from vector frames to pseudo-images and afterwards extract the features of the pseudo-images. Du et al. [164] developed an end-to-end hierarchical structure using spatial relations as an innovator of skeleton image representation. They represented the coordinates of the 3D skeleton as sequences and linked them in time. The final step was to extract and identify features from the generated pictures using a CNN. Following [164], Ke et al. [165] suggested an improved skeleton sequence representation in which 3D coordinates were divided into three grayscale images. Inspired by the RGB-based two-stream CNN [69], Li et al. proposed a skeleton-based two-stream CNN [166], in which one stream receives the initial skeleton coordinates as input, and the other stream receives the difference in joint coordinates between two subsequent frames. Ding et al. [167] employed CNN to obtain high-level semantic features from RGB textured images that were generated from the skeletal data.

The aforementioned approaches require a lot of processing work and frequently miss critical information. To get around this problem, Caetano et al. specified SkeleMotion [168] as a novel skeleton image representation to be used as an input to the neural network. Then Caetano et al. conducted

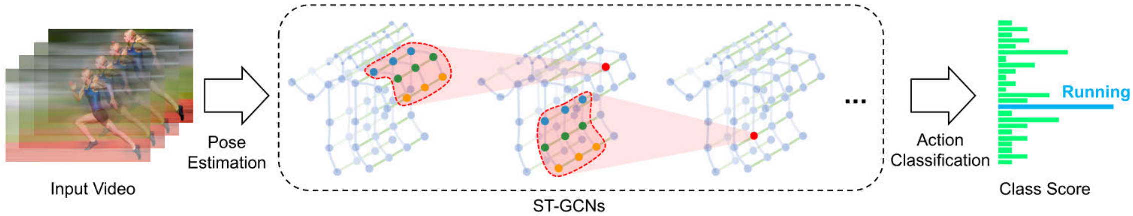

FIGURE 11. Spatio-temporal graph convolution model (ST-GCN).

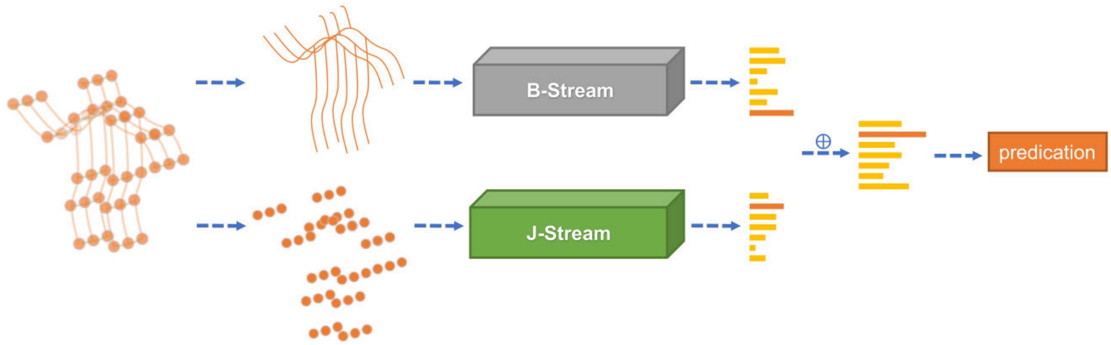

FIGURE 12. Illustration of the overall architecture of the 2s-AGCN in [163]. The scores of two streams are added to obtain the final prediction.

additional research [169] so that the input is no longer limited to the skeleton's coordinates. The tree structure reference joint image (TSRJI) was used as the skeleton representation in this research, and the reference joint and the tree structure skeleton were used together to prevent CNN's disregard of the skeleton structure.

Numerous researchers have attempted to find a solution to the long-time dependence problem because convolutional neural networks are not effective at extracting long-distance motion information. A subsequence attention network (SSAN) was suggested by Liu et al. [170] to more effectively record long-term features after applying 3DCNN to skeleton data in the initial stages. Liu et al. [171] exploited the macro-temporal correlations between skeleton joints using Fourier time pyramids, and then caught the micro-temporal interactions using a hierarchical method.

Recently, Duan et al. [172] developed a novel framework for skeleton-based HAR, PoseConv3D, as shown in Fig. 10. PoseConv3D outperforms the GCN-based method in terms of learning spatio-temporal features, resistance to pose estimation noise, and cross-dataset generalization. PoseConv3D can also handle multi-person scenes without incurring extra computation costs.

3) GCN-BASED METHODS

Both CNNs and RNNs learn with alignment regularity for euclidean data, but they are unable to deal with non-euclidean data. Gori et al. [173] first suggested GNNs in 2005 as a way to explore graph data. Later, by extending CNN on graph data, the graph convolutional neural network (GCN) was gradually suggested. GCN can be used to learn graph data directly because human skeleton data, which consists of joint points and skeletal lines, can be thought of as non-Euclidean graph data. Spectral GCN and spatial GCN are the two major branches of GCN, respectively.

- *Spectral GCN:*
  Using the eigenvalues and eigenvectors of the graph Laplacian matrix, spectral GCN converts the graph from the temporal domain to the frequency domain [174], but the computation is laborious. By only allowing the filter to operate on one neighbor node around each node, Kipf et al. [175] improved the spectral GCN method. A new spectral multi-Laplacian graph convolution network (MLGCN) was recently suggested by Mazari et al. [176] to learn the graph Laplacian, which is used as a convex combination of other basic Laplacians. Although spectral GCN has demonstrated its efficacy in HAR tasks, the computational expense makes it challenging to capture high-level information from graphs.
- *Spatial GCN:*
  Spatial GCNs are more efficient and work better than spectral GCNs in terms of computation cost. Therefore, spatial GCN is the main emphasis of the majority of the current GCN-based HAR techniques. Yan et al. [177] made the initial concept for a spatio-temporal graph convolutional network model (ST-GCN). As shown in Fig. 11, ST-GCN takes the bodily joints as the vertices and take the bodily bones in the same frame as well as the sequence frame, as the edges of the spatio-temporal graph.
  The flexibility of the graph network is somewhat reduced because each layer's parameters are fixed. Shi et al. [163] suggested a novel two-stream adaptive graph convolutional network (2sAGCN) to address this issue, shown in Fig. 12. Either the BP algorithm or an

end-to-end method can be used to learn the topology of the graph in the model. The 2sAGCN model is more adaptable to diverse data samples thanks to this data-driven methodology, which boosts its flexibility. The attention mechanism is also introduced to make the 2sAGCN more robust. In light of the fact that the joint importance varies for each action, Shiraki et al. [178] presented the spatio-temporal attentional graph convolutional network (STA-GCN). The STA-GCN method is the first to take into account the significance and interrelationship of joints, which inspired some researchers to look into drawing more focus to the GCN [179], [180].

The development of GCN-based models has been the subject of numerous studies. As an illustration, the innovative shift-graph operation in shift-GCN [181] improves the flexibility of the spatio-temporal graph's receptive domain, and the lightweight dot convolution aids in the reduction of the number of feature channels. With bottleneck structure and partial attention blocks, ResGCN [182] is an algorithm for residual graph convolution networks that boosts the efficiency, speed, and readability of GCN for HAR.

Thakkar et al. and Li et al. suggested various techniques for segmenting body parts, which were inspired by the notion that the human skeleton is a combination of numerous body parts. Thakkar et al. [183] proposed a partial-based graph convolutional network (PB-GCN). Four node-sharing subgraphs of the skeleton graph are learned using the PB-GCN algorithm. Another one is the spatio-temporal graph routing (STGR) scheme that Li et al. [184] suggested in order to untangle the semantic connections between joints.

IV. COMMON DATASET

With the continuous exploration of HAR, a large number of datasets related to action recognition have been created to evaluate and examine the performance of algorithms. Based on the types of data, the datasets are divided in this survey into RGB datasets and skeleton sequence datasets.

A. RGB DATASETS

The widely used RGB dataset, which may be gathered directly from actual situations, will be presented in this part. Table 2 lists the basic information of some commonly used RGB datasets.

- *UCF101* [185]
  There are 13,320 videos overall and 101 action categories in this compilation of real-world YouTube videos. UCF101 is the most diverse category of action, including camera movement, object shape and pose, object scale, perspective, complex backgrounds, and lighting conditions.
- *KTH* [29]
  KTH is a video intercept from a monitoring device over time that contains one or more sequences of human behavior. A distinct time step is used to represent each sequence of human actions. Each human behavior sequence is segmented, and the segmented dataset is then broken down into roughly 60,000 sub-segments with a range of 5 to 20 actions apiece.
- *HMDB51* [186]
  HMDB51 is an open source human behavior dataset that contains approximately 7000 video clips organized into 51 action categories. Each action consists of at least 101 video clips and has a different temporal and spatial scale. Each clip has a label identifying the activity as well as information about the visible body parts, camera motion, camera angle, number of participants in the action, and video quality.
- *Sports-1M* [68]
  It contains 1133,158 video URLs, automatically labeled with 487 tags. This is one of the largest video datasets containing videos of various sports, including Shaolin Temple Kung Fu and Wing Chun Kung Fu. The dataset is very complex and challenging with great variation in appearance and pose, camera motion, and background noise.
- *Kinetics*
  It contains a series of datasets, including Kinetics-400 [187], Kinetics-600 [188], Kinetics-700 [189], AVA Kinetics [190], and Kinetics 700-2020 [191]. Depending on the version of the dataset, 400/600/700 categories of human actions were covered. For each class of action, there are at least 400/600/700 video clips. With a duration of around 10 seconds, each clip is tagged by an action category. It serves as a significant benchmark in HAR, similar to ImageNet in image recognition. This dataset appears in many contexts and has the ability to pre-train some datasets before training, in addition to direct clip recognition. Kinetics is largely regarded as the first major, large video-categorization dataset. The accuracy of this dataset can be further improved.
- *ActivityNet* [192]
  The ActivityNet series has gone through various iterations since it was first made accessible in 2015. The

TABLE 2. The basic information of some commonly used RGB datasets.

Datasets	Year	Videos	Views	Actions	Subjects
KTH	2004	599	1	6	25
HMDB51	2011	7,000	-	51	-
UCF101	2012	13,320	-	101	-
Sports-1M	2014	1,133,158	-	487	-
ActivityNet	2015	28,000	-	203	-
YouTube-8M	2016	8,000,000	-	4,716	-
Kinetics	2017	500,000	-	600	-
Moments in Time	2017	1,000,000	-	339	-
HACS	2019	504,000	-	200	-
HVU	2020	572,000	-	3,142	6
AViD	2020	467,000	-	887	-

most recent version, ActivityNet 200 (V1.3), includes 200 daily life activities. It has 1024 training, 4926 validation and 5044 test videos. Per class, there are approximately 137 untrimmed films and 1.41 action occurrences.

- *YouTube8M* [193]
  With 8 million YouTube videos (500,000 hours of video in total) and 3,862 action classifications, it is the largest video database to date. The video annotation system on YouTube assigns one or more tags to each video. A training set, a validation set, and a test set were created from the dataset in the following proportions: 70:20:10. Additionally, temporal location data was included to the validation dataset.
- *HACS* [194]
  This dataset is a new large-scale dataset introduced in 2019 to track and detect human actions collected from online videos. HACS contains 504K clip videos, of which 1.4K million videos have full action videos (from the beginning to the end of the action). These videos were annotated with the 200 action categories used in ActivityNet (V1.3) [192].
- *HVU* [195]
  This dataset was released in 2020 and focuses on three tasks of video classification, video description and video clustering to help understand multi-label multi-task videos. The dataset has 3142 classes with an average of 2112 labeled data in one class, of which 481K are used for training, 31K for validation and 65K for testing. HVU describes video information with more comprehensive labels (scene, objects, actions, events, attributes, concepts).
- *AViD* [196]
  Introduced in 2020, the AViD dataset collects anonymous videos from different countries to constitute a large video dataset containing 467k videos and 887 action classes, with each video clip lasting between 3 and 15 seconds. The writers deleted the facial identify during the data gathering procedure to safeguard the privacy of the video producers. Consequently, it's possible that the AViD dataset is not the best option for detecting facially significant activities.
- *Moments-in-Time* [197]
  The MIT dataset contains 1 million tagged video clips, of which 802,264 were used for training, 33,900 for validation and 67,800 for testing, distributed across 339 categories. The visual components of the videos on MIT include individuals, animals, objects, or natural events. The information is used to create models that can abstract and make inferences about complicated behavior among individuals.

B. SKELETON-BASED ACTION RECOGNITION DATASETS

Many deep skeleton sequence datasets have also been produced with the use of some depth sensors, such as Microsoft Kinect. In this section, we present several commonly used skeleton datasets. Table 3 lists the basic information of some commonly used deep skeleton sequence datasets, including the data modality, number of captures, and number of categories of the datasets.

- *CMU Mocap* [198]
  A 3D skeleton with six degrees of freedom in each joint was created by the motion capture database at Carnegie Mellon University using 12 VICON MX-40 infrared cameras. 144 people participated in the interactive and single-subject activities. The activities were broken down into 23 subcategories encompassing context and scenario, mobility, physical activity and sport, human contact, and environment interaction.
- *HDM05* [199]
  Five amateur actors performed the action sequences in the HDM05 dataset, which was released in 2005. Each of the nearly 70 activity categories in the dataset has between 10 and 50 performers. The C3D mocap file format is used to store the produced 3D trajectory data. The VICON MX system included six RGB cameras and six IR cameras to record the videos.
- *MSR Action3D* [200]
  The dataset consisted of 20 actions of the console interaction, performed three times by each of the 7 subjects. Depth data was recorded at 15 frames per second (fps). The activities were divided into three categories: AS1, AS2, and AS3, where AS1 and AS2 represent comparable acts and AS3 represents sophisticated actions. Without RGB video, the dataset just contains depth and skeleton data.
- *CAD 60 [201]*
  RGB video and depth maps were recorded with Kinect. The dataset recorded four subjects performing 12 different activities (including several sub-activities) in five different environments. These included daily actions in the office, kitchen, bedroom, bathroom, and living room.
- *UT-Kinect* [202] 10 subjects performed 10 different indoor actions, and video was recorded with a still Kinect. Each subject performed each action twice, repeatedly. The dataset recorded RGB video, depth, and skeletons.
- *CAD-120* [203]
  After collecting 120 videos of human-object interactions, we labeled the dataset with human skeleton trajectories, object trajectories, object labels, subactivity labels, and high-level actions for each video. Four participants performed a total of 10 sub-activities in 10 different situations, including cooking oatmeal, taking medicine, and putting things away.
- *UWA3D Multiview* [204]
  The dataset contains 30 videos of daily indoor actions taken by 10 different people at different scales, all taken with Kinect. The high degree of similarity in this dataset poses an additional challenge.

TABLE 3. The basic information of some commonly used deep skeleton sequence datasets,where RGB denotes RGB data, IR denotes infrared data, S denotes skeletal data, and D denotes depth data.

Datasets	Year	Sensors	Subject	Views	Actions	Data
CMU Mocap	2003	Vicon	144	-	23	RGB+S
HDM05	2007	RRM	5	6	>70	RGB+S
MSR Action3D	2010	-	20	1	20	D+S
CAD 60	2011	Kinect v1	4	-	12	RGB+D+S
UT-Kinect	2012	Kinect v1	10	4	10	RGB+D+S
CAD-120	2013	Kinect v1	4	-	10+10	RGB+D+S
UWA3D Multiview	2014	Kinect v1	10	1	30	RGB+D+S
NTU RGB+D	2016	Kinect v2	40	80	50+10	RGB+IR+D+S
SYSU	2017	Kinect v1	40	1	12	RGB+D+S
Kinetics-Skeleton	2017	YouTube	-	-	400	RGB
UW-IOM	2019	Kinect	20	-	17	RGB+D+S
NTU RGB+D 120	2019	Kinect v2	106	155	94+26	RGB+IR+D+S
HiEve	2020	-	-	-	14	RGB+S

- *NTU RGB+D* [205]
  Three Kinect V2 cameras were used to record the 2016-created NTU RGB+D dataset, which includes 56,880 video samples and 60 action categories. Each sample includes RGB video, infrared video, depth image sequences, and 3D skeleton images. A skeleton contains 25 joints in total. 11 of the activities were interactive, while 49 of the acts were completed by a single individual.
- *SYSU* [206]
  This dataset records 40 participants' interactions between people and objects. Each participant used six different objects in 12 different manipulations. The skeleton data, depth sequences, and RGB video were all recorded by Kinect in one view.
- *Kinetics-Skeleton* [187]
  The Kinetics-Skeleton dataset is derived from the Kinetics video action recognition dataset. Using Openpose's pose estimation algorithm, they searched all major skeleton joints in the videos to create Kinetics-Skeleton, a database of nearly 300,000 videos and 400 actions that is still widely used today.
- *UW-IOM* [207]
  The University of Washington's indoor object manipulation dataset, which includes films of 20 persons classified into 17 different movement categories, is intended to identify hazards to the human body. Each participant controlled six objects in the films, which were separated into 17 action categories and averaged 12 frames per second on the Kinect.
- *NTU RGB+D 120* [208]
  The NTU RGB+D dataset was upgraded in 2019 with the addition of 60 classes and 57,600 extra video samples. The cameras and data types are identical to those of NTU RGB+D. There are 82 daily activities, 12 health-related actions (e.g., nose blowing,throwing up), and 26 interactive actions. (e.g., shaking hands, pushing each other).
- *HiEve* [209]
  The dataset focuses on human-centric analysis of a variety of people and complex events: videos of 9 different scenes and 32 different realistic environments were collected. Each subject in the videos has a bounding box, 14 joints skeletons, human identity, and human actions. Overall, there are 14 types of actions.

V. CHALLENGE

Although significant advancements have been made in HAR based on two data modalities, a number of difficulties still exist as a result of the complexity of the numerous facets of this task.

A. RGB-BASED CHALLENGES

- *Huge Amount of Calculations*
  Compared to images, RGB video offers a lot more data, necessitating the creation of strong neural network models. In real-world contexts, it is challenging to meet the demands of real-time applications due to the hardware constraints imposed by the CPU and GPU, which significantly degrade the efficiency of network computation. Also, the labor and time expenses for precise and efficient labeling of video data are enormous due to the variety and size of the data.
- *Complexity of The Environment*
  Some action recognition algorithms perform well in situations that can be controlled, while they underperform in uncontrolled outside settings. This is mostly due to the fact that motion vector noise can drastically impair resolution and that extracting action features from complicated images is extremely difficult. For instance, accurate action feature extraction is hard due to the camera's quick movement. Accurate recognition will also be impacted by other environmental issues, including poor lighting, shifting perspectives, dynamic backgrounds, etc.

- *Limitations of The Dataset*
  The dataset contains both intra-class differences and inter-class similarities. Several people present the same action in different ways, and even the same person may perform it in various ways. For different actions, there may be similar presentations. Moreover, many available datasets contain unpruned sequences, which might diminish the timeliness and lower the recognition accuracy of the network.

B. SKELETON-BASED CHALLENGES

- *Pose Preparation*
  Since the acquisition of skeleton data relies on depth cameras and sensors, it is influenced by the environment's complexity and diversity, the duration of the capture, and the exposure conditions of the capture equipment. Another common issue in daily life is occlusion, which is brought either by surrounding objects or human interaction. All of them raise the detection error for skeletons.
- *Viewpoint Variation*
  It is challenging to precisely distinguish skeleton features from one perspective from another, because some features are lost during the view change. While current RGBD cameras [210], [211], [212], [213] can normalize 3D human skeletons [214], [215] from various viewpoints to a single pose with viewpoint invariance using a pose estimation transformation matrix, some of the relative motion between the original skeletons may be lost in the process.
- *Single Data Scale*
  Since most skeleton datasets provide information based on the body joint scale, many approaches only extract human joint scale features, which results in the loss of fine joint features. Additionally, some actions, like tooth brushing, shaving, applying lipstick, etc., show similar joint interactions. Hence, it is crucial to improve local feature extraction without sacrificing holistic feature extraction [216], [217], [218], [219].

VI. FUTURE RESEARCH TRENDS

We describe a few potential future research trends after synthesizing the current situation and issues with research methodologies and applications of RGB-based and skeleton-based action recognition.

A. DEVELOPMENT OF NEW DATASETS

Data are just as crucial to deep learning as model building. It is still challenging to generalize to realistic scenes when using existing datasets because of aspects like realistic surroundings and dataset size. Moreover, the majority of datasets are oriented toward spatial representation [220], and there aren't many that can be long-term modeled. However, due to regional limitations and privacy concerns, such as those mentioned above, YouTube dataset managers usually only provide IDs or video links for users to download, not the actual videos. As a result, some videos are no longer viewable, resulting in a loss of 5% of videos annually on average [12]. These difficulties spur us to gather fresh datasets in order to advance our research.

B. DATA AUGMENTATION

Deep neural networks perform exceptionally well when given a wide variety of datasets; hence, it is essential to incorporate data augmentation as a data space solution to address the issue of restricted data. In the field of image recognition, a variety of data augmentation methods have been proposed, including deep learning-based and basic image processing methods. These methods include kernel filters [221], random erasing [222], feature space augmentation [223], adversarial training [224], generative adversarial networks [225], and meta-learning [226], [227]. In the field of action recognition, typical data augmentation methods include horizontal flipping, clipping subclips, and video splicing [228], [229], [230].The generated videos, however, lack realism. Moreover, Zhang et al. [231] employed GAN to generate new samples and "self-paced selection" for training. Gowda recently put up the Learn2Augment [232] proposal, which chooses video synthesis of the foreground and background videos as a data augmentation technique, producing diverse and realistic new samples.

C. IMPROVEMENTS IN MODELS

HAR study is dominated by deep learning models, similar to other computer vision developments. Currently, the continual advancement of deep architectures is necessary for both RGB-based and skeleton-based methods of action recognition. The following three areas generally correspond to model improvements.

- *Long-term Dependency Modeling:*
  Long-term correlations describe the sequence of actions that take place in lengthy sequences, which are similar to the storage in our brains. One pattern evokes the next when we think back on an incident. It is crucial to concentrate on the temporal component in addition to the spatial modeling because this indicates that there are extremely strong correlations between adjacent temporal features.
- *Multi-modality Modeling:*
  Multi-modality modeling relies on the fusion of data from various devices (e.g., audiovisual data). The two major types of multi-modality video understanding are described below. One is the use of multi-modality data to improve video representations, such as scene, object, action, and audio [233], [234]. Recently, there has been an increase in interest in multi-modality fusion using depth, skeleton, and RGB data. The alternative strategy is to create a model that can be pre-trained to manage the signal using multi-modality data [235], [236], [237].
- *Efficient Modeling:*
  It is necessary to create an effective network architecture because the majority of existing methods have problems

with the complexity of the models, the enormous number of parameters, and the inability to accomplish real-time. We can use efficient methods suggested for image classification, such as distributed training [238], [239], mobile networks [240], [241], hybrid precision training, etc., as well as model compression, model quantization, and model pruning.

D. ACTIONS PREDICTION

Short-term prediction and long-term prediction are the two main kinds of action prediction tasks. The goal of short-term prediction is to infer action labels based on temporally incomplete actions, which focuses on quick action videos that typically last a few seconds. The process of making long-term predictions involves presuming that present behavior will influence future behavior. It focuses on lengthy films that continue for many minutes in an effort to simulate action changes. More formally, given an action video x_a, where x_a can be a complete or incomplete action execution, the goal is to infer the next action x_b. Here, x_a and x_b are two independent, semantically meaningful, and temporally related actions [14].

Finding and modeling temporal correlations in massive amounts of data is the key to this action prediction research. The interpretability of time scales, how to model long-term correlations, and how to use multimodal data to improve predictive models are just a few of the unexplored directions for this research.

VII. CONCLUSION

This survey provides a comprehensive overview of human action recognition methods and systematically summarizes and concludes the methods according to data types including RGB data and skeleton data. It also provides relevant analysis and discussion of various methods, indicating the advantages and disadvantages of each method. In addition, the existing popular human action datasets, including RGB datasets and skeleton datasets, are also introduced. Finally, we analyze the great challenges currently facing the task of human action recognition based on RGB and skeleton data, respectively, and summarize the promising research directions in the field of action recognition to help scholars entering the field or conducting long-term research.

REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 1–43, 2011.

[2] M. Ziaeefard and R. Bergevin, "Semantic human activity recognition: A literature review," *Pattern Recognit.*, vol. 48, no. 8, pp. 2329–2345, Aug. 2015.

[3] K. M. Sagayam and D. J. Hemanth, "Hand posture and gesture recognition techniques for virtual reality applications: A survey," *Virtual Reality*, vol. 21, no. 2, pp. 91–107, Jun. 2017.

[4] M. Schröder and H. J. Ritter, "Deep learning for action recognition in augmented reality assistance systems," in *Proc. Special Interest Group Comput. Graph. Interact. Techn. Conf. (SIGGRAPH)*, Los Angeles, CA, USA, Jul. 2017, p. 75.

[5] T. Bates, K. Ramirez-Amaro, T. Inamura, and G. Cheng, "On-line simultaneous learning and recognition of everyday activities from virtual reality performances," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 3510–3515.

[6] M. Meng, H. Drira, and J. Boonaert, "Distances evolution analysis for online and off-line human object interaction recognition," *Image Vis. Comput.*, vol. 70, pp. 32–45, Feb. 2018.

[7] R. Zhen, W. Song, Q. He, J. Cao, L. Shi, and J. Luo, "Human–computer interaction system: A survey of talking-head generation," *Electronics*, vol. 12, no. 1, p. 218, Jan. 2023.

[8] G. Guo and A. Lai, "A survey on still image based human action recognition," *Pattern Recognit.*, vol. 47, no. 10, pp. 3343–3361, Oct. 2014.

[9] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," *Frontiers Robot. AI*, vol. 2, p. 28, Nov. 2015.

[10] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *Vis. Comput.*, vol. 29, no. 10, pp. 983–1009, Oct. 2013.

[11] S.-R. Ke, H. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, Jun. 2013.

[12] Y. Zhu, X. Li, C. Liu, M. Zolfaghari, Y. Xiong, C. Wu, Z. Zhang, J. Tighe, R. Manmatha, and M. Li, "A comprehensive study of deep video action recognition," 2020, *arXiv:2012.06567*.

[13] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, Feb. 2019.

[14] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," 2018, *arXiv:1806.11230*.

[15] N. Ma, Z. Wu, Y. Cheung, Y. Guo, Y. Gao, J. Li, and B. Jiang, "A survey of human action recognition and posture prediction," *Tsinghua Sci. Technol.*, vol. 27, no. 6, pp. 973–1001, Dec. 2022.

[16] G. Saleem, U. I. Bajwa, and R. H. Raza, "Toward human activity recognition: A survey," *Neural Comput. Appl.*, vol. 35, no. 5, pp. 4145–4182, Feb. 2023.

[17] Y. Xing and J. Zhu, "Deep learning-based action recognition with 3D skeleton: A survey," *CAAI Trans. Intell. Technol.*, vol. 6, no. 1, pp. 80–92, 2021.

[18] M. Feng and J. Meunier, "Skeleton graph-neural-network-based human action recognition: A survey," *Sensors*, vol. 22, no. 6, p. 2091, Mar. 2022.

[19] L. Feng, Y. Zhao, W. Zhao, and J. Tang, "A comparative review of graph convolutional networks for human skeleton-based action recognition," *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 4275–4305, Jun. 2022.

[20] P. Gupta, A. Thatipelli, A. Aggarwal, S. Maheshwari, N. Trivedi, S. Das, and R. K. Sarvadevabhatla, "Quo vadis, skeleton action recognition?" *Int. J. Comput. Vis.*, vol. 129, no. 7, pp. 2097–2112, Jul. 2021.

[21] L. Song, G. Yu, J. Yuan, and Z. Liu, "Human pose estimation and its application to action recognition: A survey," *J. Vis. Commun. Image Represent.*, vol. 76, Apr. 2021, Art. no. 103055.

[22] M. B. Shaikh and D. Chai, "RGB-D data-based action recognition: A review," *Sensors*, vol. 21, no. 12, p. 4246, Jun. 2021.

[23] S. Majumder and N. Kehtarnavaz, "Vision and inertial sensing fusion for human action recognition: A review," 2020, *arXiv:2008.00380*.

[24] L. Wang, D. Q. Huynh, and P. Koniusz, "A comparative review of recent Kinect-based action recognition algorithms," *IEEE Trans. Image Process.*, vol. 29, pp. 15–28, 2020.

[25] H. Wang, A. Kläser, C. Schmid, and C. Liu, "Action recognition by dense trajectories," in *Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Colorado Springs, CO, USA: IEEE Computer Society, Jun. 2011, pp. 3169–3176.

[26] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3551–3558.

[27] X. Peng, C. Zou, Y. Qiao, and Q. Peng, "Action recognition with stacked Fisher vectors," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*. Zurich, Switzerland: Springer, Sep. 2014, pp. 581–595.

[28] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj, "Beyond Gaussian pyramid: Multi-skip feature stacking for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 204–212.

[29] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, 2004, pp. 32–36.

[30] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, Jun. 2010.

[31] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, May 2013.
[32] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int. Workshop Vis. Surveill. Perform. Eval. Tracking Surveill.*, Oct. 2005, pp. 65–72.
[33] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition," *Int. J. Comput. Vis.*, vol. 119, no. 3, pp. 219–238, Sep. 2016.
[34] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. 15th ACM Int. Conf. Multimedia*, Sep. 2007, pp. 357–360.
[35] Y. Kong, Z. Tao, and Y. Fu, "Deep sequential context networks for action prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3662–3670.
[36] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *Proc. 11th Eur. Conf. Comput. Vis. (ECCV)*. Heraklion, Crete, Greece: Springer, Sep. 2010, pp. 392–405.
[37] L.-P. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[38] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Proc. CVPR*, Jun. 2011, pp. 3337–3344.
[39] C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Conditional models for contextual human motion recognition," *Comput. Vis. Image Understand.*, vol. 104, nos. 2–3, pp. 210–220, Nov. 2006.
[40] Q. Shi, L. Cheng, L. Wang, and A. Smola, "Human action segmentation and recognition using discriminative semi-Markov models," *Int. J. Comput. Vis.*, vol. 93, no. 1, pp. 22–32, May 2011.
[41] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.
[42] A. Klaeser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 1–275.
[43] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2004–2011.
[44] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2929–2936.
[45] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[46] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2005, pp. 1395–1402.
[47] I. Laptev and P. Pérez, "Retrieving actions in movies," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[48] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *Proc. 10th Eur. Conf. Comput. Vis. (ECCV)*. Marseille, France: Springer, Oct. 2008, pp. 548–561.
[49] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1521–1527.
[50] L. Wang and D. Suter, "Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[51] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh, "Activity recognition and abnormality detection with the switching hidden semi-Markov model," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2005, pp. 838–845.
[52] S. Rajko, G. Qian, T. Ingalls, and J. James, "Real-time gesture recognition with minimal training requirements and on-line learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[53] N. Ikizler and D. Forsyth, "Searching video for complex activities with finite state models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[54] K. Tang, L. Fei-Fei, and D. Koller, "Learning latent temporal structure for complex event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1250–1257.
[55] Z. Wang, J. Wang, J. Xiao, K. Lin, and T. Huang, "Substructure and boundary modeling for continuous action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1330–1337.
[56] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *Proc. CVPR*, Jun. 2011, pp. 489–496.
[57] C. Yuan, X. Li, W. Hu, H. Ling, and S. J. Maybank, "Modeling geometric-temporal context with directional pyramid co-occurrence for action recognition," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 658–672, Feb. 2014.
[58] B. Wu, C. Yuan, and W. Hu, "Human action recognition based on context-dependent graph kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2609–2616.
[59] C. Fanti, L. Zelnik-Manor, and P. Perona, "Hybrid models for human motion recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 1166–1173.
[60] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *Proc. CVPR*, Jun. 2011, pp. 3273–3280.
[61] Y. Kong, Y. Jia, and Y. Fu, "Interactive phrases: Semantic descriptionsfor human interaction recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 9, pp. 1775–1788, Sep. 2014.
[62] G. Luo, S. Yang, G. Tian, C. Yuan, W. Hu, and S. J. Maybank, "Learning human actions by combining global dynamics and local appearance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2466–2482, Dec. 2014.
[63] C. Yuan, W. Hu, G. Tian, S. Yang, and H. Wang, "Multi-task sparse learning with beta process prior for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 423–429.
[64] S. Yang, C. Yuan, B. Wu, W. Hu, and F. Wang, "Multi-feature max-margin hierarchical Bayesian model for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1610–1618.
[65] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7445–7454.
[66] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, Jun. 2018.
[67] A. Kar, N. Rai, K. Sikka, and G. Sharma, "AdaScan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5699–5708.
[68] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.
[69] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–11.
[70] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6201–6210.
[71] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.
[72] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[73] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream ConvNets," 2015, *arXiv:1507.02159*.
[74] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[75] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941.
[76] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal residual networks for video action recognition," 2016, *arXiv:1611.02155*.
[77] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[78] Y. Wang, M. Long, J. Wang, and P. S. Yu, "Spatiotemporal pyramid network for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2097–2106.
[79] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[80] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[81] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[82] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5534–5542.

[83] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.

[84] Y. Zhao, Y. Xiong, and D. Lin, "Trajectory convolution for action recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–10.

[85] Y. Zhou, X. Sun, Z. Zha, and W. Zeng, "MiCT: Mixed 3D/2D convolutional tube for human action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 449–458.

[86] L. Wang, W. Li, W. Li, and L. Van Gool, "Appearance-and-Relation networks for video classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1430–1439.

[87] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 305–321.

[88] D. Tran, H. Wang, M. Feiszli, and L. Torresani, "Video classification with channel-separated convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5551–5560.

[89] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 200–210.

[90] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 588–597.

[91] 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 *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 20–36.

[92] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 803–818.

[93] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[94] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[95] S. Zhang, S. Guo, W. Huang, M. R. Scott, and L. Wang, "V4D:4D convolutional neural networks for video-level representation learning," 2020, *arXiv:2002.07442*.

[96] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[97] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4694–4702.

[98] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4305–4314.

[99] A. Diba, V. Sharma, and L. Van Gool, "Deep temporal linear encoding networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1541–1550.

[100] 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, pp. 319–331, Jul. 2021.

[101] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. M. Snoek, "VideoLSTM convolves, attends and flows for action recognition," *Comput. Vis. Image Understand.*, vol. 166, pp. 41–50, Jan. 2018.

[102] 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 *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2166–2175.

[103] R. Girdhar, J. J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 244–253.

[104] X. Li, Y. Hou, P. Wang, Z. Gao, M. Xu, and W. Li, "Trear: Transformer-based RGB-D egocentric action recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 1, pp. 246–252, Mar. 2022.

[105] 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 *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 716–725.

[106] L. Zhu, D. Tran, L. Sevilla-Lara, Y. Yang, M. Feiszli, and H. Wang, "FASTER recurrent networks for efficient video classification," in *Proc. 34th AAAI Conf. Artif. Intell. (AAAI), 32nd Innov. Appl. Artif. Intell. Conf. (IAAI), 10th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI).* New York, NY, USA: AAAI Press, Feb. 2020, pp. 13098–13105.

[107] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: A review," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1318–1334, Oct. 2013.

[108] C. Ionescu, F. Li, and C. Sminchisescu, "Latent structured models for human pose estimation," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2220–2227.

[109] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.

[110] Z. Geng, K. Sun, B. Xiao, Z. Zhang, and J. Wang, "Bottom-up human pose estimation via disentangled keypoint regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14671–14681.

[111] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4733–4742.

[112] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2621–2630.

[113] Y. Chen, C. Shen, X. Wei, L. Liu, and J. Yang, "Adversarial PoseNet: A structure-aware convolutional network for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1221–1230.

[114] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 483–499.

[115] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5669–5678.

[116] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2659–2668.

[117] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3711–3719.

[118] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[119] H. Fang, S. Xie, Y. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2353–2362.

[120] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.

[121] K. Zhang, P. He, P. Yao, G. Chen, C. Yang, H. Li, L. Fu, and T. Zheng, "DNANet: De-normalized attention based multi-resolution network for human pose estimation," to be published.

[122] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4929–4937.

[123] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.

[124] D. Osokin, "Real-time 2D multi-person pose estimation on CPU: Lightweight OpenPose," 2018, *arXiv:1811.12004*.

[125] S. Kreiss, L. Bertoni, and A. Alahi, "PifPaf: Composite fields for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11969–11978.

[126] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5385–5394.

[127] S. Li and A. B. Chan, "3D human pose estimation from monocular images with deep convolutional neural network," in *Proc. 12th Asian Conf. Comput. Vis. (ACCV)*. Singapore: Springer, Nov. 2014, pp. 332–347.

[128] S. Park, J. Hwang, and N. Kwak, "3D human pose estimation using convolutional neural networks with 2D pose information," in *Computer Vision—ECCV*. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 156–169.

[129] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua, "Structured prediction of 3D human pose with deep neural networks," 2016, *arXiv:1605.05180*.

[130] B. Tekin, P. Márquez-Neila, M. Salzmann, and P. Fua, "Learning to fuse 2D and 3D image cues for monocular body pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3961–3970.

[131] K. Zhou, X. Han, N. Jiang, K. Jia, and J. Lu, "HEMlets pose: Learning part-centric heatmap triplets for accurate 3D human pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2344–2353.

[132] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1263–1272.

[133] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei, "Deep kinematic pose regression," in *Computer Vision—ECCV*. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 186–201.

[134] C. Chen and D. Ramanan, "3D human pose estimation = 2D pose estimation + matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5759–5767.

[135] F. Moreno-Noguer, "3D human pose estimation from a single image via distance matrix regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1561–1570.

[136] M. Wang, X. Chen, W. Liu, C. Qian, L. Lin, and L. Ma, "DRPose3D: Depth ranking in 3D human pose estimation," 2018, *arXiv:1805.08973*.

[137] C. Li and G. H. Lee, "Generating multiple hypotheses for 3D human pose estimation with mixture density network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9879–9887.

[138] Y. Cheng, B. Yang, B. Wang, Y. Wending, and R. Tan, "Occlusion-aware networks for 3D human pose estimation in video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 723–732.

[139] L. Chen, H. Ai, R. Chen, Z. Zhuang, and S. Liu, "Cross-view tracking for multi-human 3D pose estimation at over 100 FPS," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3276–3285.

[140] E. Remelli, S. Han, S. Honari, P. Fua, and R. Wang, "Lightweight multi-view 3D pose estimation through camera-disentangled representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6039–6048.

[141] E. Jahangiri and A. L. Yuille, "Generating multiple diverse hypotheses for human 3D pose consistent with 2D joint detections," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 805–814.

[142] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3D human pose estimation in the wild: A weakly-supervised approach," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 398–407.

[143] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proc. 20th ACM Int. Conf. Multimedia*, Oct. 2012, pp. 1057–1060.

[144] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 804–811.

[145] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 1092–1099.

[146] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 588–595.

[147] X. Cai, W. Zhou, L. Wu, J. Luo, and H. Li, "Effective active skeleton representation for low latency human action recognition," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 141–154, Feb. 2016.

[148] G. T. Papadopoulos, A. Axenopoulos, and P. Daras, "Real-time skeleton-tracking-based human action recognition using Kinect data," in *Proc. MultiMedia Modeling, 20th Anniversary Int. Conf. (MMM)*, Dublin, Ireland: Springer, Jan. 2014, pp. 473–483.

[149] B. Su, H. Wu, M. Sheng, and C. Shen, "Accurate hierarchical human actions recognition from Kinect skeleton data," *IEEE Access*, vol. 7, pp. 52532–52541, 2019.

[150] M. Kamyab, G. Liu, A. Rasool, and M. Adjeisah, "ACR-SA: Attention-based deep model through two-channel CNN and bi-RNN for sentiment analysis," *PeerJ Comput. Sci.*, vol. 8, p. e877, Mar. 2022.

[151] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu, "Online human action detection using joint classification-regression recurrent neural networks," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 203–220.

[152] S. Ma, L. Sigal, and S. Sclaroff, "Learning activity progression in LSTMs for activity detection and early detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1942–1950.

[153] B. Ni, X. Yang, and S. Gao, "Progressively parsing interactional objects for fine grained action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1020–1028.

[154] 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 *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 461–470.

[155] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.

[156] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–11.

[157] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (IndRNN): Building a longer and deeper RNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5457–5466.

[158] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1012–1020.

[159] J. Liu, G. Wang, L. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention LSTM networks," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, Apr. 2018.

[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 *Proc. 30th AAAI Conf. Artif. Intell.*, D. Schuurmans and M. P. Wellman, Eds. Phoenix, AZ, USA: AAAI Press, Feb. 2016, pp. 3697–3704.

[161] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1227–1236.

[162] W. Zheng, L. Li, Z. Zhang, Y. Huang, and L. Wang, "Relational network for skeleton-based action recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 826–831.

[163] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12018–12027.

[164] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 579–583.

[165] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4570–4579.

[166] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 597–600.

[167] Z. Ding, P. Wang, P. O. Ogunbona, and W. Li, "Investigation of different skeleton features for CNN-based 3D action recognition," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 617–622.

[168] 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 *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2019, pp. 1–8.

[169] C. Caetano, F. Brémond, and W. R. Schwartz, "Skeleton image representation for 3D action recognition based on tree structure and reference joints," in *Proc. 32nd SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2019, pp. 16–23.

[170] H. Liu, J. Tu, and M. Liu, "Two-stream 3D convolutional neural network for skeleton-based action recognition," 2017, *arXiv:1705.08106*.

[171] J. Liu, N. Akhtar, and A. Mian, "Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition," in *Proc. CVPR Workshops*, 2019, pp. 10–19.

[172] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2959–2968.

[173] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Mar. 2005, pp. 729–734.

[174] R. Li, M. Tapaswi, R. Liao, J. Jia, R. Urtasun, and S. Fidler, "Situation recognition with graph neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4183–4192.

[175] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[176] A. Mazari and H. Sahbi, "MLGCN: Multi-Laplacian graph convolutional networks for human action recognition," in *Proc. 30th Brit. Mach. Vis. Conf. (BMVC)*. Cardiff, U.K.: BMVA Press, Sep. 2019, p. 281.

[177] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI), 30th Innov. Appl. Artif. Intell. (IAAI), 8th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, S. A. McIlraith and K. Q. Weinberger, Eds. New Orleans, LA, USA: AAAI Press, Feb. 2018, pp. 7444–7452.

[178] K. Shiraki, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Spatial temporal attention graph convolutional networks with mechanics-stream for skeleton-based action recognition," in *Proc. 15th Asian Conf. Comput. Vis. (ACCV)*, in Lecture Notes in Computer Science, vol. 12626, H. Ishikawa, C. Liu, T. Pajdla, and J. Shi, Eds. Kyoto, Japan: Springer, Dec. 2020, pp. 341–357.

[179] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Trans. Image Process.*, vol. 29, pp. 9532–9545, 2020.

[180] J. Huang, Z. Huang, X. Xiang, X. Gong, and B. Zhang, "Long-short graph memory network for skeleton-based action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 634–641.

[181] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 180–189.

[182] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1625–1633.

[183] K. Thakkar and P. J. Narayanan, "Part-based graph convolutional network for action recognition," 2018, *arXiv:1809.04983*.

[184] B. Li, X. Li, Z. Zhang, and F. Wu, "Spatio-temporal graph routing for skeleton-based action recognition," in *Proc. 33rd AAAI Conf. Artif. Intell. (AAAI), 31st Innov. Appl. Artif. Intell. Conf. (IAAI), 9th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*. Honolulu, HI, USA: AAAI Press, Jan. 2019, pp. 8561–8568.

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

[186] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.

[187] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

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

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

[190] A. Li, M. Thotakuri, D. A. Ross, J. Carreira, A. Vostrikov, and A. Zisserman, "The AVA-Kinetics localized human actions video dataset," 2020, *arXiv:2005.00214*.

[191] L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu, and A. Zisserman, "A short note on the kinetics-700-2020 human action dataset," 2020, *arXiv:2010.10864*.

[192] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 961–970.

[193] 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," 2016, *arXiv:1609.08675*.

[194] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, "HACS: Human action clips and segments dataset for recognition and temporal localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8667–8677.

[195] A. Diba, M. Fayyaz, V. Sharma, M. Paluri, J. Gall, R. Stiefelhagen, and L. Van Gool, "Large scale holistic video understanding," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, pp. 593–610.

[196] A. Piergiovanni and M. Ryoo, "Avid dataset: Anonymized videos from diverse countries," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 16711–16721.

[197] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick, and A. Oliva, "Moments in time dataset: One million videos for event understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 502–508, Feb. 2020.

[198] C. MoCap, "Carnegie Mellon University graphics lab motion capture database," Apr. 30, 2007. [Online]. Available: http://mocap.cs.cmu

[199] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation mocap database HDM05," Institut für Informatik II, Univ. Bonn, Bonn, Germany, Tech. Rep. CG-2007-2, Jun. 2007.

[200] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 9–14.

[201] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from RGBD images," in *Proc. Plan, Activity, Intent Recognit., Papers AAAI Workshop*, San Francisco, CA, USA, Aug. 2011, pp. 1–8.

[202] L. Xia, C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 20–27.

[203] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *Int. J. Robot. Res.*, vol. 32, no. 8, pp. 951–970, Jul. 2013.

[204] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*. Zurich, Switzerland: Springer, Sep. 2014, pp. 742–757.

[205] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[206] J. Hu, W. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5344–5352.

[207] B. Parsa, E. U. Samani, R. Hendrix, C. Devine, S. M. Singh, S. Devasia, and A. G. Banerjee, "Toward ergonomic risk prediction via segmentation of indoor object manipulation actions using spatiotemporal convolutional networks," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 3153–3160, Oct. 2019.

[208] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[209] W. Lin, H. Liu, S. Liu, Y. Li, R. Qian, T. Wang, N. Xu, H. Xiong, G.-J. Qi, and N. Sebe, "Human in events: A large-scale benchmark for human-centric video analysis in complex events," 2020, *arXiv:2005.04490*.

[210] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel(R) realsense(TM) stereoscopic depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Aug. 2017, pp. 1–11.

[211] M.-A. Drouin and L. Seoud, "Consumer-grade RGB-D cameras," in *3D Imaging, Analysis and Applications*. Cham, Switzerland: Springer, 2020, pp. 215–264.

[212] A. Grunnet-Jepsen, J. N. Sweetser, and J. Woodfill, "Best-known-methods for tuning Intel realsense D400 depth cameras for best performance," Intel Corp., Satan Clara, CA, USA, Tech. Rep. 1, 2018.

[213] A. Zabatani, V. Surazhsky, E. Sperling, S. B. Moshe, O. Menashe, D. H. Silver, Z. Karni, A. M. Bronstein, M. M. Bronstein, and R. Kimmel, ''Intel RealSense SR300 coded light depth camera,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2333–2345, Oct. 2020.
[214] X. Yang and Y. Tian, ''Effective 3D action recognition using EigenJoints,'' *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 2–11, Jan. 2014.
[215] H. Pazhoumand-Dar, C.-P. Lam, and M. Masek, ''Joint movement similarities for robust 3D action recognition using skeletal data,'' *J. Vis. Commun. Image Represent.*, vol. 30, pp. 10–21, Jul. 2015.
[216] T. Li, R. Zhang, and Q. Li, ''Multi scale temporal graph networks for skeleton-based action recognition,'' 2020, *arXiv:2012.02970*.
[217] B. Parsa, A. Narayanan, and B. Dariush, ''Spatio-temporal pyramid graph convolutions for human action recognition and postural assessment,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1069–1079.
[218] G. Zhu, L. Zhang, H. Li, and P. Shen, ''Topology-learnable graph convolution for skeleton-based action recognition,'' *Pattern Recognit. Lett.*, vol. 135, pp. 286–292, Jul. 2020.
[219] 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,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3316–3333, Jun. 2022.
[220] Y. Li, Y. Li, and N. Vasconcelos, ''Resound: Towards action recognition without representation bias,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 513–528.
[221] G. Kang, X. Dong, L. Zheng, and Y. Yang, ''PatchShuffle regularization,'' 2017, *arXiv:1707.07103*.
[222] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, ''Random erasing data augmentation,'' in *Proc. 34th AAAI Conf. Artif. Intell. (AAAI), 32nd Innov. Appl. Artif. Intell. Conf. (IAAI), 10th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*. New York, NY, USA: AAAI Press, Feb. 2020, pp. 13001–13008.
[223] T. DeVries and G. W. Taylor, ''Dataset augmentation in feature space,'' 2017, *arXiv:1702.05538*.
[224] S. Li, Y. Chen, Y. Peng, and L. Bai, ''Learning more robust features with adversarial training,'' 2018, *arXiv:1804.07757*.
[225] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert, ''GAN augmentation: Augmenting training data using generative adversarial networks,'' 2018, *arXiv:1810.10863*.
[226] E. Real, S. Moore, A. Selle, S. Saxena, Y. L. Suematsu, J. Tan, Q. V. Le, and A. Kurakin, ''Large-scale evolution of image classifiers,'' in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2902–2911.
[227] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, ''Regularized evolution for image classifier architecture search,'' in *Proc. 33rd AAAI Conf. Artif. Intell. (AAAI), 31st Innov. Appl. Artif. Intell. Conf. (IAAI), 9th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*. Honolulu, HI, USA: AAAI Press, Jan. 2019, pp. 4780–4789.
[228] D. Sun, D. Vlasic, C. Herrmann, V. Jampani, M. Krainin, H. Chang, R. Zabih, W. T. Freeman, and C. Liu, ''AutoFlow: Learning a better training set for optical flow,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10088–10097.
[229] S. Yun, S. J. Oh, B. Heo, D. Han, and J. Kim, ''VideoMix: Rethinking data augmentation for video classification,'' 2020, *arXiv:2012.03457*.
[230] Y. Zou, J. Choi, Q. Wang, and J.-B. Huang, ''Learning representational invariances for data-efficient action recognition,'' *Comput. Vis. Image Understand.*, vol. 227, Jan. 2023, Art. no. 103597.
[231] Y. Zhang, G. Jia, L. Chen, M. Zhang, and J. Yong, ''Self-paced video data augmentation by generative adversarial networks with insufficient samples,'' in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1652–1660.
[232] S. N. Gowda, M. Rohrbach, F. Keller, and L. Sevilla-Lara, ''Learn2Augment: Learning to composite videos for data augmentation in action recognition,'' in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*. Tel Aviv, Israel: Springer, Oct. 2022, pp. 242–259.
[233] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, ''Multi-modal transformer for video retrieval,'' in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, pp. 214–229.
[234] A. J. Piergiovanni and M. S. Ryoo, ''Learning multimodal representations for unseen activities,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 506–515.
[235] L. Zhu and Y. Yang, ''ActBERT: Learning global-local video-text representations,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8743–8752.
[236] A. Rouditchenko, A. Boggust, D. Harwath, B. Chen, D. Joshi, S. Thomas, K. Audhkhasi, H. Kuehne, R. Panda, R. Feris, B. Kingsbury, M. Picheny, A. Torralba, and J. Glass, ''AVLnet: Learning audio-visual language representations from instructional videos,'' 2020, *arXiv:2006.09199*.
[237] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman, ''Self-supervised multimodal versatile networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 25–37.
[238] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, ''Accurate, large minibatch SGD: Training ImageNet in 1 hour,'' 2017, *arXiv:1706.02677*.
[239] J. Lin, C. Gan, and S. Han, ''Training kinetics in 15 minutes: Large-scale distributed training on videos,'' 2019, *arXiv:1910.00932*.
[240] A. Howard, M. Sandler, B. Chen, W. Wang, L. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, ''Searching for MobileNetV3,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.
[241] X. Zhang, X. Zhou, M. Lin, and J. Sun, ''ShuffleNet: An extremely efficient convolutional neural network for mobile devices,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

CAILING WANG received the B.S. degree from Tianjin University, Tianjin, China, in 2006, and the Ph.D. degree in signals and information processing from the Xi'an Institute of Optics and Precision Mechanics, China Academy of Sciences, Xi'an, Shaanxi, in 2011. She is currently an Associate Professor with the School of Computer Science, Xi'an Shiyou University. Her major research interests include remote sensing image processing and artificial intelligence.

JINGJING YAN received the B.S. degree in computer science and technology from Tianjin Normal University, Tianjin, China, in 2021. She is currently pursuing the M.S. degree in electronics and information technology with the School of Computer Science, Xi'an Shiyou University, Xi'an, Shaanxi, China. Her research interests include computer vision and human action recognition.