Neurocomputing 537 (2023) 164–186

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Transformer for Skeleton-based action recognition: A review of recent advances ☆

Wentian Xin [a,b], Ruyi Liu [a,b], Yi Liu [a,b], Yu Chen [a,b], Wenxin Yu [a,b], Qiguang Miao [a,b,*]

[a] Xi'an Key Laboratory of Big Data and Intelligent Vision, Xi'an 710071, Shaanxi, China
[b] School of Computer Science and Technology, Xidian University, 2 Taibainan Road, Xi'an 710071, Shaanxi, China

ARTICLE INFO

Article history:
Received 15 December 2022
Revised 27 January 2023
Accepted 4 March 2023
Available online 30 March 2023

Keywords:
Transformer
Graph convolution network
Skeleton-based action recognition
Spatial temporal structure
Survey

ABSTRACT

Skeleton-based action recognition has rapidly become one of the most popular and essential research topics in computer vision. The task is to analyze the characteristics of human joints and accurately classify their behaviors through deep learning technology. Skeleton provides numerous unique advantages over other data modalities, such as robustness, compactness, noise immunity, etc. In particular, the skeleton modality is extremely lightweight, which is especially beneficial for deep learning research in low-resource environments. Due to the non-European nature of skeleton data, Graph Convolution Network (GCN) has become mainstream in the past few years, leveraging the benefits of processing topological information. However, with the explosive development of transformer methods in natural language processing and computer vision, many works have applied transformer into the field of skeleton action recognition, breaking the accuracy monopoly of GCN. Therefore, we conduct a survey using transformer method for skeleton-based action recognition, forming of a taxonomy on existing works. This paper gives a comprehensive overview of the recent transformer techniques for skeleton action recognition, proposes a taxonomy of transformer-style techniques for action recognition, conducts a detailed study on benchmark datasets, compares the algorithm accuracy of standard methods, and finally discusses the future research directions and trends. To the best of our knowledge, this study is the first to describe skeleton-based action recognition techniques in the style of transformers and to suggest novel recognition taxonomies in a review. We are confident that Transformer-based action recognition technology will become mainstream in the near future, so this survey aims to help researchers systematically learn core tasks, select appropriate datasets, understand current challenges, and select promising future directions.

1. Introduction

Human action recognition covers an extremely large number of research topics in computer vision and has a wide range of applications in visual surveillance [1–4], industrial control [5–8], autonomous driving[9–12], intelligent transportation [13–16] and human–computer interaction [17–20]. To take full advantage of multi-source data and analyze the action recognition problem from multiple perspectives, researchers have used visual appearance (RGB) [21–24], depth information [25–28], optical flow [21,29–31], and even sound [32,33] to assist the recognition task. However, extracting action representation or mapping features directly from video sequences ignores the inherent interaction between human joints, which makes the algorithm vulnerable to the interference of environmental background and natural light [34]. Based on RGB images, 2D & 3D skeleton data are readily available, which can be generated by pose estimation methods [35,36] or acquired by devices such as Microsoft Kinetics [37,38]. Skeleton data provide detailed position and motion information of human joints, which facilitates the construction of spatio-temporal and motion features [39]. The dataset composed of skeletons helps the algorithm to focus on the essential characteristics of the action behavior, avoid the interference of the background, and effectively improve the accuracy of action recognition in complex environments.

The core of human action recognition is to obtain discriminative feature representation [34]. Different from the RGB representation, the human action in a video not only describes the appearance in

☆ The work was jointly supported by the National Natural Science Foundations of China (No. 62272364, 61902296, 62002271), the Province Key R&D Program of Shaanxi (No. 2020LSFP3-15), Guangxi Key Laboratory of Trusted Software (No. KX202061), the Key R&D Projects of Qingdao Science and Technology Plan (No. 21-1-2-18-xx), Key Project of the 14th Five-Year Plan for Adult Continuing Education Research Program of China Adult Education Association (No. 2021-414ZA), Shaanxi Higher Continuing Education Teaching Reform Research Project (No. 21XJZ004).
* Corresponding author

https://doi.org/10.1016/j.neucom.2023.03.001
0925-2312/

space but also extracts the dynamic changes of posture. With the application of deep learning in image classification and object detection, many works [40–42] apply popular algorithms to human action recognition to study the action semantics of the learning context from video. In the field of action recognition based on human skeleton, from the overall human body abstraction to the skeleton hierarchy, the focus is to deeply explore the construction and representation of data structure. Combined with advanced artificial intelligence algorithms such as deep learning and reinforcement learning, the performance is greatly improved [43].

Observing the overall development process, skeleton-based action recognition methods can be summarized into two stages: traditional machine learning and deep learning [44]. In the first stage, the features of actions need to be manually designed by experts based on prior knowledge [45]. These designs include motion trajectories, viewpoints, geometric shapes, etc. The inflexibility of manual rules becomes an obstacle to the further development of such methods. The deep learning methods break through the limitations, and researchers quickly move the stage of traditional machine learning into the stage of deep learning [46]. In the second stage, four categories adopting different deep learning frameworks are divided: RNN-based methods [47], CNN-based methods [48], GCN-based methods [39], and Transformer-based methods [49]. A brief demonstration of these methods is provided on the left of Fig. 1.

There have been several reviews that describe the existing skeleton-based action recognition methods. Ahmad et al. [50] outline the skeleton action recognition algorithms based on graph convolution and divide the action recognition into five subcategories, which describe the corresponding algorithm network model, mathematical principles, contributions, and limitations. However, the review lacks the introduction of transformer, which is not conducive to helping researchers form a macro-comprehensive combing of skeleton action recognition methods. Feng et al. [51] expound and analyze action recognition based on GCN. Five representative baseline datasets are selected to evaluate the performance of several kinds of iconic skeleton-based GCN methods for human action recognition. The shortcomings are that the classification of skeleton action recognition methods based on GCN is not clear, and the introduction of transformer is also lacking. Yue et al. [52] introduce the main models and algorithms of traditional manual method, RGB-based end-to-end action recognition method, and skeleton-based action recognition method. The review lacks explicit classification of recent popular transformer methods, which cannot form a comprehensive understanding of transformer. Ulhaq et al. [53] introduce the generation and development of transformer as well as the rise of transformer in the field of computer vision, which analyzes the application of transformer in the four basic action recognition tasks, and two data modes of single mode and multi-mode data. However, there is no explicit classification study on the skeleton action recognition methods based on transformer. Xing et al. [54] summarize practices related to 3D skeleton action recognition based on deep learning methods, including relevant algorithms in RNN-based, CNN-based, and GCN-based technologies, and compare the performance of existing motion recognition methods. The survey discusses improvements and related structures of several classic algorithm models, but the refined classification is not carried out. Simultaneously, classification and introduction of methods based on transformer are lacking. Qing et al. [55] describe the existing works on skeleton-based action recognition and categorizes them from the perspective of spatial feature extraction, temporal pattern capture, and

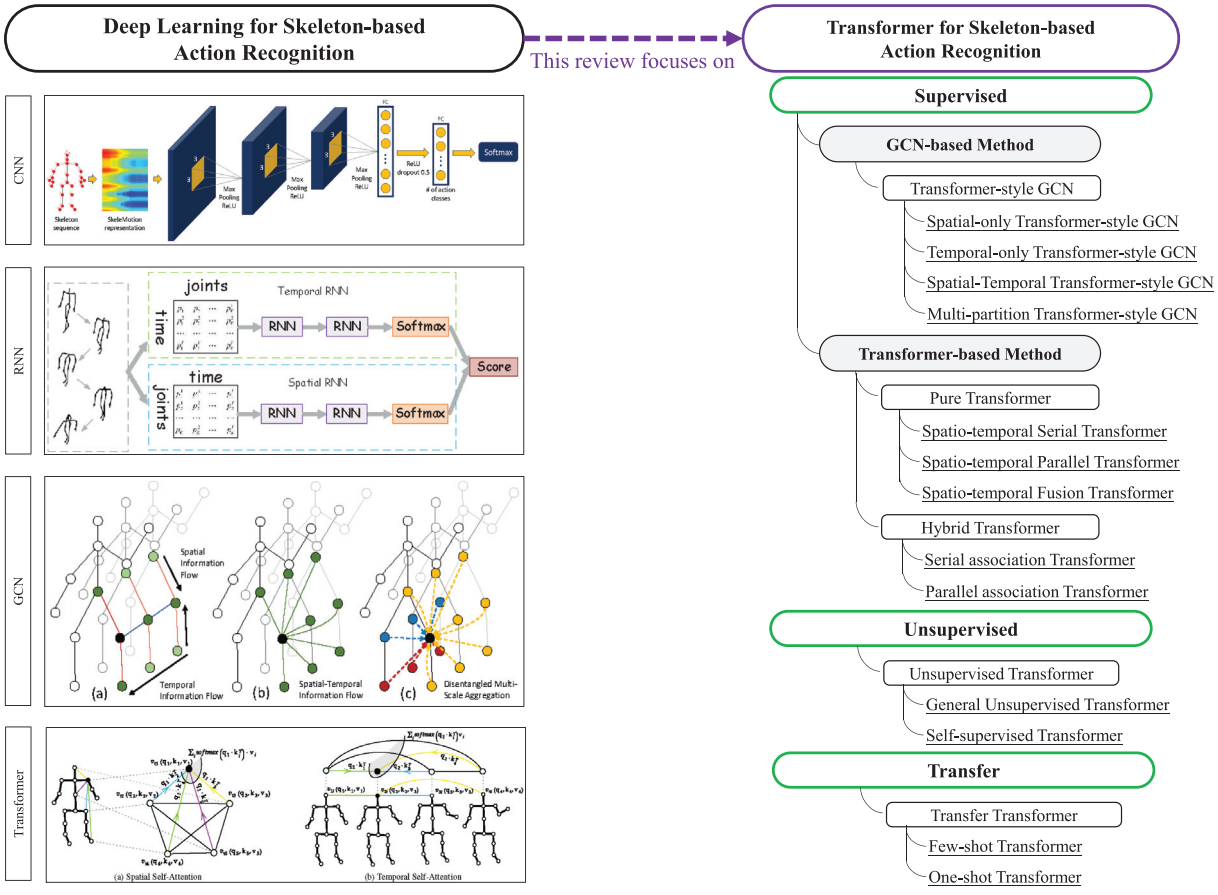


Fig. 1. Classification of Transformer for skeleton-based action recognition methods.

improved signal input quality. A large-scale human skeleton dataset (ANUBIS) is collected to facilitate a fair and comprehensive evaluation of existing methods. However, the review also lacks categorization and a comprehensive introduction to transformer-based approaches.

In summary, the reviews mentioned above lack exploration and discussion on the taxonomy of transformer-based skeleton action recognition. As an epoch-making algorithm, transformer is emerging explosively in the field of point-centric research [56]. Both in terms of computational cost and accuracy, transformer-based algorithms are gradually breaking the monopoly of GCN methods. Through analysis, we are convinced that Transformer-based methods still have strong potential and will become mainstream in the future.

We first briefly review GCN-based methods, which can effectively deal with the interaction of topological graph networks, but rely on manually designed or self-learning weighted edges for global information acquisition. We could divide the innovation directions of existing representative GCN-based methods into three types: (1) Adjacency matrix modeling [57–64], (2) Convolution strategy [39,65–70], and (3) Data reconstruction [71–77]. Through extensive exploration and analysis, we find that the adjacency matrix based on the GCN is constructed in a similar way as the transformer-based self-attention matrix in the spatial domain (see Fig. 3 for details). Considering the close association with the self-attention mechanism of transformer, we segment a subset of the GCN-based algorithms that adopt the self-attention module and classify them as Transformer-style GCN.

Compared with GCN-based methods, Transformer [78,79] can quickly obtain global topology information and increase the correlation strength of non-physical joints through network updates, but it is weak in extracting discriminative information from local features and short-term temporal information [49]. So after the initial introduction of Pure Transformer into skeleton action recognition, GCN and CNN as complements make up for the defects of the basic Transformer, called Hybrid Transformer. In particular, Transformer-style GCN and incorporating GCN into transformer are distinct. The Transformer-style GCN focuses on constructing the adjacency matrix through the self-attention mechanism, while the latter uses GCN as a supplement to improve the ability of local feature extraction, and both the infrastructure and tricks follow the rules of standard transformers, such as Positional Embedding and Multi-Heads Attention. In addition, we also introduce the datasets of existing transformer-based skeleton action recognition methods and compare the accuracies. For the latest UAV-Human dataset, according to [60], we provide a new data preprocessing method that can be easily replaced and used in the source code. Please refer to the following link for details: https://github.com/back330/UVA-Human-Skeleton-Preprocessing.

The above two parts are mainly about supervised learning. However, most existing supervised human skeleton action recognition methods have three serious drawbacks [80]: (1) Relying on large amounts of labeled data. (2) Highly sensitive to noise. (3) Difficulty in recognizing actions in complex scenes. Since learning on unlabeled data can improve the generalization ability of features, un-/semi-supervised and self-supervised human skeleton-based action recognition have attracted more attention. Unsupervised (including semi-/self-supervised) methods are beneficial to alleviate the overfitting problem and reduce the need for a massive amount of labeled data. Unsupervised action recognition learns the global context of the whole action sequence to infer the motion details of local joints and the overall topology [81]. In skeleton-based action recognition tasks, unsupervised and semi-supervised methods mostly utilize encoder-decoder models [82–85], contrastive learning [86–89], and clustering strategies [90,91]. Self-supervised methods mostly utilize contrastive learning models [92–107]. However, the performance of these algorithms is highly dependent on the choice of encoder model, so it is challenging to design networks that can accommodate global context information and simultaneously focus on detailed joint features. In the past two years, Transformer has been applied to unsupervised [108,109] and self-supervised [110–113] tasks for skeleton-based action recognition, showing excellent performance in capturing global context and local joint dynamics.

Recently, with the increase of the number of categories in action recognition, transfer learning has been widely studied. Transfer learning can apply knowledge learned in one domain or task to a different but related domain. Currently, skeleton-based zero-shot action recognition (ZSAR) [114–118], few-shot action recognition (FSAR) [119–122], and one-shot action recognition (OSAR) [123–125] are being deeply explored. It can be applied in environments with no significant amount of labeled data, greatly reducing the cost of data resources. So far, skeleton-based FSAR, ZSAR, and OSAR are implemented by algorithms such as clustering [119,114], generation [123,115], prototype mapping [125,116], matching alignment and augmentation [120,121], as well as specific tasks oriented to temporal action segmentation [122] and classroom action recognition [118]. Compared with other networks, transformer utilizes the global spatial and temporal correlation representation, which is more conducive to extracting class-specific prototypes and maintaining latent space consistency. At the same time, transfer learning based on transformer network has the ability to deal with human pose occlusion and transfer to new scarce data action classes [126,127].

In this paper, we summarize recent research on human action recognition using transformers. For the literature survey and data collection, strictly follow:

(1) This paper is a serious, detailed, and realistic introduction to transformer-based skeleton action recognition.
(2) The cited papers are selected from journals and conferences, including the latest published papers.
(3) The experiments and results have been verified by enough published papers and the corresponding methods have been properly explained.

Research contributions. Human skeleton action recognition has attracted much attention in computer vision and artificial intelligence applications in recent years. In particular, the Transformer has been the most popular basic feature extraction structure in the past five years. However, to the best of our knowledge, there is no comprehensive introduction and comparison of recent transformer-based skeleton action recognition methods in any literature. The main contributions of this paper are as follows:

1. *Overview:* This paper presents a detailed and comprehensive review of transformer-based skeleton action recognition algorithms. In addition, the corresponding datasets and accuracies are introduced, evaluated, and compared to purposefully discuss the data modality, network structure, weight updates, and information changes. Finally, the challenges faced by the current skeleton-based action recognition methods are discussed, and the future development directions are prospected.
2. *Taxonomy:* We classify Transformer-based Skeleton Action Recognition into three main categories: Supervised learning, Unsupervised learning, and Transfer learning. Further, all transformer-based methods are classified into five sub-classes.
   (a) *(Supervised) Pure Transformer:* The standard transformers simultaneously extract joint interaction information in space and inter-frame feature interaction information in time.

(1) *Spatio-temporal Serial Transformer:* The spatial transformer and temporal transformer are used alternately, and the two modules are arranged in serial in the network.
(2) *Spatio-temporal Parallel Transformer:* The spatial transformer and temporal transformer are used together, and the two modules are arranged in parallel in the network.
(3) *Spatio-temporal Fusion Transformer:* The spatial transformer and temporal transformer are integrated into one module, and there is no obvious division between the two transformers in the network.

(b) *(Supervised) Hybrid Transformer:* The Transformer, GCN, and CNN are mixed as the complementary method for feature extraction.
(1) *Serial Association Transformer:* GCN, CNN, spatial transformer, and temporal transformer are used alternately, and each module in the network is arranged in series.
(2) *Parallel Association Transformer:* GCN, CNN, spatial transformer, and temporal transformer are used together, and each module in the network is arranged in parallel.

(c) *(Supervised) Transformer-style GCN:* The correlation matrix is generated through the self-attention mechanism or the construction of interactive matrix strategy.
(1) *Spatial-only Transformer-style GCN:* The transformer strategy is used only on the adjacency matrix construction in the spatial domain.
(2) *Temporal-only Transformer-style GCN:* The transformer strategy is used only on the correlation matrix construction in the temporal domain.
(3) *Spatial–Temporal Transformer-style GCN:* The transformer strategy is adopted in the construction of the relationship matrix in the temporal and spatial domain.
(4) *Multi-partition Transformer-style GCN:* The transformer strategy is applied to build multi-level or multi-modal relationship matrices.

(d) *(Unsupervised) Unsupervised Transformer:* The transformer model is used for unsupervised skeleton action learning.
(1) *General Unsupervised Transformer:* The transformer is applied to capture global context information.
(2) *Self-supervised Transformer:* The transformer is applied to obtain spatio-temporal global abstract information.

(e) *(Transfer) Transfer Transformer:* The transformer model is used for transfer skeleton action learning.
(1) *Few-shot Transformer:* The transformer is applied to improve the global spatio-temporal position and index alignment ability.
(2) *One-shot Transformer:* The transformer is applied to alleviate the universal occlusion interference that often exists in the real world.

3. *Datasets:* This paper introduces the datasets involved in the above classification methods, which include multiple newly proposed action recognition datasets, appearing in the review papers for the first time.
4. *Comparisons:* This paper collects and compares the accuracies of all transformer-based skeleton action recognition algorithms.
5. *Challenges and Future Directions:* We analyze the limitations of the existing methods and propose possible future research directions and innovative ways to provide a reference for more researchers.

The structure of the article is arranged as follows. In Section 2, the research background is presented. In Section 3, the proposed taxonomy and the advantages and disadvantages of different methods are presented. In Section 4 and Section 5, the dataset and accuracy comparisons are presented. In Section 6, necessary remarks and future directions of research are analyzed. Finally, Section 7 concludes the paper.

2. Preliminaries

In this section, we first give a brief background on GCN-based and Transformer-based skeleton action recognition and then analyze the similarity of their construction of the relationship matrix.

2.1. Transformer on Skeleton data

In computer vision, the rise of transformer is mainly due to the paper VIT [79,128], which successfully applies the natural language processing encoder module in [78] to image encoding. As the extractor of features, the application conditions of CNN and transformer are the same in RGB image encoding. As a different data modality, CNN can also be naturally replaced by transformer in the processing of skeleton data. The CNN-based skeleton action recognition technology maps the skeleton into RGB images and then processes them by the corresponding image classification algorithm. However, as standard non-Euclidean spatial data, the skeleton mapping method will lose important spatial correlation information. Positional embedding, which is widely applied in transformers, alleviates the problem of reduced discrimination caused by the loss of spatial correlation [129]. On the other hand, the number of joints is thousands of times less than the number of image pixels, and transformer can be extracted and applied naturally without resorting to patch partitioning. Furthermore, transformer also has advantages in the acquisition and processing ability of global information, which is exactly the key point in improving the classification ability of skeleton data.

Providing a standard transformer implementation makes it easier for researchers to transfer ideas to skeleton tasks. As the encoder of VIT [79] in Fig. 2 shows, the architecture of existing articles is generally derived from the standard transformer module. Multi-Head Attention (MHA) is an essential part of Encoder, and Self-Attention is the core of MHA. In addition, the Multi-Layer Perceptron (MLP), Normalization (Norm), Skip-connections, and addition operation are also important components in optimizing the transformer architecture. Firstly, the input joints are weighted with additional information by position embedding to retain special spatial discrimination. The mainstream method in skeleton processing employs the absolute position encoding module to label each joint, and adopts sine and cosine functions with different frequencies as the encoding functions, as:

$$\begin{aligned} PE(p,2i) &= sin\left(p/10000^{2i/C_{in}}\right), \\ PE(p,2i+1) &= cos\left(p/10000^{2i/C_{in}}\right), \end{aligned} \tag{1}$$

where p and i denote the position of the joint and the dimension of the position encoding vector, respectively, and then the calculated PE vector is added to the original vector. In this way, all joints are assigned different IDs. Some methods apply tuples or topological grouping, but the core operation of PE is the same, and the position encoding is performed on different regions after the change. Next,

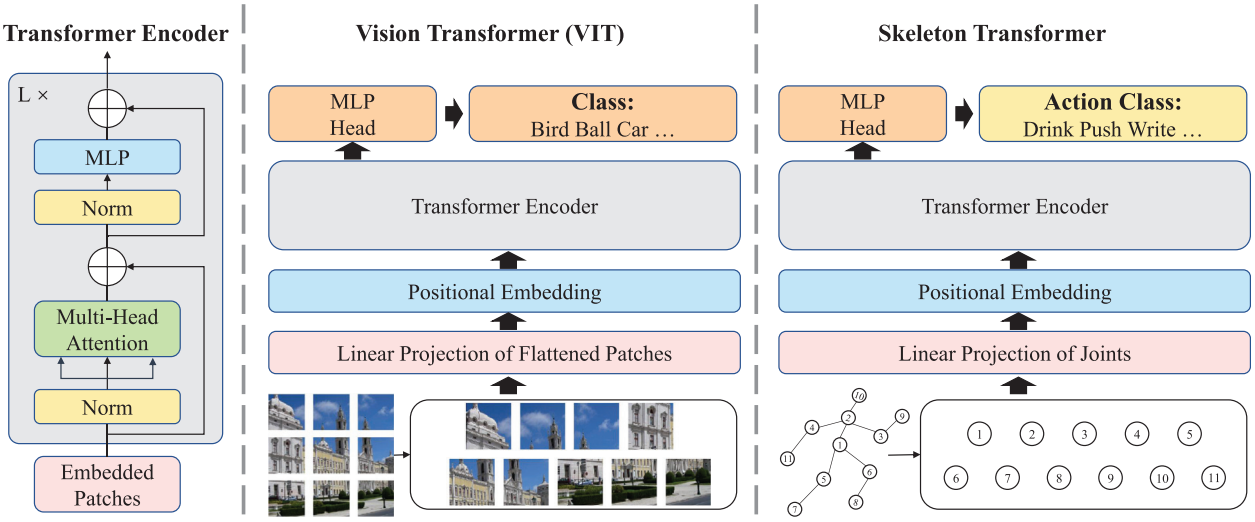


Fig. 2. Illustrating the comparison between Vision Transformer and Skeleton Transformer in implementation.

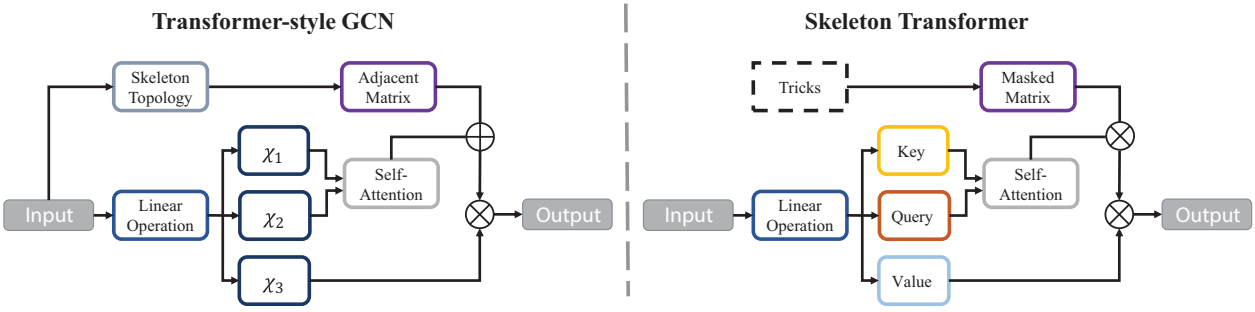


Fig. 3. Illustrating the comparison between Transformer-style GCN and Skeleton Transformer in implementation.

the vectors after PE are projected into the embedding subspace through the mapping operation, which is similar to the word vector embeddings of NLP and the patch embeddings of VIT. In the specific encoding process, some methods adopt class tokens as BERT [130], and splice them to the starting position of the feature vector. More methods adopt the strategy in [78], which directly concatenates and uses GAP for dimensionality reduction and then maps to the number of classes by FC. We have not found any special designs of multi-head, most of the papers that adopt the method of [78] and will not be specifically described anymore. In the papers on transformer-based skeleton action recognition, the self-attention module usually behaves in the following style:

$$Attention(Q,K,V) = V \cdot softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{2}$$

where Q, K, and V denote the *query*, *key*, and *value* of vectors, respectively, and this step represents the context relation computation and update of the overall skeleton data. In practice, most methods take the input and use $Conv_{1\times1}$ to map the number of channels to three times the original dimension, and then crop it back to the original channel size. If the object manipulated by the input vector is transformed from joint to time, the matrix computed by the self-attention mechanism represents the frame-to-frame contextual association. In particular, if the operands are joints, d_k generally computes the product of channels and T; if the operating object is time, then d_k generally computes the product of channel and V. Finally, additional $Conv_{1\times1}$ will be used as the adjustment of the number of channels and the secondary fusion of features between channels. In practice, ST-GCN [39] treats space and time in turn and most GCN-based methods adopt this strategy and achieve excellent performance. In addition, short-term features play an important role in action recognition, and the standard transformer is difficult to balance between global and local information in the temporal domain. Since the imbalanced performance is also reflected in the spatial domain, GCN or CNN are introduced as an additional supplement for feature extraction in subsequent work. Information fluidity in time and space is the characteristic of skeleton data, so we perform a secondary classification based on spatio-temporal association relationships. By looking for the commonality and individuality in the information interaction process of all algorithms, the structure of our survey is more objective and clear, which is fundamentally different from the existing classification according to the operation of algorithms.

2.2. Transformer-style GCN on Skeleton data

In this paper, we study the similarity between adjacency matrices in GCNs and self-attention matrices in Transformers by exploring their respective implementation methods. First, we need a brief introduction to GCN. Given the body joint sequence in 2D or 3D coordinates, the skeleton of the human body can be denoted as $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, where $\mathscr{V} = (v_1, v_2, \ldots, v_N)$ represents the joint set of N vertices, $\mathscr{E}$ represents the bone set of the edges. In the adjacency

matrix $A \in \mathbb{R}^{N\times N}$ (assuming $\mathcal{G}$ is an undirected graph), if the V_i and V_j have a skeleton directly connected, $A_{i,j} = 1$, otherwise, $A_{i,j} = 0$. In the action sequence of the graph, the joint features set is denoted as $\mathcal{X} = \{x_{t,n} \in \mathbb{R}^C | t, n \in \mathbb{Z}, 1 \leqslant t \leqslant T, 1 \leqslant n \leqslant N\}$, and the input can be denoted as $X \in \mathbb{R}^{T\times N\times C}$, where $x_{t,n}$ represents the C-dimensional feature vector of vertex v_n at time t in T frames. If the skeleton sequence is represented by X and A, the layer-wise iteration and weights update can be formulated as:

$$X^{l+1} = \sigma\left(\tilde{\Lambda}^{-\frac{1}{2}}\tilde{A}\tilde{\Lambda}^{-\frac{1}{2}}X^l W^l\right), \tag{3}$$

where $\tilde{A} = A + I$ is the skeleton graph with added self-loops to keep identity features, $\tilde{\Lambda}$ is the diagonal degree matrix of $\tilde{A}$, and $\sigma(\cdot)$ is an activation function. $W^l \in \mathbb{R}^{C_l\times C_{l+1}}$ represents the learnable matrix of the network at layer l. Graph convolutional networks can be equivalent to compute the average feature aggregation of the joints.

In development, a series of algorithms represented by 2s-AGCN [57] employ self-attention methods to build data-adaptive adjacency matrices:

$$A_{atten} = softmax\left(f_{in}^{\mathrm{T}} W_\theta^{\mathrm{T}} W_\theta f_{in}\right). \tag{4}$$

where W_θ and $f(\cdot)$ represent learnable matrix and mapping operations, respectively. In Fig. 3, it is clear that these methods are similar to the transformer construction. The difference lies in how the original vector is processed to generate Q and K. The fully parameterized adjacency matrix A_m in GCN is consistent with the construction method of Attention Masks in Transformer. Both of them declare that the parameterized matrix with the same dimension as the original matrix is autonomously learned through network updates. Based on the above analysis and comparison of formulas, we believe that it is more appropriate to merge these GCN methods into Transformer methods. So in the taxonomy, Transformer-style GCN is added as the final supplement.

3. Taxonomy for transformer-based Skeleton action recongnition

Compared with modalities such as RGB and optical flow, the skeleton data is compact and has the advantage of high computational efficiency. In addition, the skeleton data are robust to illumination changes and background noise and invariant to camera viewpoint [44]. As an emerging feature extraction method, the transformer has made a lot of contributions in the field of skeleton action recognition. And PE has the ability to encode locations without losing spatial correlation information due to feature mapping. Moreover, the self-attention mechanism has a powerful ability to obtain global information, which is conducive to grasping the spatio-temporal feature changes of the skeleton from an overall perspective. Compared with GCN, Transformer facilitates the acquisition of distal joint association relations and does not require the assistance of additional learnable parameter edges. Given the strong potential of transformers in skeleton action recognition, we propose a new taxonomy—Transformer-based Skeleton Action Recognition Taxonomy.

The structure of this section is arranged as follows. Section 3.1 contains the definition, classification criteria, and method content of Pure Transformer. Section 3.2 contains the definition, classification criteria and subclassification rules of Hybrid Transformer. Section 3.3 contains the definition, classification criteria, and method content of Transformer-sytle GCN. Section 3.4 contains the definition, classification criteria, and method content of Unsupervised Transformer. Section 3.5 contains the definition, classification criteria, and method content of Transfer Transformer. In Section 3.6, we uniformly introduce the innovation points and the main problems solved by these methods.

3.1. Pure transformer for Skeleton-based action recognition

As shown in the Fig. 4, the module composition of Pure Transformer is mainly composed of: spatial transformer, temporal transformer and spatial–temporal fusion transformer. The Pure Transformer only uses transformer structure as the feature extractor in the temporal and spatial domain processing, and the standard Transformer extracts the joint information in space and the inter-frame feature interaction information in time. According to the arrangement of spatial and temporal transformers in the overall framework, all pure transformers are divided into three categories: spatio-temporal serial transformer, spatio-temporal parallel transformer, and spatio-temporal fusion transformer. In particular, the spatio-temporal order between modules is not considered for completeness of classification. In addition, we will introduce the strategies of these methods, such as multi-level, grouping, information assistance, etc. However, in contrast to the large differences in the spatio-temporal structure, we will not classify the above strategies separately because they are inclusion relations.

3.1.1. Spatio-temporal serial transformer

To fuse the human joint and part interactions simultaneously, Wang et al. [131] propose a novel spatial–temporal transformer network (IIP-Former), which can efficiently capture both joint-level (intra-part) dependencies and part-level (inter-part) dependencies. Compared with the standard Transformer network, there are three main improvements: (1) Intra-Inter-Part self-attention mechanism simultaneously captures intra-part features and inter-part features without increasing much computational complexity. (2) Such kind of method introduce a learnable class-token instead of pooling all features extracted by the backbone. (3) Instead of using two individual transformers to model spatial and temporal dependencies, IIP-Former reduces model size while increasing the generalization of the model. Apart from the relationship between body parts, the long-distance dependence of joints is also crucial. Yan et al. [132] propose MSST-RT and make the following improvements: (1) The spatial–temporal relative transformer (ST-RT) is formed based on a lightweight transformer called the relative transformer mechanism. In spatial dimension, the relative transformer mechanism is able to establish a relationship between

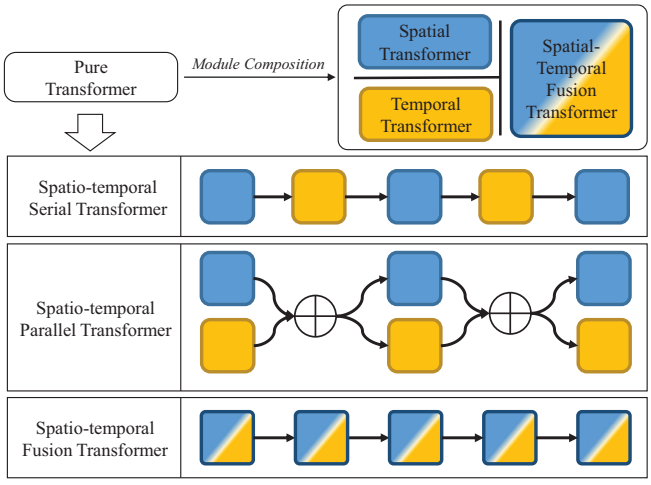


Fig. 4. Illustration of diagrams and ketches in Pure Transformer, which explains the basic architecture of the categories.

two joints that are far apart in the physical structure. In temporal dimension, the relative transformer mechanism can establish a relationship between two distant frames to achieve interaction. (2) The multi-stream spatial–temporal relative transformer (MSST-RT) is formed based on the ST-RT. Each pathway of the MSST-RT contains an ST-RT to extract the spatio-temporal features from four skeleton sequences respectively, and the extracted features are fused to improve the performance of behavior recognition. (3) An effective dynamic representation is produced, which can obtain more abundant information from the skeleton sequence in ST-RT by fusing the information of three different motion scales.

If the simple fusion model is taken as the baseline, the cross-view fusion model introduces the cross-view fusion mechanism based on the simple fusion model. Momal et al. [133] propose a multi-modal transformer-based network, which combines the characteristics of joints and acceleration: (1) Two kinds of single-modality models based on the transformer named Spatial Temporal Skeleton Model and Acceleration Model are proposed. (2) By combining two single-modality models, two dual-modality models named Simple Fusion Model and Cross-view Fusion Model are created, which are trained based on skeleton data and acceleration data. The decoupling of the skeleton is introduced to emphasize the specific characteristics of space/time and different motion scales, which provides a more comprehensive understanding of human actions. Shi et al. [134] represent DSTANet to improve the traditional recognition network in the following several aspects: (1) They propose a novel decoupled spatial–temporal attention network (DSTA-Net) for skeleton-based action recognition. Such kind of method is a unified framework based entirely on the attention mechanism, which allows modeling spatial–temporal dependencies between joints without the requirement of knowing their positions or mutual connections. (2) By introducing three techniques in building DSTA-Net to meet the specific requirements for skeletal data, including spatial–temporal attention decoupling, decoupled position encoding, and spatial global regularization.

3.1.2. Spatio-temporal parallel transformer

By adopting different joint organization strategies to model the skeleton sequence in spatial and temporal dimensions and effectively capture the motion of the skeleton, Zhang et al. [135] propose the Spatial–Temporal Specialized Transformer (STST). This method explicitly extract coordinate information, semantic information and temporal information into three kinds of tokens, in order to make full use of the skeleton data without losing information. In the encoder, the authors design the Spatial Transformer Block to model the posture represented by the skeleton in each frame separately, and the Directional Temporal Transformer Block to model the action based on the movements of the entire skeleton in the temporal dimension with direction-aware strategy. In addition, the two-stream structure leads to the extension of the feature dimension and enables the network to capture richer information. Zhang et al. [136] propose a two stream Transformer encoder network (TSTE) that includes motion spatio-temporal feature embedding and shape transformation. The inputs of the network are coordinate data and coordinate changes, the coordinate data contains joint space features, and the coordinate changes include motion time features. By extracting and merging the features of the two inputs, the network consider all spatial and temporal features as a whole to calculate correlation, thus allowing it to accept input data of various shapes without the need for adjacency matrix. Furthermore, prior to the Transformer encoder, a full connection layer known as the embedding layer is introduced to distribute features more densely.

Transformer often requires large computational resources, Shi et al. [137] design an efficient sparse transformer-based action recognition model—STAR. The STAR model greatly improved the traditional skeleton-based action recognition tasks in the following two aspects: (1) Sparse self-attention module is proposed to capture spatial correlations, which greatly reduces extra computation by performing sparse matrix multiplications. (2) STAR applies a segmented linear self-attention module to captures temporal correlations to further reduce the computation and memory usage by processing variable length of sequences. Segmented positional encoding is accordingly used to the data embedding to convey the idea of timeseries ordering along the temporal dimension of variable-length skeletal data.

3.1.3. Spatio-temporal fusion transformer

Semantic-based graph is learned by attention mechanism in a data-driven way to capture the semantic correlations of interactive body parts. Pang et al. [138] propose the Interaction Graph Transformer (IGFormer) contained GI-MSA module and SPM module: (1) The Graph Interaction Multi-head Self-Attention (GI-MSA) module is proposed to learn the relationship between persons at both semantic and distance levels. GI-MSA consists of semantic-based graph and distance based interaction graph. The distance based graph is constructed by measuring the distance between body parts to mine the distance information between interactive body parts. This innovation better represent the interrelationship between the body parts of the interactive personnel. (2) The Semantic Partition Module (SPM) converts each subject's skeleton sequence into the Body Part Time (BPT) sequence to facilitate effective modeling of interactive body parts. To balance feature representation for cross modal data, Ahn et al. [140] propose START. STAR transformer consists of encoder and decoder. The encoder called STAR transformer encoder contains two modules, namely full spatial temporary attention (FATTN) module and zigzag spatial temporary attention (ZATTN) module. The continuous decoder is composed of FATTN module and binary spatial temporary attention (BATTN) module. The STAR transformer learn the spatio-temporal features efficiently by arrange the pairs of FATTN, ZATTN and BATTN modules in a appropriate way. To solve the problem of combining cross-modal action data, the multi-class token is proposed to aggregate cross-modal data of spatio-temporal video and skeleton.

By exploiting 2D skeletal representations of short time sequences, Mazzia et al. [139] introduce a simple, fully self-attention architecture named Action Transformer (Ac-Former): (1) An architecture derived from the standard transformer encoder is proposed for the action recognition task, the encoder consists of alternating multi-head self-attention and feedforward blocks. (2) In addition, a real-time short-time HAR dataset named MPOSE2021 is introduced, which is suitable for pose-based and RGB-based methods. By continuously classifying actions within short past time steps, the dataset has shorter time steps than other publicly available datasets.

3.2. Hybrid transformer for skeleton-based action recognition

As shown in the Fig. 5, the modules of Hybrid Transformer are mainly composed of: spatial transformer, spatial graph convolution (Spatial GraFormer), temporal transformer, temporal convolution and temporal convolution & transformer (Temporal ConvFormer). Hybrid Transformer blends GCN and CNN in the standard Transformer as a complementary feature extraction method. The convergence of technologies helps give full play to the advantages of different base models and promotes the migration and application of Transformer on different widely used algorithms. We overcome the difficulty of distinguishing between various models. According to the arrangement of different models in the overall framework, all Hybrid Transformers are clearly divided into two main categories—serial association transformer,

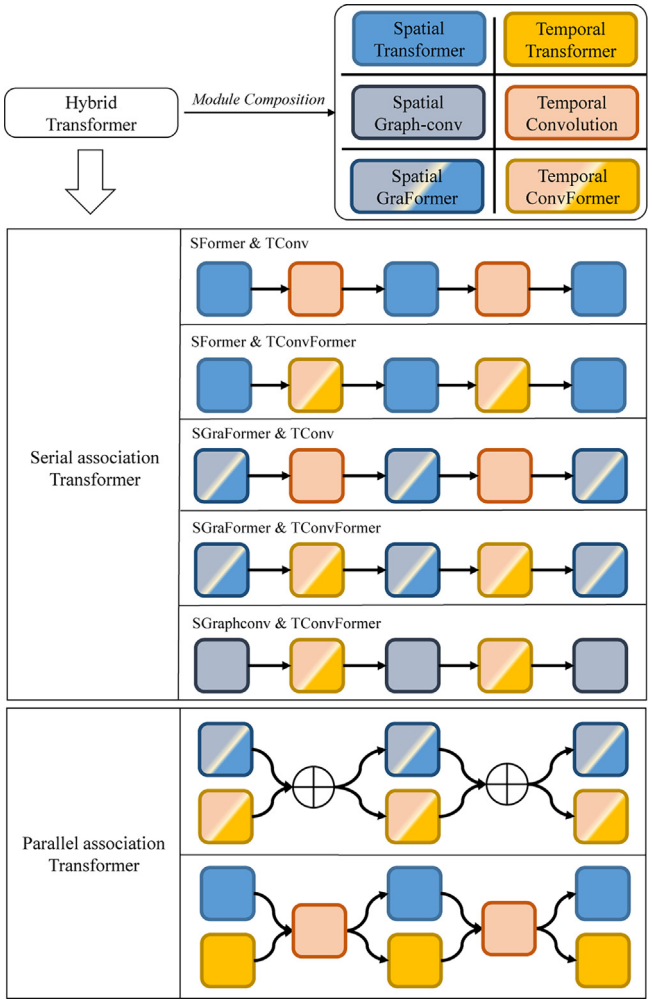


Fig. 5. Illustration of diagrams and ketches in Hybrid Transformer, which explains the basic architecture of the categories.

parallel association transformer, and five small categories—SFormer & TConv, SFormer & TConvFormer, SGraFormer & TConv, SGraFormer & TConvFormer, and SGraphconv & TConvFormer. Detailed labels are proposed to assist readers to clearly obtain the essential information of the classification.

3.2.1. Serial association transformer

(1) Spatial Transformer and Temporal Convolution (SFo-rmer & TConv). Hypergraphs can be used to describe potential higher-order relationships at joint points. Zhou et al. [141] design H-Former to narrow the performance gap between the transformer and GCN in several ways: (1) By introducing a new variant of Self-Attention (SA) named the Hypergraph Self-Attention (HyperSA), the inherent higher-order relationships are integrated into the model. (2) A strong k-hop relative position embedding (k-Hop RPE) based on the shortest path distance (SPD) is proposed, which enables the model to automatically search for the optimal segmentation strategy. (3) The MLP layers are removed. This change makes the intra-joint modeling of simple 3D coordinates omitted, thus reducing the calculation and memory consumption.

Many existing models focus on deep and complex neural networks and train large-scale coarse-grained action datasets. Yuan et al. [142] designed a spatial transformer model recognizing small-scale fine-grained Tai Chi action dataset specially. Due to the weakness of overfitting in the indeterminacy of inter-class distances and similarity of action between Tai Chi action classes, they proposed a transformer network named TC-Former, which merged CNN, GCN and transformer. The model firstly uses spatial transformer and large-scale coarse-grained action datasets to pre-train the model to capture common human motion features and then use small-scale Tai Chi action dataset to retrain the weight in specific layer.

Between non-adjacent frames, feature aggregation module improve the capacity to recognize similar actions. Qiu et al. [143] construct a novel spatio-temporal tuples Transformer (STTFormer) model. The proposed structure can capture the correlation between joints in the same frame and between joints in consecutive frames by applying attention mechanisms to all nodes in the tuple to capture. Spatio-temporal tuple self-attention (STTA) module is used to extract the related features of joints in each short sequence simply and effectively.

The extension of human skeleton-based single-person action recognition to multi-person group activities has a wide range of applications in practice. Yang et al. [144] represent Zoom-Former for improving the traditional graph convolutional network in the following several aspects: (1) Designing a Relation-aware Attention mechanism, which comprehensively leverages the prior knowledge of the human body structure and the global characteristic of the human motion to fully exploit the multi-level features of group activities. (2) Zoom Transformer with Relation-aware Attention is explored to extract the low-level motion information of a single person and the high-level interaction information of multiple persons from the skeleton sequence hierarchically.

To learn the optimal dependency matrix from the uniform distribution based on the multi-head attention mechanism, Yang et al. [146] introduce a novel skeleton-based action recognition method (UNIK). The proposed method is not only effective to learn spatio-temporal features on human skeleton sequences but able to generalize across datasets. Such kind of method leverages information from several representation sub-spaces at different positions of the dependency matrix to effectively learn the spatio-temporal features on skeletons. The proposed UNIK does not rely on any topology related to the human skeleton, making it easier to transfer to other skeleton datasets. Opening up a great design space to further improve the recognition performance by transferring a model pretrained on a sufficiently large dataset. In addition, by integrating the adjacency graph inherent to the human skeleton and transformer, Graph Skeleton Transformer network (GSTN) [145] is designed to remain natural connection information in the skeleton graph while existing transformer-based methods most focus on multi-stream data rather than natural connections of the human body. The authors introduce a grid search method to determine the best multi-stream fusion weights, whereas most Transformers rely on manually setting each input weight.

(2) Spatial Transformer and Temporal Convolution-Transformer (SFormer & TConvFormer). Gao et al. [147] propose a novel end-to-end Focal and Global Spatial–Temporal Transformer network (FG-STForm) to effectively capture relations of the crucial local joints and the global contextual information in both spatial and temporal dimensions. FG-STForm forces the network to focus on modelling correlations for both the learned discriminative spatial joints and human body parts respectively. The selective focal joints eliminate the negative effect of non-informative ones during accumulating the correlations. Furthermore, dilated temporal convolutions are integrated into a global self-attention mechanism to explicitly capture local temporal motion patterns of joints or body parts.

The use of graph transformer operators facilitates the modeling of higher-order spatial dependencies between joints. Liu et al. [148] propose a powerful kernel attention adaptive graph transformer network (KA-AGTN) for skeleton-based action recognition.

Such kind of method introduce a Skeleton Graph Transformer (SGT) block with graph transformer operators to learn the spatio-temporal modes between joints accurately. In addition, an effective skeleton temporal feature enhancer, called temporal kernel attention (TKA) block, is designed by using kernel attention. The global channel-level attention score is calculated to help SGT blocks focus on more noteworthy temporal features. The authors fully consider the time factor and propose a more robust adaptive graph strategy, so that the graph embedding adaptively reflects the interaction of nodes over time.

(3) Spatial Graph-Convolution-Transformer and Temporal Convolution-Transformer (SGraFormer & TConvFormer). As graph embedding and temporal embedding modules, GCN and CNN have efficient local feature exploration capabilities. Chen et al. [149] propose a backbone with a local–global alternation pyramid structure called PGT. The proposed model contains two types of transformer blocks: spatial–temporal transformer blocks and joint transformer blocks. The spatial–temporal transformer block calculates the connection of the global nodes of the graph by a proposed spatial–temporal separated attention(STSA). The spatial self-attention and temporal self-attention can be performed separately with long-range temporal and large-scale spatial aggregation. The joint transformer block discovers the global self-attention of all nodes in the graphs by flattening the tokens in both the spatial and temporal domains.

(4) Spatial Graph-Convolution-Transformer and Temporal Convolution (SGraFormer & TConv). Yue et al. [150] propose an optimization ST-GCN based on the limitation of weak feature extraction and deficient generalization called TAG. They design a novel skeleton-based action recognition framework integrated with transformer structure, which contains adaptive graph convolutional layer and adds attention mechanisms to facilitate to the adaptive graph convolutional layer.

(5) Spatial Graph-Convolution and Temporal Convolution-Transformer (SGraphconv & TConvFormer). Kong et al. [151] improve the traditional skeleton-based action recognition tasks in the following three aspects: (1) Multi-scale temporal embedding modules (MT-EMs) are proposed to capture features with multiple branches in different temporal scales. Meanwhile, MT-EMs are integrated with GCN to embed the original skeleton data to improve recognition ability. (2) Transformer encoders (TE) are adopted to incorporate embeddings and describe the long-term temporal pattern. (3) They suggest a lateral connection (Lac) structure fill the semantic gap between feature extractors and integrators. (4) Finally, the authors design a novel multi-scale temporal transformer (MTT) to realize the global attention mechanism and create long-term dependency across the entire skeleton sequences.

3.2.2. Parallel association transformer

Shi et al. [49] apply parallel association transformer to spatial–temporal skeleton-based action recognition in the following several aspects: (1) A two-stream Spatial–Temporal Transformer network (ST-TR) is proposed for skeleton activity recognition tasks, which use self-attention mechanism on both spatial and temporal dimensions to models dependencies between joints. (2) Spatial Self-Attention (SSA) module extracts low-level features embedding the relations between body parts by employing self-attention inside the frame to compute correlations between each pair of joints in every single frame independently. (3) Temporal Self-Attention Module (TSA) models inter-frame correlation to extract the long-term dependence of the whole action, overcoming the strict locality of standard convolution.

To improve the limitations of entangled spatio-temporal feature representation, Bai et al. [152] propose disentangled spatio-temporal transformer (DSTT) block. The DSTT block is based on Transformer and shows a strong ability in alleviating the local constraints of GCN, separating spatial features from temporal features. And the DSTT block puts dynamic global spatio-temporal attention on all joints and frames and enhances local information to get better learning ability. Meanwhile, The authors introduce a novel Hierarchical Graph Convolutional Skeleton Transformer framework for Action Recognition, namely HGCT, which promotes the disentanglement of spatio-temporal feature representations and remains the advantage of GCN and transformer.

3.3. Transformer-style GCN

As shown in the Fig. 6, the modules of Transformer-style GCN are mainly composed of: spatial graph convolution, temporal convolution, spatial self-attention, temporal self-attention, structure grouping and partition transformer. Fig. 3 clearly shows that the adjacency matrix construction strategy is consistent with Transformer, only the application of adjacency matrix and mask matrix is different. We divide these papers into four categories: Spatial-only Transformer-style GCN, Temporal-only Transformer-style GCN, Spatial–Temporal Transformer-style GCN and Multi-partition Transformer-style GCN. Related papers include generating spatial adjacency matrix and temporal relation matrix in GCN using self-attention.

3.3.1. Spatial-only transformer-style GCN

In 2s-AGCN [57], the elements of C_k are calculated by self-attention, where each element represents the similarity between nodes. Shi et al. [57] combine bone-stream and joint-stream as a two-stream framework, which is effective for simultaneously modeling the first-order and second-order information and promises a considerable increase in action recognition accuracy. In this method, the topology of graph is optimized together with other parameters of the network in an end-to-end learning manner. Sun et al. [154] improve 2s-AGCN and propose a SlowFast graph convolution network (SF-GCN) inspired by the SlowFast convolutional network. It facilitates spatial–temporal feature extraction more efficiently by introducing Fast and Slow pathways. Hu et al. [156] design a Forward-reverse adaptive graph convolutional network for skeleton-based action recognition (FR-AGCN) which is an improvement of 2s-AGCN. The FR-AGCN emphasizes the reversibil-

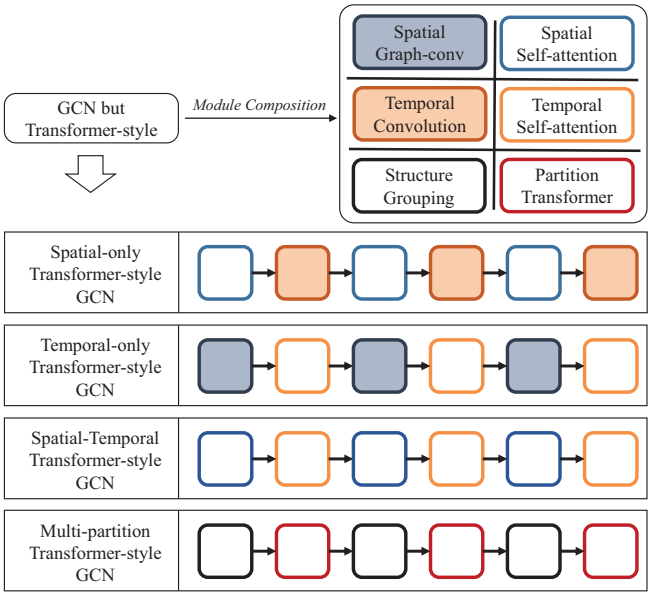


Fig. 6. Illustration of diagrams and ketches in Transformer-style GCN, which explains the basic architecture of the categories.

ity of the skeleton data in the temporal dimension and combines forward and reverse sequences to fully utilize the deep information of the skeleton data. In addition, Xie et al. [155] present a novel attention adjacency matrix (AAM) to design graph convolution kernels and a dimension-attention block to improve the robustness of the model. And The ATM module is obtained by self-attention and used as an additional spatial information aggregation strategy. To improve the ability of 2s-AGCN to aggregate features in the temporal domain, Li et al. [158] propose a multi-stream and enhanced spatial–temporal graph convolution network (MS-ESTGCN). The self-attention strategy of 2s-AGCN is also used in space, but an additional spatial GCL branch is added to enhance the spatial feature extraction ability.

In order to fuse the channel correlation modeling in the realignment stage and the time evolution modeling in the motion interaction stage, Taman et al. [153] apply self-attention and propose the Enhanced Discriminative Graph Convolutional Network (ED-GCN). In the spatial domain, the Squeeze and Exception (SE) module is used to extract discriminatory channel-wise features and integrate the adaptive enhanced feature map into the graph convolution layer to learn better action representation.

Adopting transformers to capture relevant information from neighboring joint points can improve self-occlusion and avoid depth blur. Zeng et al. [157]. design a novel learning skeletal graph neural network named HCSF for Hard 3D Pose estimation to more effectively handle difficult hard poses in 3D pose estimation. The network introduces dynamic skeletal graphs which express correlation connections between joints by integrating all input joints for contextual learning. Zhang et al. [159]. propose the SATD-GCN, which consists of the spatial attention pooling module (SAP) and the temporal dilated graph convolution module (TDGC), the two modules can be integrated into ST-GCN-based models. To mitigate the impact of data redundancy and noise, the SAP module selects human joints that benefit action recognition via the self-attention mechanism.

Self-attention methods can also facilitate feature representations that blend relative angles and relative distances of different spaces. Xing et al. [74] introduce a novel hybrid spatial attention mechanism named HA-GCN to capture implicit connections between joints. Meanwhile, they create a new spatial graph with connections between the head, hands, and feet to increase the accuracy of action recognition.

To further improve the spatial discrimination of the model, Chi et al. [76] present InfoGCN, an information bottleneck-based representation learning framework. An information bottleneck-based learning objective is proposed to instruct the model on how to learn informative but compact latent representations. Furthermore, a graph convolution module based on self-attention, is proposed to capture discriminative information efficiently. They also design a multi-modal representation of the skeleton to provide complementary spatial information to the joints.

Transformers can effectively capture global features of actions. Xie et al. [160] propose a global co-occurrence feature and local spatial feature learning model (GCLS). The Vertex Attention Mechanism branch (VAM-branch) captures the global co-occurrence feature of actions effectively. And the Cross-kernel Feature Fusion branch (CFF-branch) extracts the local spatial structure features of adjacent bones, and restricts the channels unrelated to action recognition. Shahid et al. [161] propose an Adaptive and Self-Attentive graph convolution neural network (ASA-GCN) for skeleton-based action recognition. The network employs local and global adaptive graph topology to distinguish important joints and bones in each frame, and adaptively learn topology of the graph. To capture the important features of each dimension separately, the self-attentive layer mechanism is proposed and embedded in each graph convolution layer.

The self-attention mechanism in transformer is utilized to construct the learnable skeleton adjacency matrix, which can effectively represent the correlation information between joints. Zhang et al. [69] propose a simple and effective semantically guided neural Network (SGN) for skeleton-based action recognition that exploits spatial and temporal correlations at both joints and frames in a hierarchical manner.

3.3.2. Temporal-only transformer-style GCN

Li et al. [162] design a Temporal Enhanced Graph Convolutional Network (TE-GCN) to capture the correlations between non-adjacent temporal distances. The TE-GCN develops a temporal self-attention graph that explicitly creates connections between temporal characteristics that are semantically connected to depict temporal relations between both adjacent and non-adjacent time steps. Meanwhile, multi-stream TE-GCN is proposed to integrate multiple information of skeleton to fully utilize data information and improve the recognition performance.

3.3.3. Spatial–temporal transformer-style GCN

Transformer has achieved excellent results in global spatio-temporal feature extraction. Bai et al. [163] propose a novel Graph Convolutional skeleton Transformer, namely GCsT, which combines Transformer networks and Graph convolutional networks to enhance the flexibility of spatiotemporal map convolution feature extraction. Spatial–temporal Transformer block (STT) is designed to capture global spatial–temporal dependencies by stacking a spatial module (GCN or Spatial Transformer Attention module) and a temporal module (TCN or Temporal Transformer Attention module). The TTA module introduces long-short term attention (LSTA) to extract the correlation between frames in the temporal dimension. In addition, Ke et al. [61] propose a skeleton-based action recognition framework called To-a-T Spatio-Temporal Focus (STF). The STF module with learnable gradient-enforced and instance-dependent adjacency matrices is designed to model high-order spatial–temporal dynamics.

As the time series data, the transformer-based method is difficult to achieve better context learning by using only the spatial module. Hu et al. [164] design a spatial–temporal graph attention network for skeleton-based action recognition, namely STGAT, to extract short-term dependencies. The STGAT continues to use temporal-only modules for long-term modeling while granting spatial-only modules wider latitude to carry out local spatial–temporal modeling. Gao et al. [165] propose a Focusing-Diffusion Graph Convolutional Network (FDGCN). In the focusing process, to extract spatial context, a supernode is generated by an attention module for each frame to model inter-frame and intra-frame connections. Furthermore, a transformer-based context-aware module is proposed to capture the temporal context.

3.3.4. Multi-partition transformer-style GCN

Some skeleton action recognition methods based on GCN will manually divide the whole body into several regions, and then improve the overall recognition ability by calculating the correlation of sub-parts. Self-attention happens to be a good tool for computing association relations. Hou et al. [166] propose the self-attention-based skeleton anchor proposal (SAP) module. The SAP module explores the potential relationship of the human body using triplet representation rather than fixed pair-wise bone connection, which can extract complementary features to the current ST-GCN-based method. To extract the contextual information efficiently, a self-attention-based method is proposed to automatically extract the root points for encoding the target joint's angle information named skeleton-anchors. Qian et al. [63] propose a novel structural attention method for channel-wise adaptive graph convolution which improve the existing methods in the following two

respects: (1) The part relation attention (PRA) module is introduced to measure part relation importance of adaptive adjacency matrix that is formed by GCN. (2) The body symmetry trajectory attention (STA) module is designed and stacked behind PRA to capture symmetric similarity by dividing joints of all frames into the left part and right part and create correlation connection between two parts with cross-attention and linear transformation.

In reducing the computational cost, the transformer can also be applied to multi-modal fusion methods. Yang et al. [77] propose the multi-modal knowledge-embedded graph convolutional networks (MKE-GCN) for action recognition in the wild. The MKE-GCN is able to extract the multi-modal skeleton representations by using the adaptive multi-modal aggregation (AMA) module, and the ability of the MKE-GCN to recognize human action "in the wild" is improved by a multi-modal knowledge distillation (MKD) strategy.

3.4. Unsupervised transformer for Skeleton-based action recognition

Over the past two years, transformers have been applied to both unsupervised and pre-trained tasks for skeleton-based action recognition, showing excellent performance in capturing global context and local joint dynamics. Specifically, transformers are often used to optimize and abstract whole-body motion correlations, long-term relationships in sequences, and dynamic dependencies of spatio-temporal structures. According to specific application scenarios, all unsupervised transformers are divided into two categories: general unsupervised transformer and self-supervised transformer. In particular, the general unsupervised transformer only focuses on the traditional unsupervised skeleton action recognition task.

3.4.1. General unsupervised transformer

To address the limitations of existing transformer models for unsupervised skeleton-based action learning methods in capturing global environment and local joint dynamics, GL-Transformer [108] designs a global and local attention mechanism for learning whole-body motions, long-range temporal dynamics, and human–human interactions. The model takes human interaction into account so that global body motion and local joint motion focus on each other.

Unsupervised methods focus on modeling temporal dependencies in sequences and lack the ability to model spatial structures in human actions. To better capture the spatial and temporal structure in skeleton sequences, Cheng et al. [109] propose H-Transformer, a novel hierarchically aggregated self-attention module. Furthermore, H-Transformer predicts the motion between adjacent frames as a pre-training strategy to better capture the long-term dependencies in the overall sequence.

3.4.2. Self-supervised transformer

When contrastive learning processes the temporal and discriminative information of video, it is easy to ignore the hierarchical spatio-temporal characteristics of human skeleton. Chen et al. [111] propose a self-supervised hierarchical pre-training scheme combined into a hierarchical Transformer-based Skeleton Sequence Encoder (Hi-TRS) to explicitly capture spatial, short-term, and long-term temporal dependencies at the frame, clip, and video levels, respectively. Given a skeleton sequence, Frame Transformer (F-TRS) and Clip Transformer (C-TRS) use self-attention to learn the spatial structure and short-term fine-grained temporal dynamic dependencies between skeleton joints. Video Transformers (V-TRS) then summarize the long-term abstract information from the segments and produce a feature representation of the skeleton sequence. In addition, to explore the temporal dependencies in human action sequences, Motion-Transformer [112] adopts a self-supervised pre-training strategy to automatically capture long-term and short-term temporal dependencies through the transformer's self-attention mechanism during the reconstruction process. In addition, the authors also propose a method to predict the motion stream of human skeleton, which can learn the temporal correlation relationship in the sequence more efficiently.

Inspired by Masked Autoencoders (MAE) [167], Wu et al. [113] propose skeleton MAE for a spatio-temporal masked autoencoder framework for self-supervised 3D skeleton action recognition. The authors leverage a skeleton-based encoder-decoder transformer architecture to reconstruct the masked skeleton sequence. Skeleton MAE employs spatio-temporal masking at both joint and frame level, and this pre-training strategy enables the encoder to output generic skeleton features with spatial and temporal dependencies. In terms of specific applications, inspired by the recent progress of self-supervised large-scale language models such as GPT-3, Endo et al. [110] propose Gaitformer to apply human motion prediction as a practical self-supervised pre-training task. Gaitformer is designed to estimate the severity of movement disorders to cope with the difficulty of Parkinson's Disease (PD) data scarcity. The proposed method is pre-trained on a publicly available dataset to predict the motion of the gait and applied to clinical data to predict the severity of MDS-UPDRS gait impairment.

Due to occlusion or shooting, the extracted skeleton data often has noise in both temporal and spatial dimensions, which will reduce the recognition ability of the model. To adapt to similar situations with incomplete information, Zhang et al. [135] propose a multi-task self-supervised learning method STST to improve the robustness of the model by providing confounding samples in different situations. In particular, since the transformer strategy adopted in this paper is not for self-supervision, it is classified as *Pure Transformer*.

3.5. Transfer transformer for Skeleton-based action recognition

In terms of extracting class-specific prototypes and maintaining consistency in the latent space, the Transformer-based network helps transfer learning to better utilize the global spatial and temporal association representation. In particular, thanks to the transformer's global relational dependency, the network also has the ability to deal with human pose occlusion. According to specific application scenarios, all transfer transformers are divided into two categories: few-shot transformer and one-shot transformer. Few-shot and one-shot methods represent skeleton action recognition tasks that provide a small number or a single labeled sample during the training phase, respectively.

3.5.1. Few-shot transformer

Joint temporal location and simulated viewpoint indexes facilitate meta-learning with finite samples of new categories. Wang et al. [127] propose JEANIE, a few-shot learning pipeline method for 3D skeleton action recognition based on joint temporal and camera viewpoint alignment. In particular, this paper proposes a dynamic time-warping strategy to jointly align the temporal block and the viewpoint index of the simulated skeleton between the support query sequence to select the smoothest path matching the temporal position and the view index. Finally, a similarity-based loss is proposed to encourage alignment of similar sequences while preventing alignment of unrelated sequences.

3.5.2. One-shot transformer

Research on the recognition of skeleton sequences with sparse data, such as one-shot action recognition, does not explicitly consider occlusion. To mitigate the pervasive occlusion interference

that often exists in the real world, Peng et al. [126] propose Trans4SOAR, which utilizes multiple data streams and hybrid attention fusion mechanisms to mitigate the adverse effects caused by occlusion. Specifically, the proposed Trans4SOAR utilizes different types of inputs and performs information exchange through a hybrid attention fusion mechanism (MAFM). Then, Trans4SOAR iteratively augments the transformer feature representations of the intermediate layers by predicting class-specific prototypes and Latent Space Consistency (LSC) loss.

3.6. Innovations and improvements

We construct Table 1, Table 2, Table 3, Table 4, and Table 5 to make a unified display of the innovation points and problems solved by all the above methods. By studying the specificity and consistency, it is obvious that most of the innovations of the algorithms focus on structural morphology and self-attention mechanism. Firstly, a large number of recent methods integrate GCN and CNN methods, which is beneficial to obtain local discriminant joint associations and short-term keyframe features. Dilation Temporal Convolution [59] is adopted by many papers and achieves outstanding results. Secondly, inspired by GCN in the adaptive method of adjacency matrix exploration, the Transformer-based method improves the localization on the self-attention mechanism and performs the secondary processing of Q and K to adapt to the characteristics of the small amount of skeleton data, strong topological relationship and long time dimension. Simultaneously, to avoid the huge computational cost of Transformer, [77,132,154] explore the field of lightweight, reducing the number of parameters while ensuring the accuracy of the calculation. Finally, thanks to the ability to efficiently capture the correlations and dependencies of global spatio-temporal features, unsupervised and transfer skeleton action recognition methods based on transformers have attracted increasing attention, alleviating the overfitting problem and reducing the need for large amounts of labeled training data [81]. It is conducive to optimizing the migration and landing application ability of the overall ecology in the field of action recognition based on the transformer skeleton.

4. Datasets

The labels for action recognition cover a rich class of actions in life and sports. Action recognition datasets can be divided into several categories according to their size, including large datasets NTU-RGB + D [37,38], Kinetics [168], medium datasets UAV-Human [169], PKU-MMD [170], NW-UCLA [171], UWA3D [172], small datasets SBU [173]. Besides, for realistic tasks, nursing oriented NCRC [133] and Tai-Chi Action [142] have also been proposed. Table 6 presents a summary of commonly used datasets.

4.1. NTU RGB + D

The NTU-RGB + D dataset [37] contains 56,880 action videos. The dataset is captured by Microsoft Kinect-v2 for 40 participants

Table 1
Innovations and improvements in Pure Transformer.

Method	Taxonomy	Innovations	Improvements
IIP-Former [131]	Pure Serial Transformer	According to the topology of the human body, the human body are aggregated into multiple parts. Based on the transformer, the concept of human body parts is introduced into the skeleton-based action recognition.	IIP-Former improves existing methods limitations in: (1) Transformer-based networks brings quadratic computation and memory cost on action recognition tasks. (2) Previous studies mainly focus on the relationships among individual joints, which often suffers from the noisy skeleton joints.
MSST-RT [132]	Pure Serial Transformer	MSST-RT completely relies on a relative transformation mechanism is proposed to learn long-range dependencies. A multi-scale dynamic representation (DR) is introduced to combine the multi-scale features in skeletons.	By employing the relative transformer mechanism to learn correlations between joints, they improve the limitation in catching long-range dependence.
NCRC [133].	Pure Serial Transformer	A dual modality transformer for bone joint signals and acceleration signals is proposed, which combines a new fusion technology based on attention to fuse spatio-temporal skeletal features and acceleration features.	By proposing cross-attention-based fusion between skeletal joints and acceleration data, it makes up for the limitation that some GCN-based methods only focus on single-scale features.
DSTANet [134]	Pure Serial Transformer	DSTANet is the first to propose a decoupled spatial–temporal attention networks (DSTA-Net) for skeleton-based action recognition, which is built with pure attention modules without manual designs of traversal rules or graph topologies.	DSTANet improves the limitations of traditional manual design of traversal rules or graph topologies, and adopts pure attention modules with better performance and versatility.
STST [135]	Pure Parallel Transformer	STST explicitly defines three types of encoding strategies to consider main situations of joints and design specific Transformers for the temporal and spatial dimensions.	Most of the existing works treat skeleton sequences in the temporal and spatial dimension in the same way, ignoring the difference between the temporal and spatial dimension in skeleton data.
TSTE [136]	Pure Parallel Transformer	By combining motion spatio-temporal feature embedding and shape transformation, the two-stream transformer encoder (TSTE) network is 30% of the general recognition method.	The transformer proposed by TSTE incorporates the inherent higher-order relations into the model instead of computing the adjacency matrix, thus reducing the computational cost.
STAR [137]	Pure Parallel Transformer	STAR turns their attention from parameterized and complex models to sparse transformer.	STAR greatly reduces existing methods' computational complexity and memory usage.
IGFormer [138]	Pure Fusion Transformer	The GI-MSA module is introduced to learn the relationship between persons at both semantic and distance levels. The Semantic Partition Module (SPM) is proposed to transform each skeleton sequence into a BPT sequence to retain interactive body part information.	IGFormer improves the limitation of existing methods in two ways: (1) Model the relationship between interaction personnel from the semantic and distance levels. (2) Transform each skeleton sequence into a BPT sequence to enhance the modeling of interactive body parts.
Ac-Former [139]	Pure Fusion Transformer	A new model called AcT based on fully self-attentional architecture is represented, which is superior to the existing CNN and RNN models. Furthermore, a new dataset named MPOSE-2021 is proposed for HAR tasks.	For human motion recognition task, the solution that completely relies on self focus block has not been studied. This paper focuses on this problem and proposes an architecture derived from pure transformer encoder.
START [140]	Pure Fusion Transformer	Based on cross-modal learning and the STAR-transformer attention mechanism, this paper proposes a multi-feature representation method, in which the cross modal learning method can flexibly aggregate skeleton features and video frames.	By aggregating cross-modal data of spatio-temporal video and skeleton into a multi-class token, they improve the existing limitations in separating model and balancing feature representation for cross modal data.

Table 2
Innovations and improvements in Hybrid Transformer.

Method	Taxonomy	Innovations	Improvements
H-Former [141]	Hybrid Serial Transformer	The Hyperformer contains the HyperSA and an effective temporary cooperation module. The HyperSA can model joint co-occurrences, while the temporary cooperation module is introduced for temporary modeling.	They improve the limitations of the methods based on the transformer in: (1) Learning the bone connectivity of bone data and (2) Extracting the potential relationship between the body joints that has a distinct physical functionality for human action.
TC-Former [142]	Hybrid Serial Transformer	TC-Former constructs a small-scale dataset that captures fine-grained Tai Chi action especially.	TC-Former greatly improves the overfitting problem of the model in the fine-grained dataset training.
STTForm [143]	Hybrid Serial Transformer	By introducing spatio-temporal tuple to model the correlation of nodes in multiple consecutive frames of the tuple, the relationship between different joints in consecutive frames can be captured.	Traditional methods only focus on the same connection between frames, which makes the extracted motion features too simple to capture the correlation of different joints between frames.
Zoom-Former [144]	Hybrid Serial Transformer	Zoom-Former focuses on multi-person group activities. It uses the characteristics of people as input to capture the spatio-temporal interaction information between people in the scene, to realize the recognition of group activities.	They improve existing methods' limitations in: (1) Focusing on single-person action recognition while neglecting the group activity of multiple people. (2) Unable to mine high-level semantic information from the skeleton data.
GSTN [145]	Hybrid Serial Transformer	GSTN introduces the adjacency graph structure of the skeleton joints into the transformer attention map and proposes a grid-search method to assign weight to each input stream automatically.	This paper improves the disadvantage that the existing transformer network uses data as input and ignores the characteristics of the natural connection structure of the human body.
UNIK [146]	Hybrid Serial Transformer	Previous methods trained from scratch without taking advantage of fine-tuning on a pre-trained model. They are the first to explore the pre-training and fine-tuning strategies for real-world videos.	GCN-based methods are difficult to generalize across domains, UNIK improves the adaptability of existing methods under different human topologies.
FG-STForm [147]	Hybrid Serial Transformer	FG-STForm is designed to enhance spatial dependency by fusing interactions between focal joint points and body parts employing mutual cross-attention.	This paper improves the following aspects: (1) Undervaluing the effect of discriminative local joints and the short-range temporal dynamics. (2) Directly modeling effective temporal relations of joints globally over a long input sequence.
KA-AGTN [148]	Hybrid Serial Transformer	A novel attempt to apply the graph transformer operator to skeleton-based action recognition tasks, learning the spatio-temporal modes between joints accurately. Such kind mode can both greatly mitigate the oversmoothing problem and capture long-distance dependencies more effectively.	This paper improves limitations in: (1) Accurately capture dependencies between vertices representing joints, especially vertices with long distances. (2) Allowing the model to focus on more informative temporal features and further generate skeleton graph embeddings.
PGT [149]	Hybrid Serial Transformer	By combining the spatial–temporal transformer blocks and joint transformer blocks, the fusion module can learn robust features for skeleton graphs with dynamic attention effectively.	This paper improves limitations in: (1) Catching the temporal and spatial correlation between all nodes. (2) Dynamic fusions of the global joints.
TAG [150]	Hybrid Serial Transformer	The authors propose an adaptive graph convolutional layer integrated with attention mechanisms based on Transformer to optimize the original ST-GCN model.	They improve the ability of weak feature extraction of bone, enhancing the generalization ability of the model greatly and extracting the weak features of bones more effectively.
MTT [151]	Hybrid Serial Transformer	This paper designs an all-encompassing multi-scale temporal transformer for skeleton-based action recognition. The proposed transformer can learn temporal patterns at different scales.	MTT improves the shortcomings in: (1) Difficulty to get long-term temporal information. (2) Ignoring information in various temporal scales.
STTR [49]	Hybrid Parallel Transformer	The proposed method discards any predefined skeleton structure, and can automatically discover the joint relationship related to the prediction of the current action.	STTR has the following advantages: (1) Effectively encodes the hidden information of the underlying 3D skeleton. (2) Collect effective information from joint relations.
HGCT [152]	Hybrid Parallel Transformer	The authors suggest a unique design called Hierarchical Graph Convolutional Skeleton Transformer (HGCT) to combine the benefits of GCN and Transformer, which include the advantages of local topology, global context, and dynamic attention.	The proposed method has the following advantages: (1) Limitation of feature aggregation in the immediate spatial–temporal vicinity. (2) Sequential spatiotemporal features learning for stacked STGC block.

aged 10 to 35. The dataset mainly contains four types of Data Modalities, namely RGB videos, depth sequences, infrared frames, and skeleton data. The resolution of RGB videos is 1920 × 1080, while the resolution of depth sequences and infrared frames is 512 × 424. The 3D skeleton data includes the 3D positions of 25 main body joints in the human body. The dataset contains 60 action categories, including 40 daily activities (drinking, eating, reading, etc.), 9 health-related behaviors (sneezing, stumbling, falling, etc.), and 11 mutual actions (punching, kicking, hugging, etc.). For each action, three cameras with the same height are used to capture three different horizontal attempts from three different horizontal angles ($-45°, 0°, +45°$). The subject should perform each action twice, facing the left and right cameras respectively. In this way, two front views, one left side view, one right side view, one left side 45 degrees view, and one right side 45 degrees view are captured. Two Benchmark evaluations named cross-subject (CS) and cross-view (CV) are recommended. In cross-subject, 40 subjects are equally divided into two groups. The training set and the testing set contain 40320 and 16560 samples respectively. In cross-view, the training set consists of 37920 samples, including the front view and two side views of the action. The testing set consists of 18,960 samples, including two 45-degree views on the left and right of the action.

4.2. NTU RGB + D 120

NTU-RGB + D 120 [38] is currently the largest dataset with 3D joint annotations which is extended from NTU RGB + D 60. It contains 57367 new skeleton sequences and additional 60 new action categories. It has 114,480 videos in total and 120 classes which are captured from 32 different camera setups and 106 distinct subjects. The dataset follows two evaluation protocols: Cross-Subject and Cross-Setup. Cross-Subject divides subjects into two parts, one for training and the other for dataset testing. Cross-Setup splits samples by the camera setup, similarly, one part is for training and the other part for testing.

4.3. SBU interaction

SBU Interaction Dataset [173] contains eight classes of human interactions including approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands. Each interaction is performed by two subjects. The dataset contains 282 videos recorded by seven participants. Each frame of the video contains the 3D coordinates of 15 joints for each participant. All videos are recorded in the same lab environment. The dataset was divided into 21 groups, each of which consisted of a different

Table 3
Innovations and improvements in Spatial-only Transformer-style GCN.

Method	Taxonomy	Innovations	Improvements
2s-AGCN [57]	Spatial-only Transformer-style GCN	The authors propose an adaptive GCN structure to build flexible topological graphs with the self-attention mechanism. The two-stream network is proposed by introducing the skeleton second-order information.	In existing methods, the topology of the graph is set manually, and it is fixed over all layers and input samples. In addition, the second-order information (the lengths and directions of bones) of the skeleton data is rarely investigated in existing methods.
ED-GCN [153]	Spatial-only Transformer-style GCN	The SE module is proposed to obtain discriminative channel features, and the ATB is proposed to flexibly model the temporal features in two stages.	The existing GCN-based methods mainly express actions by learning appropriate graphs, while the proposed ED-GCN introduces SE modules to enhance the significant features to learn better representation.
SFGCN [154]	Spatial-only Transformer-style GCN	Inspired by SlowFast convolutional network, SF-GCN contains the Fast and Slow pathway that focuses on learning slow or fast temporal changes respectively.	This paper significantly improves the significant temporal dependency restrictions of the skeleton sequence in prior approaches.
AAM-GCN [155]	Spatial-only Transformer-style GCN	The proposed method effectively solves the problem of over-smoothing and uses self-attention to get rid of the dependence on the manually designed center of gravity.	AAM-GCN improves the partitioning strategy of neighbor set for graph vertices on the gravity center designed manually, which is limited in generalizability to diverse skeletons in action recognition.
FR-AGCN [156]	Spatial-only Transformer-style GCN	Reverse AGCN is introduced to emphasize the reversibility of the skeleton data in the temporal dimension and integrate the forward and reverse sequences of the skeleton data to improve recognition performance.	They improve the existing methods' limitation in deep information utilization of the skeleton data and in overfitting due to multiple copies of the motion sequences with fewer frames.
HCSF [157]	Spatial-only Transformer-style GCN	This paper introduces an adaptively dynamic skeletal graphs to capture implicit connection and proposes a novel hop-aware hierarchical channel-squeezing fusion layer to overcome the limitation of skeletal graph recognition.	This paper improves the existing methods' shortcomings in skeletal recognition preference with problems of self-occlusion, deep ambiguity, and difficult or uncommon positions in hard 3D Pose.
MS-ESTGCN [158]	Spatial-only Transformer-style GCN	The proposed MS-ESTGCN is the first six-stream network that uses six modalities of data as input.	By connecting multiple temporal GCLs with different kernel sizes densely, they improve the limitation that ST-GCNs utilize only one fixed kernel.
SATD-GCN [159]	Spatial-only Transformer-style GCN	The SAP module is proposed to extract discriminative joint features and the TDGC module is proposed to learn hierarchical temporal features.	The authors improve the limitations of the ST-GCN-based methods in: (1) High-order spatial–temporal importance is not embedded in spatial connection topology. (2) The attention module cannot effectively capture the spatio-temporal changes.
HA-GCN [74]	Spatial-only Transformer-style GCN	A novel hybrid spatial attention mechanism is represented to capture implicit connections between joints which integrates transformers to perform better context learning.	HA-GCN improves the limitation in feature extraction of implicit connection between joints while most existing methods focus on physical skeleton graphs.
Info-GCN [76]	Spatial-only Transformer-style GCN	A self-attention graph convolution module and an information bottleneck strategy are proposed to model the context-dependent intrinsic topology and learn the action latent representation.	Introducing an information bottleneck objective and incorporating a self-attention mechanism improves the latent representation of human actions that most works ignore.
GCLS [160]	Spatial-only Transformer-style GCN	They propose a Cross-kernel feature Fusion (CFF), instead of using the traditional feature fusion based on the same convolution kernel.	Existing GCN-based methods cannot effectively capture the global co-occurrence features between joints and bones.
ASA-GCN [161]	Spatial-only Transformer-style GCN	A self-attention module is proposed to distinguish and learn important features from spatial, temporal, and channel dimensions respectively.	Local and global adaptive graph topologies are introduced to enhance the ability to flexibly model multi-layer semantics.
SGN [69]	Spatial-only Transformer-style GCN	This paper explicitly explores the joint semantics (frame index and joint type) and presents a semantics-guided neural network (SGN).	The self-attention mechanism is used to construct the correlation matrix in the joint-level module to represent the correlation degree between joints.

pair of people engaging in all 8 interactions. In most interactions, one of the subjects is acting, and the other response.

4.4. Northwestern-UCLA

Northwestern UCLA dataset [171] consists of 1494 video sequences. The dataset contains 10 action categories, which are completed by 10 subjects and simultaneously shot by 3 Kinect cameras from different directions. The samples taken by the first two cameras are used as training sets, and the samples taken by the third camera are used as testing machines.

4.5. NCRC

NCRC [133] is a dataset specially designed for the identification of nursing activities of nurses. The dataset is from the Nurse Care Activity Recognition Challenge. The NCRC dataset consists of 498 samples, including 6 different actions completed by 8 nurses in a monitored environment. The actions include vital signs measurement, blood collection, blood glucose measurement, indwelling drill retention and connection, oral care, diameter exchange, and cleaning of the area. There are 282 action samples in the training set, including all the above actions performed by 6 nurses. There are 216 action samples in the testing set, including actions performed by 2 nurses. NCRC is not a widely used dataset, the limitations of the dataset include that the sampling rate of the sensor varies greatly, the acceleration data is noisy, and the overall size of the dataset is very small.

4.6. Tai-Chi Action

Tai Chi Action dataset [142] is a small-scale fine-grained dataset captured by the wearable device Perception Neuron. It contains 10 Tai Chi action classes, each with 20 samples. The skeleton of each frame contains 72 joints.

4.7. Kinetics

Kinetics [168] is a large-scale human action dataset with a total of 300000 video clips. All these video clips are downloaded from

Table 4
Innovations and improvements in Temporal-only, Spatial–Temporal and Multi-partition Transformer-style GCN.

Method	Taxonomy	Innovations	Improvements
TE-GCN [162]	Temporal-only Transformer-style GCN	This paper designs a Temporal Augmented Graph Convolutional Network (TE-GCN), which constructs a correlation matrix for context learning in the temporal dimension.	The authors improve the existing methods' opacity in exploring the temporal dynamics of skeleton sequence due to the key temporal information dilution resulting from the repeat of massive local convolutions.
GCsT [163]	Spatial–Temporal Transformer-style GCN	This paper designs a novel Graph Convolutional skeleton Transformer that combines the advantages of both GCNs and Transformer to efficiently extract local–global contexts.	The authors improve the existing methods' limitation in feature aggregation of neighborhood and deficiency of flexibility in feature extraction.
STF [61]	Spatial–Temporal Transformer-style GCN	The STF module that generates dynamic connection topology is designed and three loss terms called STF exploration, STF divergence, and STF coherence are proposed.	STF improves the drawbacks of ST-GCN-based methods: (1) Fixing the temporal kernel size over all layers. (2) Capturing global graph features from joint features via average pooling.
ST-GAT [164]	Spatial–Temporal Transformer-style GCN	A novel spatial–temporal graph attention network (ST-GAT) is designed to extract short-term dependencies by connecting joints in local spatial–temporal neighborhoods and dynamic capturing the implicit connection between joints.	ST-GAT improves the existing methods' limitation of feature extraction in short-term dependencies which are important to distinguish some similar behaviors while most existing methods focus on long-term temporal dependencies.
FDGCN [165]	Spatial–Temporal Transformer-style GCN	A new architecture called FDGCN is proposed, which decomposes the skeleton into the focusing and diffusion graphs and combines the transformer to explore the spatial–temporal context efficiently.	This paper improves the limitation of existing methods in exploring the spatial–temporal context by generating the super-node to model inter-frame and intra-frame connections.
SAP-GCN [75]	Multi-partition Transformer-style GCN	A high-order triplet representation is proposed to replace the pairwise skeleton connection, and the anchor location is automatically determined by the anchor location learning method of the self-attention mechanism.	Exploring angular information through a higher-order triplet representation rather than using a pairwise representation with fixed topological constraints to extract human relations improves the limitations of potential joint correlations.
MKE-GCN [77]	Multi-partition Transformer-style GCN	A lightweight AMA module and MKD strategy are proposed, which can improve the accuracy of action recognition in the wild without increasing the complexity of the model.	MKE-GCN improves the limitations of action recognition in the wild by proposing the adaptive multi-modal aggregation strategy and knowledge distillation strategy.
SA-GCN [63]	Multi-partition Transformer-style GCN	A novel body symmetry trajectory attention (STA) module is designed to capture symmetric similarity and a part relation attention (PRA) module is proposed to measure the local correlation weight of adaptive adjacency matrix.	SA-GCN enhances the shortcomings of current approaches in the extraction of implicit structural characteristics, which are often concealed in the geometric properties of the skeleton in the spatial domain.

Table 5
Innovations and improvements in Unsupervised Transformer and Transfer Transformer.

Method	Taxonomy	Innovations	Improvements
GL-Transformer [108]	General Unsupervised Transformer	To model local joint dynamics and capture global context from skeletal motion sequences of multiple people, this paper designs a novel transformer architecture including global and local attention (GLA) mechanism.	The authors introduce a novel pretraining strategy, multi-interval displacement prediction (MPDP), to learn both global and local attention in diverse time ranges. The proposed model successfully learns local dynamics of the joints and captures global context.
H-Transformer [109]	General Unsupervised Transformer	The authors propose a novel unsupervised representation learning framework by introducing a hierarchical structure of body part connections.	To facilitate the modeling of long-term temporal dependencies, the paper introduces the prediction of 3D bone motion as a new pre-training task.
Gait-Transformer [110]	Self-supervised Transformer	This paper proposes a new method, Gaitformer, to predict motion and gait while estimating injury severity. The proposed method significantly outperforms previous methods that rely on clinical data.	Based on recent advances in Transformer and attention models, gaitformer can leverage large-scale public datasets of human motion and behavior to learn reliable motion representations.
Hi-TRS [111]	Self-supervised Transformer	The authors propose a novel approach to encode skeleton sequences with hierarchical transformer encoders and design a pre-training scheme consisting of three different levels of pre-training tasks. Experiments show that the proposed method is an effective way to learn the skeleton representation.	The prior knowledge learned through hierarchical pre-training shows strong transfer learning ability for downstream tasks at different levels. For both supervised and semi-supervised settings, this algorithm achieves state-of-the-art performance compared to competing baselines.
Motion-Transformer [112]	Self-supervised Transformer	This paper proposes the Motion-Transformer model to take full advantage of temporal dependencies in sequential inputs for skeleton-based action recognition: (1) Pre-train the proposed model using a self-supervised reconstruction algorithm. (2) Predict the optical flow between adjacent skeleton frames as another pre-training task.	Due to the inherent properties of Transformers, existing methods often rely on the ability to analyze temporal dependencies. This model focuses more on the information of motion dynamics and behavioral consistency within human actions, thus enhancing the exploration of more fine-grained temporal dependencies.
Skeleton-MAE [113]	Self-supervised Transformer	This paper proposes a simple and efficient skeleton-based masked autoencoder architecture that aims to learn comprehensive and general skeleton-based feature representations.	To better understand the skeleton mask approach, the authors propose a novel joint and frame level spatio-temporal mask for skeleton data and explored different masking methods.
JEANIE [127]	Few-shot Transformer	This paper presents a few-shot action recognition method for learning 3D skeletons via JEANIE, which jointly aligns the viewpoint indices of temporal blocks and simulated skeletons between suppose-query sequences.	The authors optimize a smooth path between jointly modeled queries and support frames to achieve optimal alignment in both time and simulated camera viewpoint space, enabling end-to-end learning with limited few-shot training data.
Trans4SOAR [126]	One-shot Transformer	This is the first action recognition benchmark for human skeleton data considering occlusion. It breaks through the limitation that the standard recognition of 3D bones only considers random joint missing. This benchmark is able to resolve more realistic occlusions produced by everyday objects.	This paper proposes realistic synthesis and random occlusion to better solve the occlusion problem. The authors propose a novel skeleton architecture, Trans4SOAR, to provide discriminative representations for skeleton inputs and enhance robustness to different scenes.

Table 6
Summary of commonly used datasets based on transformer skeleton action recognition methods.

Dataset	Size	#Classes	#Videos	#Joints	#Subjects	Web.	Papers
NTU-60 [37]	Large	60	56,880	25	40	NTU-60	[131,132,134–137], [138,140,141,143–145], [146–151], [49,152,57,153–155], [156–159,74,76], [160,161,69,162,163,61], [164–166,77,63]
NTU-120 [38]	Large	120	114,480	25	106	NTU-120	[131,132,134,136–138], [140,141,143–146], [147–149,151,49,152], [154,156,157,76,161,69], [162,163,61,164,77,63]
Kinetics [168]	Large	400	300000	18	-	KNT-400	[135,148,151,57,153,154], [155,158,159,74,160,161], [164,165]
UAV-Human [169]	Medium	155	22,476	17	119	UAV-155	[132,57,156,77]
PKU-MMD [170]	Medium	41/51	2085	25	66/13	PKUMMD	[111]
NW-UCLA [171]	Medium	10	1,494	20	10	UCLA-10	[145,141,76,163,63]
UWA3D [172]	Medium	30	1075	15	10	UWA-3D	[127]
SBU [173]	Small	8	282	15	7	SUB-8	[138]
NCRC [133]	Small	6	498	29	8	NCRC-6	[133]
Tai-Chi [142]	Small	10	200	72	-	TAI-10	[142]

YouTube, containing 400 categories and each video lasts for 10 s. The 234619 video clips are used to train the model, and the remaining 19761 video clips are used to verify the model accuracy. Then we use the OpenPose toolbox to estimate the skeleton. Each frame includes 3D coordinates and joint confidence of 18 joints.

4.8. UAV-Human

UAV-Human [169] is a 155-classes action recognition dataset containing 22,476 video clips. The dataset was collected by a UAV in multiple urban and rural areas during the day and night, and covers a wide variety of diverse backgrounds, illumination, weather, occlusions, and also includes multiple UAV movements and flight poses. Action data are collected from 119 different subjects and 155 different activity categories at 45 different environmental locations. The authors suggest the following evaluation method: 89 subjects for training and 30 subjects for testing.

4.9. PKU-MMD

PKU-MMD [170] is a medium-scale dataset focusing on long continuous sequences action detection and multi-modality action analysis. The dataset is captured via the Kinect v2 sensor. PKU-MMD contains 2 parts, for action detection tasks with increasing difficulty. Part 1 is large-margin action detection task. Part 2 is small-margin action detection task. Compared to Part I, Part II is more challenging due to short action intervals, concurrent actions and heavy occlusion. Part I contains 1,076 untrimmed video sequences with 51 action classes performed by 66 subjects. Part II contains 1,009 untrimmed video sequences with 41 action classes performed by 13 subjects. Joint information consists of 3-dimensional locations of 25 major body joints for detected and tracked human bodies in the scene.

4.10. UWA3D

UWA3D Multiview Activity dataset [172] consists of 30 activities performed by 10 human subjects of varying scales. To capture depth videos from front view, each subject performed two or three random permutations of the 30 activities in a continuous manner. For cross-view action recognition, 5 subjects performed 15 activities from 4 different side views. The total number of action sequences is 1075. The dataset The dataset is challenging due to self-occlusions and high similarity. It uses the Kinect to emphasize three factors: (1) Scale variations between subjects. (2) View-point variations. (3) All actions were performed in a continuous manner with no breaks or pauses. Thus, the start and end positions of body for the same actions are different.

5. Comparisons

In Table 7 and Table 8, we list the papers presented above, including the algorithms, classifications and accuracies on different datasets. Most methods continue the spatio-temporal graph convolution architecture in ST-GCN, seeking innovative points in the temporal domain and the spatial domain. Some methods are keen to change the model structure, blur the spatio-temporal boundaries, and use the holistic method to update and obtain features. We are excited to see the emergence of many new action datasets that will benefit the practice and development of this field.

6. Challenges and future directions

The current social life has an increasing demand for artificial intelligence technology, and one of the centers of artificial intelligence technology is the human-centered multimedia interaction work. Human action recognition is the key to decrypt human biological passwords and explore and analyze the mysteries of action. In the current society with high demand for recognition technology, if we cannot accurately understand the behavior of the human body, it will greatly affect the development of human-oriented AI fields. Although skeleton-based action recognition has gained a lot of attention, we still believe that the following challenges will need to be overcome in the future:

(1) *The challenge in end-to-end module.* The key point of skeleton action recognition task is the acquisition and preprocessing of skeleton data. As a typical upstream task, the corresponding downstream task is pose estimation. Although many attempts try to overcome the influence of illumination, occlusion, noise and Angle in the prediction task, it is still difficult to obtain pure and accurate joint coordinates [174]. Precisely accurate joint coordinates will have a decisive impact on the accuracy of skeleton action recognition.

Table 7
Performance comparison for Skeleton Transformer and Transformer-style GCN methods for NTU RGB + D, NTU RGB + D 120, and Kinetics-400 datasets. P. is short for Pure, H. is short for Hybrid, S. is short for Spatial, T. is short for Temporal, S.T. is short for Spatio-Temporal, and M. is short for Multi-partition.

Method	*Category*	*Year*	NTU RGB + D 60		NTU RGB + D 120		Kinetics	
			C-Sub (%)	C-View (%)	C-Sub (%)	C-Set (%)	Top-1 (%)	Top-5 (%)
DSTA-Net [134]	P. Serial Trans.	2020	91.5	96.4	86.6	89.0	-	-
MSST-RT [132]	P. Serial Trans.	2021	88.4	93.2	79.3	82.3	-	-
IIP-Former [131]	P. Serial Trans.	2022	92.3	96.4	88.4	89.7	-	-
STST [135]	P. Parallel Trans.	2021	91.9	96.8	-	-	38.3	61.2
STAR [137]	P. Parallel Trans.	2021	83.4	89.0	78.3	80.2	-	-
TSTE [136]	P. Parallel Trans.	2022	80.5	85.3	66.6	67.5	-	-
IGFormer [138]	P. Fusion Trans.	2022	93.6	96.5	85.4	86.5	-	-
START [140]	P. Fusion Trans.	2022	92.2	96.5	90.3	92.7	-	-
UNIK [146]	H. Serial Trans.	2021	86.8	94.4	80.8	86.5	-	-
H-Former [141]	H. Serial Trans.	2022	92.6	96.5	89.9	91.2	-	-
STTFormer [143]	H. Serial Trans.	2022	92.3	96.5	88.3	89.2	-	-
Zoom [144]	H. Serial Trans.	2022	90.1	95.3	84.8	86.5	-	-
GSTN [145]	H. Serial Trans.	2022	91.3	96.6	86.4	88.7	-	-
FG-STForm [147]	H. Serial Trans.	2022	92.6	96.7	89.0	90.6	-	-
KA-AGTN [148]	H. Serial Trans.	2022	90.4	96.1	86.1	88.0	38.1	61.0
PGT [149]	H. Serial Trans.	2022	90.9	95.9	86.5	88.8	-	-
TAG [150]	H. Serial Trans.	2022	82.1	90.0	-	-	-	-
MTT [151]	H. Serial Trans.	2022	90.8	96.7	86.1	87.6	37.9	61.3
STTR [49]	H. Parallel Trans.	2021	88.7	95.6	81.9	84.1	-	-
HGCT [152]	H. Parallel Trans.	2022	92.2	96.5	89.2	90.6	-	-
2s-AGCN [57]	S. Trans-style GCN	2019	88.5	95.1	-	-	36.1	58.7
MS-GCN [158]	S. Trans-style GCN	2020	91.4	96.8	-	-	39.4	62.1
GCLS [160]	S. Trans-style GCN	2020	89.5	96.1	-	-	37.5	60.5
SGN [69]	S. Trans-style GCN	2020	89.0	94.5	79.2	81.5	-	-
SF-GCN [154]	S. Trans-style GCN	2021	89.5	96.2	84.4	86.1	37.0	59.5
AAM-GCN [155]	S. Trans-style GCN	2021	90.4	96.2	-	-	37.5	60.5
FR-AGCN [156]	S. Trans-style GCN	2021	90.5	95.8	86.6	87.0	-	-
HCSF [157]	S. Trans-style GCN	2021	91.6	96.7	87.5	89.2	-	-
ED-GCN [153]	S. Trans-style GCN	2022	88.7	95.2	-	-	36.9	59.0
SATD [159]	S. Trans-style GCN	2022	89.3	95.5	-	-	36.6	59.8
HA-GCN [74]	S. Trans-style GCN	2022	92.1	97.0	-	-	38.2	61.1
Info-GCN [76]	S. Trans-style GCN	2022	93.0	97.1	89.8	91.2	-	-
ASA-GCN [161]	S. Trans-style GCN	2022	90.4	95.6	87.1	89.8	38.7	56.1
TE-GCN [162]	T. Trans-style GCN	2020	90.8	96.2	84.4	85.9	-	-
ST-GAT [164]	S.T. Trans-style GCN	2020	92.8	97.3	88.7	90.4	39.2	62.8
GCsT [163]	S.T. Trans-style GCN	2021	91.6	96.2	87.7	89.3	-	-
FDGCN [165]	S.T. Trans-style GCN	2021	90.9	96.5	-	-	37.7	60.8
STF [61]	S.T. Trans-style GCN	2022	92.5	96.9	88.9	89.9	39.9	-
SAP-GCN [166]	M. Trans-style GCN	2021	92.5	96.9	-	-	-	-
MKE-GCN [77]	M. Trans-style GCN	2022	92.5	96.9	89.7	91.1	-	-
SA-GCN [63]	M. Trans-style GCN	2022	92.6	96.7	89.0	90.7	-	-

Table 8
Performance comparison for Skeleton Transformer and Transformer-style GCN methods for NW-UCLA and UAV-Human datasets.

Method	*Category*	NW-UCLA Top-1 (%)
GSTN [145]	M. Serial Trans.	95.9
H-Former [141]	M. Serial Trans.	96.6
Info-GCN [76]	S. Trans-style GCN	97.0
GCsT [163]	S.T. Trans-style GCN	96.1
SA-GCN [63]	M. Trans-style GCN	97.0

Method	*Category*	UAV-Human	
		Cv-1 (%)	Cv-2 (%)
MSST-RT [132]	P. Serial Trans.	41.22	-
2s-AGCN [57]	S. Trans-style GCN	34.84	66.68
FR-AGCN [156]	S. Trans-style GCN	43.98	69.50
MKE-GCN [77]	M. Trans-style GCN	44.60	-

So how to apply skeleton action recognition algorithm to build the efficient end-to-end module is a serious challenge. Some semi-supervised [107,175,87,176,105] techniques hope to make up for the loss of the model by reducing the amount of data acquisition, but the experimental results can find that the key is the high quality of the data. The jitter or disappearance of skeleton data caused by the existing pose estimation algorithm will have a disastrous impact on the results of action recognition. Perhaps there is a need for photography devices such as kinetics to algorithmically optimize and balance costs for better skeleton acquisition and preprocessing at the bottom of the raw data.

(2) *The challenge in multi-person recognition.* Most of the existing skeleton-based action recognition datasets are for single and double people. In practical applications, more scenarios are for multi-person behavior recognition. Multi-person skeleton extraction algorithms have made some progress [177–179], but how to properly process these complex skeleton sequences is still rarely studied. Therefore, it is a difficult problem to transfer skeleton action recognition algorithm from single or two-person to multi-person.

(3) *The challenge in multi-action recognition.* Most of the current skeleton-based action recognition datasets are for single action recognition. In practice, more scenarios are oriented

to multi-action recognition. [180,181] use sliding windows and memory network to deal with multi-action sequences, but they still face the problems of single action segmentation, false detection and missed detection, and generate a large burden of computing resources. So how to design skeleton action recognition algorithm from single action to multi-action sequence, while effectively reducing the computational cost, is a problem to be overcome.

(4) *The challenge in application.* Although the skeleton data is compact and lightweight, in order to meet the application in practical operation, it is still necessary to compress the computing resources of the recognition task and provide them to the upstream pose estimation task. In GCNs, [70,182] adopt the lightweight strategy in CNN, [68] changes the original input topology to adapt to different data scales, and [64] pre-fuses the input modalities to reduce the overall computational cost. The Transformer algorithm needs to calculate the global association relationship, which causes a lot of computational overhead. In addition, the generalization ability of the model is highly required for complex real-world scenarios. Therefore, how to lightweight the model while ensuring the classification accuracy is a serious challenge to face.

In view of the above challenges, we summarize a series of future directions for researchers in action recognition:

(1) *Multi-person and multi-action Skeleton Transformer.* The application of the model in the real world is very different from the testing in the experimental environment. In real life, the algorithm is more oriented to the scene of multi-person and multi-action recognition. Researchers can refer to multi-person and multi-action recognition methods in RGB video surveillance and social activities. It takes the advantages of transformer in processing multi-dimensional skeleton sequences and solves the dual difficulties of multi-person and multi-action recognition, which greatly promotes the development of skeleton action recognition.

(2) *Robust Transformers for Skeleton-based Action Recognition.* The number of joint points in different datasets is different, and how to overcome the limitations of different joint acquisition rules is a challenging research. The Transformer structure is designed to adapt to the dataset to improve the topology robustness. In addition, complex actions in the face of complex scenes will make the intra-class difference larger and the inter-class difference smaller. By exploiting the potential of transformer in global information acquisition ability, the discriminative power of key features is enhanced and the recognition robustness is improved.

(3) *Multimodal Action Recognition with Transformers.* Most of the existing multi-stream methods calculate each stream independently and fuse the accuracy at the end. This strategy produces a great waste of resources and is not conducive to modal augmentation. Using the multimodal interaction ability of transformer, RGB, optical flow, joint, skeleton and other modalities are fused to explore a comprehensive transformer-based framework. It is conducive to improving the computational efficiency and information fusion ability of the multimodal model, reducing the time cost and improving the final result.

(4) *Interpretability of Hybrid Transformers for Skeleton-based Action Recognition.* Although Hybrid Transformer fuses multiple architectures and improves performance, existing methods lack reliable validity explanations. Specifically, the location and number of Transformer, CNN, and GCN applications can only be adjusted by experiments, which greatly wastes computing resources. Therefore, it is necessary to further study the theoretical support of the hybrid model, further explore the processing methods of features between different models, and reflect the various advantages between structures. In addition, Transformer has the ability of global information acquisition and interaction, but it is not good at exploring local details. Researchers can design the adaptive correlation matrix of the fusion transformer, combine the related technologies of CNN and GCN, and improve the global and local discrimination from the bottom of module.

(5) *Action Recognition with Lightweight Transformers for Skeletons.* The model needs to balance parameters and accuracy in practical applications. Researchers can refer to the Transformer lightweight method in RGB images and ignore the modules in the structure that have a small impact on the result. Transformers can be used to flexibly process data with different levels of complexity and break through the limitations of traditional fixed topologies. The lightweight technology is transferred to the Transformer-based skeleton action recognition task to greatly reduce the computational cost while ensuring the accuracy.

(6) *Domain Adaptation Skeleton-based Action Recognition with Transformers.* In practical surveillance video analysis, action recognition modules are usually trained in one environment (source domain) and used in another environment with different viewpoints and features (target domain). This task, known as domain adapted action recognition, is becoming a popular research topic [183]. Most methods apply the fully supervised learning paradigm, where training and testing data are drawn from the same domain. However, action labels are only available on the source dataset, but performance evaluation cannot be performed on the target dataset [184]. Therefore, there is a lack of exploration of this field in the UDA setting. Furthermore, graph convolutional networks (GCNs) have made rapid progress in various graph-based data-rich video analysis tasks. However, exploration of transferable knowledge between different graphs, a direction with broad and potential applications, has rarely been studied [185]. In addition, transformer-based domain adaptation methods have achieved excellent results in RGB-based action recognition tasks [186], but transformers still lack effective application in skeleton data. In summary, we believe that Domain Adaptation Skeleton-based Action Recognition with Transformers is a direction with great potential.

(7) *Adversarial Attack on Skeleton-based Action Recognition with Transformers.* Human skeleton-based action recognition has received increasing attention due to its potential wide range of applications [187]. However, in real life, a potential enemy may easily fool an action recognition model by performing actions with imperceptible perturbations [188–191]. Deploying such a model without understanding its adversarial vulnerabilities can lead to severe consequences [192]. [193–195] have shown that a general vulnerability of skeleton-based HAR is found in various classifiers and data. Despite these security issues, there are still few studies on the vulnerability of human skeleton-based action recognition. To the best of our knowledge, no transformer-based approach has been adopted for skeleton-based adversarial learning, so this direction has strong potential.

(8) *Continual Learning on Skeleton-based Action Recognition with Transformers.* The human brain can transfer new knowledge into long-term memory and avoid forgetting. Moreover, humans do not need to retrain using historical information when learning each new knowledge. In deep learning, con-

tinuous learning refers to the way a system seamlessly learns from a continuous stream of information while preventing catastrophic forgetting, that is, situations where new incoming information strongly interferes with previously learned representations [196]. Continuous learning for action recognition refers to the continuous learning of various types of new actions over time. However, the difference between the previously learned action and the new action to be learned can lead to catastrophic forgetting. At present, there are only few skeleton-based action recognition methods for this challenging task [197,198], especially no transformer-based method. We believe this is a very promising research direction.

7. Conclusion

Graph convolution and Transformer are two of the most popular technologies for action recognition, which have powerful skeleton data information interaction capabilities and topological structure extraction capabilities. Through analysis, to break through the limitations of the traditional ST-GCN structure, many GCN-based methods apply the self-attention mechanism or the correlation matrix auxiliary network in the spatio-temporal processing. In this study, we explore Transformer-style GCN and the standard Transformer skeleton action recognition and summarize these two parts into the Transformer-based skeleton action recognition. In addition, we investigate unsupervised and transfer learning methods for transformer-based skeletal action recognition, and explore the impact of global dependency association relationships in clustering, self-attention, few-/one-shot learning, etc.

In this survey, our main contents include: (1) We summarize the state-of-the-art skeleton transformers and propose the corresponding taxonomy. (2) The Transformer-style GCN algorithm is summarized and the corresponding classification method is proposed. (3) The unsupervised and transfer learning algorithms with Transformers are summarized, and the corresponding classification methods are proposed. (4) Comprehensively expounds the introduction, innovation, and improvement of the papers in the above five classifications. (5) This paper collects and compares the accuracy of a large number of Transformer-based skeleton-based action recognition algorithms. And we provide a new data preprocessing method that can be easily replaced and used in the source code. (6) The challenges faced by skeleton action recognition and the future research direction of the transformer method in this field are introduced, which provides some reference and help for the majority of researchers.

Declaration of Competing Interest

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

References

[1] R. Singh, A.K.S. Kushwaha, R. Srivastava, Multi-view recognition system for human activity based on multiple features for video surveillance system, Multimedia Tools Appl. 78 (12) (2019) 17165–17196.

[2] A. Prati, C. Shan, K.I.-K. Wang, Sensors, vision and networks: From video surveillance to activity recognition and health monitoring, J. Ambient Intell. Smart Environ. 11 (1) (2019) 5–22.

[3] M. Shorfuzzaman, M.S. Hossain, M.F. Alhamid, Towards the sustainable development of smart cities through mass video surveillance: A response to the covid-19 pandemic, Sustain. Cities Soc. 64 (2021).

[4] N. Khalid, M. Gochoo, A. Jalal, K. Kim, Modeling two-person segmentation and locomotion for stereoscopic action identification: a sustainable video surveillance system, Sustainability 13 (2) (2021) 970.

[5] J. Yang, M. Xi, B. Jiang, J. Man, Q. Meng, B. Li, Fadn: fully connected attitude detection network based on industrial video, IEEE Trans. Industr. Inf. 17 (3) (2020) 2011–2020.

[6] T. Liu, Y.-F. Li, H. Liu, Z. Zhang, S. Liu, Risir: Rapid infrared spectral imaging restoration model for industrial material detection in intelligent video systems, IEEE Trans. Industr. Inf. (2019).

[7] R. Kumar, R. Tripathi, N. Marchang, G. Srivastava, T.R. Gadekallu, N.N. Xiong, A secured distributed detection system based on ipfs and blockchain for industrial image and video data security, J. Parallel Distrib. Comput. 152 (2021) 128–143.

[8] C. Dai, X. Liu, H. Xu, L.T. Yang, M.J. Deen, Hybrid deep model for human behavior understanding on industrial internet of video things, IEEE Trans. Industr. Inf. 18 (10) (2021) 7000–7008.

[9] D. Liu, Y. Cui, Y. Chen, J. Zhang, B. Fan, Video object detection for autonomous driving: Motion-aid feature calibration, Neurocomputing 409 (2020) 1–11.

[10] M. Siam, A. Kendall, M. Jagersand, Video class agnostic segmentation benchmark for autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2825–2834.

[11] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, R. Yang, The apolloscape open dataset for autonomous driving and its application, IEEE Trans. Pattern Anal. Mach. Intell. 42 (10) (2019) 2702–2719.

[12] P. Li, J. Jin, Time3d: End-to-end joint monocular 3d object detection and tracking for autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3885–3894.

[13] S. Wan, X. Xu, T. Wang, Z. Gu, An intelligent video analysis method for abnormal event detection in intelligent transportation systems, IEEE Trans. Intell. Transp. Syst. 22 (7) (2020) 4487–4495.

[14] J. Liang, H. Zhu, E. Zhang, J. Zhang, Stargazer: A transformer-based driver action detection system for intelligent transportation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3160–3167.

[15] P. Sharma, A. Singh, K.K. Singh, A. Dhull, Vehicle identification using modified region based convolution network for intelligent transportation system, Multimedia Tools Appl. 81 (24) (2022) 34893–34917.

[16] M.H. Alkinani, A.A. Almazroi, M. Adhikari, V.G. Menon, Design and analysis of logistic agent-based swarm-neural network for intelligent transportation system, Alexandria Eng. J. 61 (10) (2022) 8325–8334.

[17] S. Nayak, B. Nagesh, A. Routray, M. Sarma, A human–computer interaction framework for emotion recognition through time-series thermal video sequences, Comput. Electr. Eng. 93 (2021).

[18] M. Kashef, A. Visvizi, O. Troisi, Smart city as a smart service system: Human-computer interaction and smart city surveillance systems, Comput. Hum. Behav. 124 (2021).

[19] T. Vuletic, A. Duffy, L. Hay, C. McTeague, G. Campbell, M. Grealy, Systematic literature review of hand gestures used in human computer interaction interfaces, Int. J. Hum Comput Stud. 129 (2019) 74–94.

[20] A. Kashevnik, A. Ponomarev, N. Shilov, A. Chechulin, Threats detection during human-computer interaction in driver monitoring systems, Sensors 22 (6) (2022) 2380.

[21] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Advances in neural information processing systems 27 (2014).

[22] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6202–6211.

[23] S. Buch, V. Escorcia, C. Shen, B. Ghanem, J. Carlos Niebles, Sst: Single-stream temporal action proposals, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 2911–2920.

[24] G. Varol, I. Laptev, C. Schmid, Long-term temporal convolutions for action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2017) 1510–1517.

[25] C. Lu, J. Jia, C.-K. Tang, Range-sample depth feature for action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 772–779.

[26] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, P.O. Ogunbona, Action recognition from depth maps using deep convolutional neural networks, IEEE Trans. Human-Mach. Syst. 46 (4) (2015) 498–509.

[27] C. Chen, K. Liu, N. Kehtarnavaz, Real-time human action recognition based on depth motion maps, J. Real-time Image Process. 12 (1) (2016) 155–163.

[28] H. Basak, R. Kundu, P.K. Singh, M.F. Ijaz, M. Woźniak, R. Sarkar, A union of deep learning and swarm-based optimization for 3d human action recognition, Scientific Rep. 12 (1) (2022) 1–17.

[29] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, W. Zhang, Optical flow guided feature: A fast and robust motion representation for video action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1390–1399.

[30] L. Wang, P. Koniusz, D.Q. Huynh, Hallucinating idt descriptors and i3d optical flow features for action recognition with cnns, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8698–8708.

[31] J. Chen, C.M. Ho, Mm-vit: Multi-modal video transformer for compressed video action recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1910–1921.

[32] E. Kazakos, A. Nagrani, A. Zisserman, D. Damen, Epic-fusion: Audio-visual temporal binding for egocentric action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5492–5501.

[33] Y. Li, T. Tu, H. Zhang, J. Li, Z. Jin, Q. Wen, Sound can help us see more clearly, Sensors 22 (2) (2022) 599.
[34] L.L. Presti, M. La Cascia, 3d skeleton-based human action classification: A survey, Pattern Recogn. 53 (2016) 130–147.
[35] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7291–7299.
[36] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 5693–5703.
[37] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1010–1019.
[38] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, A.C. Kot, Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding, IEEE Trans. Pattern Anal. Mach. Intell. 42 (10) (2019) 2684–2701.
[39] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Thirty-second AAAI conference on artificial intelligence, 2018.
[40] X. Wu, D. Xu, L. Duan, J. Luo, Action recognition using context and appearance distribution features, in: CVPR 2011, IEEE, 2011, pp. 489–496.
[41] L. Fan, S. Buch, G. Wang, R. Cao, Y. Zhu, J.C. Niebles, L. Fei-Fei, Rubiksnet: Learnable 3d-shift for efficient video action recognition, in: European Conference on Computer Vision, Springer, 2020, pp. 505–521.
[42] R. Girdhar, J. Carreira, C. Doersch, A. Zisserman, Video action transformer network, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 244–253.
[43] Y. Kong, Y. Fu, Human action recognition and prediction: A survey, Int. J. Comput. Vision 130 (5) (2022) 1366–1401.
[44] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, J. Liu, Human action recognition from various data modalities: A review, IEEE Trans. Pattern Anal. Mach. Intell. (2022).
[45] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1290–1297.
[46] B. Ren, M. Liu, R. Ding, H. Liu, A survey on 3d skeleton-based action recognition using learning method, arXiv preprint arXiv:2002.05907 (2020).
[47] H. Wang, L. Wang, Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 499–508.
[48] C. Caetano, J. Sena, F. Brémond, J.A. Dos Santos, W.R. Schwartz, Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition, in: 2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS), IEEE, 2019, pp. 1–8.
[49] C. Plizzari, M. Cannici, M. Matteucci, Spatial temporal transformer network for skeleton-based action recognition, in: International Conference on Pattern Recognition, Springer, 2021, pp. 694–701.
[50] T. Ahmad, L. Jin, X. Zhang, S. Lai, G. Tang, L. Lin, Graph convolutional neural network for human action recognition: A comprehensive survey, IEEE Trans. Artif. Intell. 2 (2) (2021) 128–145.
[51] L. Feng, Y. Zhao, W. Zhao, J. Tang, A comparative review of graph convolutional networks for human skeleton-based action recognition, Artif. Intell. Rev. (2021) 1–31.
[52] R. Yue, Z. Tian, S. Du, Action recognition based on rgb and skeleton data sets: A survey, Neurocomputing (2022).
[53] A. Ulhaq, N. Akhtar, G. Pogrebna, A. Mian, Vision transformers for action recognition: A survey, arXiv preprint arXiv:2209.05700 (2022).
[54] Y. Xing, J. Zhu, Deep learning-based action recognition with 3d skeleton: a survey (2021).
[55] Z. Qin, Y. Liu, M. Perera, T. Gedeon, P. Ji, D. Kim, S. Anwar, Anubis: Skeleton action recognition dataset, review, and benchmark., CoRR (2022).
[56] W. Zhang, Z. Dong, J. Liu, Q. Yan, C. Xiao, et al., Point cloud completion via skeleton-detail transformer, IEEE Trans. Visual Comput. Graphics (2022).
[57] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 12026–12035.
[58] F. Ye, S. Pu, Q. Zhong, C. Li, D. Xie, H. Tang, Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 55–63.
[59] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, H. Lu, Decoupling gcn with dropgraph module for skeleton-based action recognition, in: European Conference on Computer Vision, Springer, 2020, pp. 536–553.
[60] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, W. Hu, Channel-wise topology refinement graph convolution for skeleton-based action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13359–13368.
[61] L. Ke, K.-C. Peng, S. Lyu, Towards to-at spatio-temporal focus for skeleton-based action recognition, arXiv preprint arXiv:2202.02314 (2022).
[62] L. Hu, S. Liu, W. Feng, Spatial temporal graph attention network for skeleton-based action recognition, arXiv preprint arXiv:2208.08599 (2022).
[63] R. Qian, J. Wang, J. Wang, S. Liang, Structural attention for channel-wise adaptive graph convolution in skeleton-based action recognition, in: 2022 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2022, pp. 01–06.
[64] Y.-F. Song, Z. Zhang, C. Shan, L. Wang, Constructing stronger and faster baselines for skeleton-based action recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2022).
[65] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 143–152.
[66] B. Degardin, V. Lopes, H. Proença, Regina–reasoning graph convolutional networks in human action recognition, IEEE Trans. Inf. Forensics Secur. 16 (2021) 5442–5451.
[67] J. Lee, M. Lee, D. Lee, S. Lee, Hierarchically decomposed graph convolutional networks for skeleton-based action recognition, arXiv preprint arXiv:2208.10741 (2022).
[68] L. Shi, Y. Zhang, J. Cheng, H. Lu, Adasgn: Adapting joint number and model size for efficient skeleton-based action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13413–13422.
[69] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, N. Zheng, Semantics-guided neural networks for efficient skeleton-based human action recognition, in: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 1112–1121.
[70] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, H. Lu, Skeleton-based action recognition with shift graph convolutional network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 183–192.
[71] L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-based action recognition with multi-stream adaptive graph convolutional networks, IEEE Trans. Image Process. 29 (2020) 9532–9545.
[72] Z. Qin, Y. Liu, P. Ji, D. Kim, L. Wang, B. McKay, S. Anwar, T. Gedeon, Leveraging third-order features in skeleton-based action recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021.
[73] Z. Chen, S. Li, B. Yang, Q. Li, H. Liu, Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 1113–1122.
[74] H. Xing, D. Burschka, Skeletal human action recognition using hybrid attention based graph convolutional network, in: 2022 26th International Conference on Pattern Recognition (ICPR), IEEE, 2022, pp. 3333–3340.
[75] R. Hou, Y. Li, N. Zhang, Y. Zhou, X. Yang, Z. Wang, Shifting perspective to see difference: A novel multi-view method for skeleton based action recognition, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 4987–4995.
[76] H. Chi, M.H. Ha, S. Chi, S.W. Lee, Q. Huang, K. Ramani, Infogcn: Representation learning for human skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20186–20196.
[77] S. Yang, X. Wang, L. Gao, J. Song, Mke-gcn: Multi-modal knowledge embedded graph convolutional network for skeleton-based action recognition in the wild, in: 2022 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2022, pp. 01–06.
[78] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[79] A. Dosovitskiy, L. Beyer, e. a. Kolesnikov, An image is worth 16x16 words: Transformers for image recognition at scale, ICLR (2020).
[80] P. Pareek, A. Thakkar, A survey on video-based human action recognition: recent updates, datasets, challenges, and applications, Artif. Intell. Rev. 54 (3) (2021) 2259–2322.
[81] T. Özyer, D.S. Ak, R. Alhajj, Human action recognition approaches with video datasets–a survey, Knowl.-Based Syst. 222 (2021).
[82] Y. Su, G. Lin, Q. Wu, Self-supervised 3d skeleton action representation learning with motion consistency and continuity, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 13328–13338.
[83] W. You, X. Wang, View enhanced jigsaw puzzle for self-supervised feature learning in 3d human action recognition, IEEE Access 10 (2022) 36385–36396.
[84] H. Yao, S.-J. Zhao, C. Xie, K. Ye, S. Liang, Recurrent graph convolutional autoencoder for unsupervised skeleton-based action recognition, in: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2021, pp. 1–6.
[85] H. Zhang, Y. Hou, W. Zhang, Skeletal twins: Unsupervised skeleton-based action representation learning, in: 2022 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2022, pp. 1–6.
[86] X. Shu, B. Xu, L. Zhang, J. Tang, Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2022).
[87] B. Xu, X. Shu, Y. Song, X-invariant contrastive augmentation and representation learning for semi-supervised skeleton-based action recognition, IEEE Trans. Image Process. (2022).
[88] H. Zhang, Y. Hou, W. Zhang, W. Li, Contrastive positive mining for unsupervised 3d action representation learning, in: European Conference on Computer Vision, Springer, 2022, pp. 36–51.
[89] S. Xu, H. Rao, X. Hu, J. Cheng, B. Hu, Prototypical contrast and reverse prediction: Unsupervised skeleton based action recognition, IEEE Trans. Multimedia (2021).

[90] G. Paoletti, J. Cavazza, C. Beyan, A. Del Bue, Subspace clustering for action recognition with covariance representations and temporal pruning, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 6035–6042.
[91] K. Su, X. Liu, E. Shlizerman, Predict & cluster: Unsupervised skeleton based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9631–9640.
[92] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, F. Brémond, Self-supervised video pose representation learning for occlusion-robust action recognition, in: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), IEEE, 2021, pp. 1–5.
[93] Y. Su, G. Lin, R. Sun, Y. Hao, Q. Wu, Modeling the uncertainty for self-supervised 3d skeleton action representation learning, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 769–778.
[94] A.B. Tanfous, A. Zerroug, D. Linsley, T. Serre, How and what to learn: Taxonomizing self-supervised learning for 3d action recognition, in: WACV, 2022, pp. 2888–2897.
[95] P. Wang, J. Wen, C. Si, Y. Qian, L. Wang, Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition, IEEE Trans. Image Process. 31 (2022) 6224–6238.
[96] J. Zhang, L. Lin, J. Liu, Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations, arXiv preprint arXiv:2211.13466 (2022).
[97] X. Gao, Y. Yang, Y. Zhang, M. Li, J.-G. Yu, S. Du, Efficient spatio-temporal contrastive learning for skeleton-based 3d action recognition, IEEE Trans. Multimedia (2021).
[98] H. Rao, S. Xu, X. Hu, J. Cheng, B. Hu, Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition, Inf. Sci. 569 (2021) 90–109.
[99] F.M. Thoker, H. Doughty, C.G. Snoek, Skeleton-contrastive 3d action representation learning, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 1655–1663.
[100] S. Das, M.S. Ryoo, Stc-mix: Space, time, channel mixing for self-supervised video representation, arXiv preprint arXiv:2112.03906 (2021).
[101] R. Brinzea, B. Khaertdinov, S. Asteriadis, Contrastive learning with cross-modal knowledge mining for multimodal human activity recognition, arXiv preprint arXiv:2205.10071 (2022).
[102] Z. Chen, H. Liu, T. Guo, Z. Chen, P. Song, H. Tang, Contrastive learning from spatio-temporal mixed skeleton sequences for self-supervised skeleton-based action recognition, arXiv preprint arXiv:2207.03065 (2022).
[103] B. Khaertdinov, S. Asteriadis, Temporal feature alignment in contrastive self-supervised learning for human activity recognition, arXiv preprint arXiv:2210.03382 (2022).
[104] N. Lingg, M. Sarabia, L. Zappella, B.-J. Theobald, Contrastive self-supervised learning for skeleton representations, arXiv preprint arXiv:2211.05304 (2022).
[105] T. Guo, H. Liu, Z. Chen, M. Liu, T. Wang, R. Ding, Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 762–770.
[106] X. Gao, Y. Yang, S. Du, Contrastive self-supervised learning for skeleton action recognition, in: NeurIPS 2020 Workshop on Pre-registration in Machine Learning, PMLR, 2021, pp. 51–61.
[107] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, W. Zhang, 3d human action representation learning via cross-view consistency pursuit, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4741–4750.
[108] B. Kim, H.J. Chang, J. Kim, J.Y. Choi, Global-local motion transformer for unsupervised skeleton-based action learning, in: European Conference on Computer Vision, Springer, 2022, pp. 209–225.
[109] Y.-B. Cheng, X. Chen, J. Chen, P. Wei, D. Zhang, L. Lin, Hierarchical transformer: Unsupervised representation learning for skeleton-based human action recognition, in: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2021, pp. 1–6.
[110] M. Endo, K.L. Poston, E.V. Sullivan, L. Fei-Fei, K.M. Pohl, E. Adeli, Gaitforemer: Self-supervised pre-training of transformers via human motion forecasting for few-shot gait impairment severity estimation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 130–139.
[111] Y. Chen, L. Zhao, J. Yuan, Y. Tian, Z. Xia, S. Geng, L. Han, D.N. Metaxas, Hierarchically self-supervised transformer for human skeleton representation learning, in: European Conference on Computer Vision, Springer, 2022, pp. 185–202.
[112] Y.-B. Cheng, X. Chen, D. Zhang, L. Lin, Motion-transformer: self-supervised pre-training for skeleton-based action recognition, in: Proceedings of the 2nd ACM International Conference on Multimedia in Asia, 2021, pp. 1–6.
[113] W. Wu, Y. Hua, S. Wu, C. Chen, A. Lu, et al., Skeletonmae: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition, arXiv preprint arXiv:2209.02399 (2022).
[114] Y. Xu, C. Han, J. Qin, X. Xu, G. Han, S. He, Transductive zero-shot action recognition via visually connected graph convolutional networks, IEEE Trans. Neural Networks Learn. Syst. 32 (8) (2020) 3761–3769.
[115] P. Gupta, D. Sharma, R.K. Sarvadevabhatla, Syntactically guided generative embeddings for zero-shot skeleton action recognition, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 2021, pp. 439–443.
[116] A. Zhu, Q. Ke, M. Gong, J. Bailey, Adaptive local-component-aware graph convolutional network for one-shot skeleton-based action recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 6038–6047.
[117] B. Jasani, A. Mazagonwalla, Skeleton based zero shot action recognition in joint pose-language semantic space, arXiv preprint arXiv:1911.11344 (2019).
[118] B. Shi, L. Wang, Z. Yu, S. Xiang, T. Liu, Y. Fu, Zero-shot learning for skeleton-based classroom action recognition, in: 2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC), IEEE, 2021, pp. 82–86.
[119] L. Xu, Q. Wang, X. Lin, L. Yuan, X. Ma, Temporal-spatial feature fusion for few-shot skeleton-based action recognition, in: IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society, IEEE, 2022, pp. 1–6.
[120] M.-R. Tseng, A. Gupta, C.-K. Tang, Y.-W. Tai, Haa4d: Few-shot human atomic action recognition via 3d spatio-temporal skeletal alignment, arXiv preprint arXiv:2202.07308 (2022).
[121] Z. Li, X. Gong, R. Song, P. Duan, J. Liu, W. Zhang, Smam: Self and mutual adaptive matching for skeleton-based few-shot action recognition, IEEE Trans. Image Process. 32 (2022) 392–402.
[122] L. Xu, Q. Wang, X. Lin, L. Yuan, An efficient framework for few-shot skeleton-based temporal action segmentation, arXiv preprint arXiv:2207.09925 (2022).
[123] A. Sabater, L. Santos, J. Santos-Victor, A. Bernardino, L. Montesano, A.C. Murillo, One-shot action recognition in challenging therapy scenarios, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2777–2785.
[124] S. Berti, A. Rosasco, M. Colledanchise, L. Natale, One-shot open-set skeleton-based action recognition, in: 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids), IEEE, 2022, pp. 765–772.
[125] T. Chen, D. Zhou, J. Wang, S. Wang, Q. He, C. Hu, E. Ding, Y. Guan, X. He, Part-aware prototypical graph network for one-shot skeleton-based action recognition, arXiv preprint arXiv:2208.09150 (2022).
[126] K. Peng, A. Roitberg, K. Yang, J. Zhang, R. Stiefelhagen, Delving deep into one-shot skeleton-based action recognition with diverse occlusions, IEEE Trans. Multimedia (2023).
[127] L. Wang, P. Koniusz, Temporal-viewpoint transportation plan for skeletal few-shot action recognition, in: Proceedings of the Asian Conference on Computer Vision, 2022, pp. 4176–4193.
[128] X. Chen, S. Xie, K. He, An empirical study of training self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9640–9649.
[129] T. Lin, Y. Wang, X. Liu, X. Qiu, A survey of transformers, AI Open, 2022.
[130] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[131] Q. Wang, J. Peng, S. Shi, T. Liu, J. He, R. Weng, Iip-transformer: Intra-inter-part transformer for skeleton-based action recognition, arXiv preprint arXiv:2110.13385 (2021).
[132] Y. Sun, Y. Shen, L. Ma, Msst-rt: Multi-stream spatial-temporal relative transformer for skeleton-based action recognition, Sensors 21 (16) (2021) 5339.
[133] M. Ijaz, R. Diaz, C. Chen, Multimodal transformer for nursing activity recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2065–2074.
[134] L. Shi, Y. Zhang, J. Cheng, H. Lu, Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition, in: Proceedings of the Asian Conference on Computer Vision, 2020.
[135] Y. Zhang, B. Wu, W. Li, L. Duan, C. Gan, Stst: Spatial-temporal specialized transformer for skeleton-based action recognition, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 3229–3237.
[136] H. Zhang, H. Geng, G. Yang, Two-stream transformer encoders for skeleton-based action recognition, in: International Conference on Computing, Control and Industrial Engineering, Springer, 2022, pp. 272–281.
[137] F. Shi, C. Lee, L. Qiu, Y. Zhao, T. Shen, S. Muralidhar, T. Han, S.-C. Zhu, V. Narayanan, Star: Sparse transformer-based action recognition, arXiv preprint arXiv:2107.07089 (2021).
[138] Y. Pang, Q. Ke, H. Rahmani, J. Bailey, J. Liu, Igformer: Interaction graph transformer for skeleton-based human interaction recognition, in: European Conference on Computer Vision, Springer, 2022, pp. 605–622.
[139] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, M. Chiaberge, Action transformer: A self-attention model for short-time pose-based human action recognition, Pattern Recogn. 124 (2022).
[140] D. Ahn, S. Kim, H. Hong, B.C. Ko, Star-transformer: A spatio-temporal cross attention transformer for human action recognition, arXiv preprint arXiv:2210.07503 (2022).
[141] Y. Zhou, C. Li, Z.-Q. Cheng, Y. Geng, X. Xie, M. Keuper, Hypergraph transformer for skeleton-based action recognition, arXiv preprint arXiv:2211.09590 (2022).
[142] L. Yuan, Z. He, Q. Wang, L. Xu, X. Ma, Spatial transformer network with transfer learning for small-scale fine-grained skeleton-based tai chi action recognition, arXiv preprint arXiv:2206.15002 (2022).
[143] H. Qiu, B. Hou, B. Ren, X. Zhang, Spatio-temporal tuples transformer for skeleton-based action recognition, arXiv preprint arXiv:2201.02849 (2022).
[144] J. Zhang, Y. Jia, W. Xie, Z. Tu, Zoom transformer for skeleton-based group activity recognition, IEEE Trans. Circuits Syst. Video Technol. (2022).
[145] Y. Jiang, Z. Sun, S. Yu, S. Wang, Y. Song, A graph skeleton transformer network for action recognition, Symmetry 14 (8) (2022) 1547.

[146] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, F. Bremond, Unik: A unified framework for real-world skeleton-based action recognition, arXiv preprint arXiv:2107.08580 (2021).
[147] Z. Gao, P. Wang, P. Lv, X. Jiang, Q. Liu, P. Wang, M. Xu, W. Li, Focal and global spatial-temporal transformer for skeleton-based action recognition, in: Proceedings of the Asian Conference on Computer Vision, 2022, pp. 382–398.
[148] Y. Liu, H. Zhang, D. Xu, K. He, Graph transformer network with temporal kernel attention for skeleton-based action recognition, Knowl.-Based Syst. 240 (2022).
[149] S. Chen, K. Xu, X. Jiang, T. Sun, Pyramid spatial-temporal graph transformer for skeleton-based action recognition, Appl. Sci. 12 (18) (2022) 9229.
[150] Y. Meng, M. Shi, W. Yang, Skeleton action recognition based on transformer adaptive graph convolution, in: Journal of Physics: Conference Series, Vol. 2170, IOP Publishing, 2022, p. 012007.
[151] J. Kong, Y. Bian, M. Jiang, Mtt: Multi-scale temporal transformer for skeleton-based action recognition, IEEE Signal Process. Lett. 29 (2022) 528–532.
[152] R. Bai, M. Li, B. Meng, F. Li, M. Jiang, J. Ren, D. Sun, Hierarchical graph convolutional skeleton transformer for action recognition, in: 2022 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2022, pp. 01–06.
[153] T. Alsarhan, U. Ali, H. Lu, Enhanced discriminative graph convolutional network with adaptive temporal modelling for skeleton-based action recognition, Comput. Vis. Image Underst. 216 (2022).
[154] N. Sun, L. Leng, J. Liu, G. Han, Multi-stream slowfast graph convolutional networks for skeleton-based action recognition, Image Vis. Comput. 109 (2021).
[155] J. Xie, Q. Miao, R. Liu, W. Xin, L. Tang, S. Zhong, X. Gao, Attention adjacency matrix based graph convolutional networks for skeleton-based action recognition, Neurocomputing 440 (2021) 230–239.
[156] Z. Hu, Z. Pan, Q. Wang, L. Yu, S. Fei, Forward-reverse adaptive graph convolutional networks for skeleton-based action recognition, Neurocomputing 492 (2022) 624–636.
[157] A. Zeng, X. Sun, L. Yang, N. Zhao, M. Liu, Q. Xu, Learning skeletal graph neural networks for hard 3d pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11436–11445.
[158] F. Li, A. Zhu, Y. Xu, R. Cui, G. Hua, Multi-stream and enhanced spatial-temporal graph convolution network for skeleton-based action recognition, IEEE Access 8 (2020) 97757–97770.
[159] J. Zhang, G. Ye, Z. Tu, Y. Qin, Q. Qin, J. Zhang, J. Liu, A spatial attentive and temporal dilated (satd) gcn for skeleton-based action recognition, CAAI Trans. Intell. Technol. 7 (1) (2022) 46–55.
[160] J. Xie, W. Xin, R. Liu, Q. Miao, L. Sheng, L. Zhang, X. Gao, Global co-occurrence feature and local spatial feature learning for skeleton-based action recognition, Entropy 22 (10) (2020) 1135.
[161] A.R. Shahid, H. Yan, Skeleton-based action recognition with adaptive and self-attentive graph convolution network (2022).
[162] J. Li, X. Xie, Z. Zhao, Y. Cao, Q. Pan, G. Shi, Temporal graph modeling for skeleton-based action recognition, arXiv preprint arXiv:2012.08804 (2020).
[163] R. Bai, M. Li, B. Meng, F. Li, J. Ren, M. Jiang, D. Sun, Gcst: Graph convolutional skeleton transformer for action recognition, arXiv preprint arXiv:2109.02860 (2021).
[164] Q. Huang, F. Zhou, J. He, Y. Zhao, R. Qin, Spatial–temporal graph attention networks for skeleton-based action recognition, J. Electron. Imaging 29 (5) (2020).
[165] J. Gao, T. He, X. Zhou, S. Ge, Skeleton-based action recognition with focusing-diffusion graph convolutional networks, IEEE Signal Process. Lett. 28 (2021) 2058–2062.
[166] R. Hou, Z. Wang, Self-attention based anchor proposal for skeleton-based action recognition, arXiv preprint arXiv:2112.09413 (2021).
[167] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.
[168] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset, arXiv:1705.06950 (2017).
[169] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, Z. Li, Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 16266–16275.
[170] C. Liu, Y. Hu, Y. Li, S. Song, J. Liu, Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding, arXiv preprint arXiv:1703.07475 (2017).
[171] J. Wang, X. Nie, Y. Xia, Y. Wu, S.-C. Zhu, Cross-view action modeling, learning and recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 2649–2656.
[172] H. Rahmani, A. Mahmood, D.Q Huynh, A. Mian, Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part II 13, Springer, 2014, pp. 742–757.
[173] K. Yun, J. Honorio, D. Chattopadhyay, T.L. Berg, D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning, in: 2012 IEEE computer society conference on computer vision and pattern recognition workshops, IEEE, 2012, pp. 28–35.
[174] W. Liu, Q. Bao, Y. Sun, T. Mei, Recent advances of monocular 2d and 3d human pose estimation: a deep learning perspective, ACM Comput. Surv. 55 (4) (2022) 1–41.
[175] Z. Tu, J. Zhang, H. Li, Y. Chen, J. Yuan, Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition, IEEE Trans. Multimedia (2022).
[176] C. Si, X. Nie, W. Wang, L. Wang, T. Tan, J. Feng, Adversarial self-supervised learning for semi-supervised 3d action recognition, in: European Conference on Computer Vision, Springer, 2020, pp. 35–51.
[177] J.T.S. Phang, K.H. Lim, Real-time multi-camera multi-person action recognition using pose estimation, in: Proceedings of the 3rd international conference on machine learning and soft computing, 2019, pp. 175–180.
[178] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, S. Savarese, Social scene understanding: End-to-end multi-person action localization and collective activity recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4315–4324.
[179] X. Liang, H.-B. Zhang, Y.-X. Zhang, J.-L. Huang, Jtcr: joint trajectory character recognition for human action recognition, in: 2019 IEEE Eurasia conference on IOT, communication and engineering (ECICE), IEEE, 2019, pp. 350–353.
[180] A. Gilbert, J. Illingworth, R. Bowden, Fast realistic multi-action recognition using mined dense spatio-temporal features, in: 2009 IEEE 12th international conference on computer vision, IEEE, 2009, pp. 925–931.
[181] X. Shu, J. Tang, G.-J. Qi, Y. Song, Z. Li, L. Zhang, Concurrence-aware long short-term sub-memories for person-person action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1–8.
[182] K. Cheng, Y. Zhang, X. He, J. Cheng, H. Lu, Extremely lightweight skeleton-based action recognition with shiftgcn++, IEEE Trans. Image Process. 30 (2021) 7333–7348.
[183] Z. Gao, Y. Zhao, H. Zhang, D. Chen, A.-A. Liu, S. Chen, A novel multiple-view adversarial learning network for unsupervised domain adaptation action recognition, IEEE Trans. Cybern. (2021).
[184] Y. Tang, X. Liu, X. Yu, D. Zhang, J. Lu, J. Zhou, Learning from temporal spatial cubism for cross-dataset skeleton-based action recognition, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 18 (2) (2022) 1–24.
[185] Y. Tang, Y. Wei, X. Yu, J. Lu, J. Zhou, Graph interaction networks for relation transfer in human activity videos, IEEE Trans. Circuits Syst. Video Technol. 30 (9) (2020) 2872–2886.
[186] V.G.T. da Costa, G. Zara, P. Rota, T. Oliveira-Santos, N. Sebe, V. Murino, E. Ricci, Unsupervised domain adaptation for video transformers in action recognition, in: 2022 26th International Conference on Pattern Recognition (ICPR), IEEE, 2022, pp. 1258–1265.
[187] H. Wang, F. He, Z. Peng, T. Shao, Y.-L. Yang, K. Zhou, D. Hogg, Understanding the robustness of skeleton-based action recognition under adversarial attack, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14656–14665.
[188] Y. Diao, T. Shao, Y.-L. Yang, K. Zhou, H. Wang, Basar: Black-box attack on skeletal action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7597–7607.
[189] J. Liu, N. Akhtar, A. Mian, Adversarial attack on skeleton-based human action recognition, IEEE Trans. Neural Networks Learn. Syst. (2020).
[190] D. Kumar, C. Kumar, C.W. Seah, S. Xia, M. Shao, Finding achilles' heel: Adversarial attack on multi-modal action recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3829–3837.
[191] Y. Huang, C. Dai, W.-C. Chien, Sparse attack on skeleton-based human action recognition for internet of video things systems, in: Information Security Practice and Experience: 17th International Conference, ISPEC 2022, Taipei, Taiwan, November 23–25, 2022, Proceedings, Springer, 2022, pp. 197–212.
[192] T. Zheng, S. Liu, C. Chen, J. Yuan, B. Li, K. Ren, Towards understanding the adversarial vulnerability of skeleton-based action recognition, arXiv preprint arXiv:2005.07151 (2020).
[193] H. Wang, Y. Diao, Z. Tan, G. Guo, Defending black-box skeleton-based human activity classifiers, arXiv preprint arXiv:2203.04713 (2022).
[194] N. Tanaka, H. Kera, K. Kawamoto, Adversarial bone length attack on action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 2335–2343.
[195] H. Park, Z.J. Wang, N. Das, A.S. Paul, P. Perumalla, Z. Zhou, D.H. Chau, Skeletonvis: Interactive visualization for understanding adversarial attacks on human action recognition models, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 16094–16096.
[196] G.I. Parisi, R. Kemker, J.L. Part, C. Kanan, S. Wermter, Continual lifelong learning with neural networks: A review, Neural Networks 113 (2019) 54–71.
[197] G.I. Parisi, Human action recognition and assessment via deep neural network self-organization, in: Modelling Human Motion, Springer, 2020, pp. 187–211.
[198] T. Li, Q. Ke, H. Rahmani, R.E. Ho, H. Ding, J. Liu, Else-net: Elastic semantic network for continual action recognition from skeleton data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13434–13443.

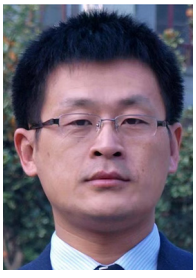
Wentian Xin received the B.S. degree in Intelligent Science and Technology from Institute of Robotics and Automatic Information System, Nankai University, Tianjin, China, in 2018. He is currently pursuing the Ph. D. degree with the School of Computer Science and Technology, Key Laboratory of Big Data and Intelligent Vision, Xidian University, Xi'an, China. His research interest is human action recognition, especially skeletonbased action recognition.

Ruyi Liu received the Ph.D. degree in School of Computer and technology, Xidian University. She is currently working as Lecturer at School of Computer Science and Technology, Xidian University. Her current interests include image classification and segmentation, and computer vision methods with applications in remote sensing.

Yi Liu received the B.S. degree in Computer Science and Technology, Anhui Normal University, Wuhu, China, in 2021. He is currently pursuing the M.E. degree with the School of Computer Science and Technology, Key Laboratory of Big Data and Intelligent Vision, Xidian University, Xi'an, China. His research interest is human skeletonbased action recognition.

Yu Chen obtained the bachelor's degree from College of Computer Science and Technology, Xidian University, Xi'an, China, in 2021. He is currently pursuing the master's degree in the Big Data and Artificial Intelligence Center of Tsingtao Institute of Computing Technology, Xidian University, Tsingtao, China. His research direction is human action recognition based on skeletons.

Wenxin Yu Wenxin Yu received the B.S. degree in software engineering from the Institute of Software Engineering at Northwest University, Xi'an, in 2020. He is currently pursuing the M.S. degree with the School of Computer Science and Technology, Key Laboratory of Big Data and Intelligent Vision, Xidian University, Xi'an, China. His research interest is human skeleton-based action recognition.



Qiguang Miao received the Ph.D. degree in computer application technology from Xidian University, Xi'an, China, in December 2005. He is currently a Professor and a Ph.D. Student Supervisor with the School of Computer Science and Technology, Xidian University. In recent years, he has published over 100 articles in the significant domestic and international journals or conferences. His research interests include machine learning, intelligent image processing, and malware behavior analysis and understanding.