Temporal-Channel Topology Enhanced Network for Skeleton-Based Action Recognition

Jinzhao Luo[1,2], Lu Zhou[1,2], Guibo Zhu[1,2,3(✉)], Guojing Ge[1], Beiying Yang[1,2], and Jinqiao Wang[1,2,3,4(✉)]

[1] Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[3] Wuhan AI Research, Wuhan 430073, China
[4] The Peng Cheng Laboratory, Shenzhen 518066, China.
luojinzhao2020@ia.ac.cn, {lu.zhou, gbzhu, guojing.ge, beiying.yang, jqwang}@nlpr.ia.ac.cn

Abstract. Skeleton-based action recognition has become popular in recent years due to its efficiency and robustness. Most current methods adopt graph convolutional network (GCN) for topology modeling, but GCN-based methods are limited in long-distance correlation modeling and generalizability. In contrast, the potential of convolutional neural network (CNN) for topology modeling has not been fully explored. In this paper, we propose a novel CNN architecture, Temporal-Channel Topology Enhanced Network (TCTE-Net), to learn spatial and temporal topologies for skeleton-based action recognition. The TCTE-Net consists of two modules: the Temporal-Channel Focus module, which learns a temporal-channel focus matrix to identify the most important feature representations, and the Dynamic Channel Topology Attention module, which dynamically learns spatial topological features, and fuses them with an attention mechanism to model long-distance channel-wise topology. We conduct experiments on NTU RGB+D, NTU RGB+D 120, and FineGym datasets. TCTE-Net shows state-of-the-art performance compared to CNN-based methods and achieves superior performance compared to GCN-based methods. The code is available at https://github.com/aikuniverse/TCTE-Net.

Keywords: human skeleton · action recognition · topology modeling.

1 Introduction

Action recognition is a crucial task with applications in various fields such as human-robot interaction and virtual reality. With the continuous development of depth sensors and pose estimators, obtaining high quality 3D skeletal data has become easier. As a result, skeleton-based action recognition received increasing attention in recent years, thanks to the compactness and robustness of human skeletal data against complicated backgrounds.

Graph Convolutional Networks (GCNs)[1–3] have become one of the most popular skeleton-based action recognition methods due to its ability to handle irregular topological information in skeletons[4, 5]. Specifically, GCNs model skeleton sequences as spatio-temporal graph topologies. They use a handcrafted graph, which represents physically connected edges among human skeleton, to extract the spatial features representing the relationships between joint nodes in the human skeleton. ST-GCN [5], the first well-known GCN-based method, constructs spatial and temporal correlations in skeletal data via graph convolution. Subsequently, Li *et al.* [6] expands the receptive field based on the self-attention mechanisms to learn topology between joints, while Wang *et al.* [7] aggregates the spatio-temporal topological feature representations to improve the modeling capacity. However, GCN-based methods have limitations. Joint nodes in skeleton are treated equally, which means important nodes and edges cannot be identified [20]. Furthermore, GCNs struggle to model the complicated correlations between distant unnaturally connected joint nodes. However, for human action recognition, relationships between structurally distant joints are as important as between adjacent joints. Besides, GCN-based methods require complex network structure designs to fuse skeleton and other modalities [8]. Different from the previous GCN methods, in this paper, we attempt to alleviate the disadvantage of GCN methods that cannot identify the relationships between distant joint nodes.

Compared with GCN, Convolutional Neural Network(CNN) can model topological features more effectively with powerful local convolution characteristics and self-attention mechanism[9, 10]. They can also be easily fused with other modalities [11]. Caetano *et al.* [12] converts the skeleton coordinates to a three channels pseudo image input, and then classifies the features extracted through the network. Such input cannot exploit the locality nature of convolution networks. Shi *et al.* [18] modeling spatial-temporal dependencies between joints without the requirement of knowing their positions or mutual connections by building the attention blocks. PoseC3D [11] generates 3D heatmap volumes from skeleton coordinates as input, and then classifies with a 3D-CNN. However, existing CNN-based methods do not utilize the natural topology of the bones. Processing in the temporal and spatial dimensions leads to the separation of spatio-temporal attributes of actions respectively, without forming a unified spatio-temporal feature representation. In addition, using the attention mechanism simply to assign different weights to each joint node cannot extract the relationship between joint nodes effectively.

In order to solve the above problems, in this paper, we propose a novel Temporal-Channel Topology Enhanced Network (TCTE-Net), which models the topological information of skeleton data effectively. Specifically, we propose a novel Temporal-Channel Focus (TCF) module, which emphasizes vital features to force the model focus on the critical joint nodes in action classification. And we propose a Dynamic Channel Topology Attention (DCTA) module, which can identify the relationships between distant joint nodes effectively and model the correlation between distant joint nodes dynamically.

Our contributions are summarized as follows:

1. We present a novel TCTE-Net for skeleton-based action recognition equipping with TCF and DCTA modules, which attempt to identify the critical joint nodes and relationships between distant joint nodes in action classification. The TCF module emphasizes the critical joint nodes with a focus matrix. DCTA learns distant channel-wise topology modeling based on the dynamic channel distance matrix and attention mechanism.
2. The extensive experiments verify the effectiveness of TCF and DCTA modules. The proposed TCTE-Net outperforms state-of-the-art CNN methods significantly and achieves remarkable performance compared to GCN-based methods on three skeleton-based action recognition datasets.

2 PROPOSED METHOD

In this section, we first propose the detailed architecture of TCTE-Net in Sec. 2.1. Then, the Temporal-Channel Focus Module and the Dynamic Channel Topology Attention Module are introduced in Sec. 2.2 and Sec. 2.3.

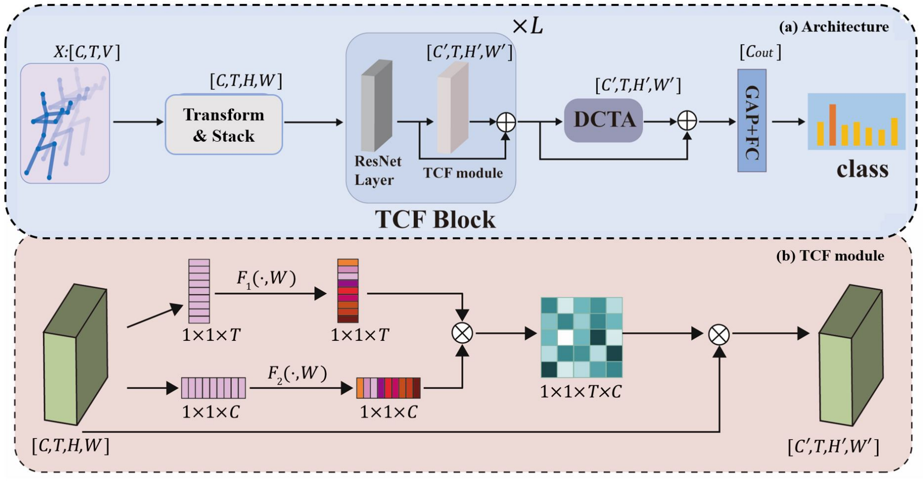

Fig. 1. (a) Pipeline of the proposed TCTE-Net, which consists of L TCF blocks and one DCTA module. We instantiate TCTE-Net with the SlowOnly backbone, where L is 3. (b) The detailed architecture of TCF module.

2.1 Network Architecture

Joint nodes in different body parts contribute to action classification differently. For example, in the case of a 'shaking hands' action, the weight of arm part is much higher than the body part. However, existing CNN-based methods are limited in identifying critical joint nodes, and CNNs cannot model the natural topology of the bones directly without an adjacency matrix, which is widely used in GCNs. To address these limitations, we propose the Temporal-Channel Topology Enhanced Network (TCTE-Net).

The TCTE-Net is illustrated in Fig. 1.(a), which consists of three TCF blocks and one DCTA module. TCTE-Net adopts lightweight SlowOnly 3D-CNN[11, 13] as the backbone. Our approach focuses more on feature representation and topology modeling. The 2d skeleton coordinates extracted from the video frames stored as coordinate-triplets $(x,\ y,\ c)$, where $(x,\ y)$ represent the coordinates of the skeleton joints and c is the corresponding probability. In order to get the input of TCTE-Net, we first transform the 2d skeleton coordinates into a pseudo-image form. Specifically, a set of 2D skeleton coordinates $(x,\ y,\ c)$ is represented as a heatmap of size $K \times H \times W$ by composing K gaussian maps centered at joint:

$$J_{kij} = e^{-\frac{(i-x_k)^2+(j-y_k)^2}{2*\sigma^2}*c_k}, \tag{1}$$

where σ controls the variance of gaussian maps, $(x_k,\ y_k)$ and c_k are respectively the location and confidence score of the k-th joint, and K is the number of joints, H and W are the height and width of the frame respectively. All heatmaps are then stacked along the temporal dimension T. We adopt the 3D heatmap volumes as input. The input joints dimension can be viewed as image channels dimension. In this case, a joint node in the original 2D skeleton is represented as a heatmap of size $H \times W$. The 3D heat map representation can retain the original human skeleton information and reduce information loss.

Firstly, The input 3D heatmap volumes are fed into L TCF blocks. The TCF block consists of ResNet layer and TCF module, where the 3D heatmap volumes converted into high level features through ResNet layer, and then extract vital joints feature representations through TCF module. At last, the high level feature representations are fed into DCTA module to model the topological relationship between distant joint nodes. Both TCF and DCTA module are residual connected. A classifier is followed to predict action labels. Based on TCF and DCTA module, TCTE-Net can model the local actions finely, and extract long-distance information between non-local joint points under different actions flexibly.

2.2 Temporal-Channel Focus Module

To enhance the critical joint nodes, we propose TCF module, which is illustrated in Fig. 1.(b). Before being fed into TCF module, the input features are transformed into high level representations $X \in \mathbb{R}^{C\times T\times H\times W}$. The features X are then fed into a two-stream structure of temporal and channel dimensions, which is implemented along the temporal dimension and channel dimension with a Global average pooling (GAP) layer and followed with a FC layer. The features X are downsampled and linearly transformed along the temporal and channel dimensions respectively. Through the above operations, we get the weighted vectors of joint node features in channel and temporal dimensions. The weighted vectors are then fused by element-wise multiplication. Activation function is applied to get the temporal-channel focus matrix.

In the early stages of the network, the heatmaps of different channels represent skeleton joint features of different body parts, which has different weights in classification. The proposed TCF module can distinguish critical local joint nodes, and extract the spatio-temporal local features that are more effective for action recognition adaptively. This makes the model pay more attention to the most discriminative local features, and addresses the difficulties of recognizing subtle and similar movements.

Finally, the joint node features X are strengthened by temporal-channel focus matrix. The overall process of TCF can be formulated as:

$$X' = \sigma\big(F_1\big(GAP(X) * F_2(GAP(X))\big)\big) * X, \tag{2}$$

where $X \in \mathbb{R}^{C\times T\times H\times W}$ and $X' \in \mathbb{R}^{C'\times T\times H'\times W'}$ are the input and output of the TCF module respectively. F_1 and F_2 mean FC layers. σ is activation function.

2.3 Dynamic Channel Topology Attention Module

To eliminate the weakness of CNN in modeling the irregular skeletal topology, we introduce the Dynamic Channel Topology Attention module (DCTA). As shown in Fig. 2, the self-attention matrix is used to extract global shared topology for all channels. Meanwhile, we learn specific relationships between joint nodes of different channels.

We utilize convolution and pooling operations on input feature X' to generate new feature representations Q, K, V, and reshape to Q, $K \in \mathbb{R}^{C'\times N}$ and $V \in \mathbb{R}^{C'\times T\times N}$, where $N = H' \times W'$. Then we calculate the spatial attention map $S \in \mathbb{R}^{N\times N}$:

$$S_{ji} = \frac{e^{q_i\cdot k_j}}{\sum_{i=1}^{N} e^{q_i\cdot k_j}}, \tag{3}$$

where S_{ji} measures the correlation between position q_i and k_j. CNN is capable of modeling the topology of joints implicitly [14]. Thus S are adopted to represent the global shared topological relationship between features of different joint nodes. Meanwhile, we calculate the channel-specific correlations between different features $M \in \mathbb{R}^{C'\times N\times N}$, which can be formulated as:

$$M(x_i, x_j) = \sigma\big(Q(x_i) - K(x_j)\big), \tag{4}$$

where $\sigma(\cdot)$ is activation function. $M(x_i, x_j)$ is dynamic channel distance matrix, which calculates distances between features x_i and x_j along channel dimension. Different channels represent different types of motion features in classification [1]. Therefore, $M(x_i, x_j)$ essentially models the topological relationship between joint nodes under different motion features. For distant unnaturally connected joint nodes, the dynamic channel distance matrix is able to capture

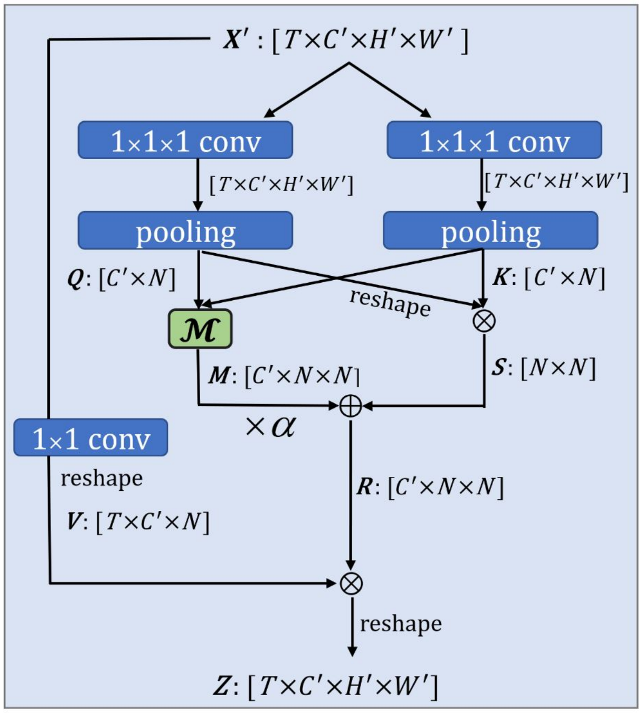

Fig. 2. The detailed architecture of DCTA module.

their specific correlations under different motion features dynamically. The final topological relation $R \in \mathbb{R}^{C' \times N \times N}$ is formulated as:

$$R = S + \alpha \cdot M. \tag{5}$$

The dynamic channel distance matrix $M(x_i, x_j)$ is utilized to enhance the global shared topological representation S with a trainable scalar α . The addition is conducted in a broadcast way. Finally, we perform a matrix multiplication between R and V , and reshape the result to $\mathbb{R}^{C' \times T \times H' \times W'}$. The output of DCTA module is formulated as:

$$Z_j = X'_j + \sum_{i=1}^{N} r_{ji} v_i, \tag{6}$$

where r_{ji} and v_i represent the corresponding elements in the matrix respectively.

The Equation (6) shows that the output feature Z is the sum of the final topological relation and original features, which models the long-range correlations between joint nodes dynamically.

3 EXPERIMENTS

3.1 Datasets and Implementation Details

NTU RGB+D NTU RGB+D is a large-scale human action recognition dataset. It contains more than 56k video samples of 60 human action classes performed by 40 distinct human subjects. Each sample is captured from different views by three Microsoft Kinect v2 cameras at the same time. The dataset has two benchmarks: Cross-subject (X-Sub), Cross-view (X-View), for which are split by action subjects, camera views in training and validation.

NTU RGB+D 120 NTU RGB+D 120 extends NTU RGB+D with 57k video samples of additional 60 action classes, which contains 113k samples over 120 human action classes performed by 106 human subjects. The authors recommend two benchmarks: Cross-subject (X-Sub) and Cross-setup (X-Set, split by camera setups).

FineGYM FineGYM is a fine-grained action recognition dataset. It contains 29K videos of 99 fine-grained action classes collected from 300 professional gymnastics competitions. The characteristic of this dataset is that the behaviors are divided according to the level from fine-grained actions to coarse-grained events, and the human skeleton of gymnastics moves has a large deformation.

Implementation details. TCTE-Net is implemented via Pytorch and trained with 8 RTX 2080 TI GPUs, where each GPU has 11 video clips in a mini-batch. The model is trained for 30 epochs with SGD optimizer. The initial learning rate is set to 0.1375 and decayed with Cosine Annealing scheduler [21]. The weight decay is set to 0.0003. For all datasets, we report the results of 10-clip testing.

3.2 Ablation Study

Table 1. Ablation study on NTU RGB+D. FM represents Focus Matrix.

Method	Param.	X-Sub (%)	X-View (%)
Baseline	2.03M	93.3	96.2
+TCF (1)	+0.10M	93.5	96.3
+TCF (3)	+0.25M	93.7	96.5
+TCF (3) w/o FM	+0.25M	93.5	96.4
+DCTA	+0.30M	93.6	96.5
+ TCF (3) + DCTA	+0.53M	**93.8**	**96.6**

Effectiveness of TCF. Table 1 illustrates the performance gains brought about by TCF and DCTA on the NTU RGB+D dataset. We use SlowOnly as the baseline model and add the TCF module to it. When adding a TCF module,the accuracy increases by 0.2% compared to the baseline on the X-view benchmark. When increasing the number of TCF modules to 3, TCF boosts accuracy by 0.4% on X-view benchmark, with little parameter increase. Our results validating the effectiveness of TCF module. To verify the validity of the Focal Matrix in the TCF module, we then remove the Focus Matrix (FM) from TCF. The TCF without FM module parallels the weighted channel and temporal vectors. Compared to TCF, the performance of TCF without FM drops by 0.2%, indicating that the weighted focus matrix can identify classification-vital features efficiently, thereby assigning different attention weights to all spatio-temporal joint nodes.

Effectiveness of DCTA. By introducing the DCTA module, we further improve accuracy by 0.3%. DCTA dynamically models the relationship between long-distance joint points under different action characteristics based on the dynamic channel distance matrix, so that the model has a better recognition effect on actions with large deformation and fast movement speed. Our proposed TCTE-Net achieves an accuracy of 93.8% with the X-Sub benchmark, which improves the baseline accuracy by 0.5% with an efficient model.

3.3 Comparison with the State-of-the-Art

Table 2. Comparative Experiment of TCTE-Net and SOTA Model on NTU RGB+D Dataset.

Type	Method	X-Sub (%)	X-View (%)
CNN	DSTA-Net [18]	91.5	96.4
	Ta-CNN+ [14]	90.7	95.1
	PoseConv3D+ [11]	93.7	96.6
GCN	MS-G3D+ [3]	91.5	96.2
	STF [9]	92.5	96.9
	HD-GCN+ [20]	93.0	**97.0**
Ours	TCTE-Net	**93.8**	96.6

In the experimental results section, we evaluate the effectiveness of TCTE-Net on three benchmark datasets: NTU RGB+D, NTURGB+D 120, and FineGYM. Many state-of-the-art methods employ multi-stream fusion models[20, 22, 23], joint, bone. For a fair comparison, we compare our model with the state-of-the-art methods obtained by the best single models on each dataset, and our model significantly outperforms the other methods.

On the NTU RGB+D dataset, the results shown in Table 2 demonstrate that our model is effective. Although the performance of X-View benchmark is nearly saturated, our model still obtains remarkable performance. Moreover, TCTE-Net achieves an accuracy rate of 93.8% in the X-Sub benchmark test, outperforming other state-of-the-art single-model methods, and achieves a performance improvement of 0.1% compared with advanced multi-stream fusion methods. Compared with the state-of-the-art GCN model, TCTE-Net achieves 0.8% performance improvement on X-Sub benchmark test.

Table 3. Comparative Experiment of TCTE-Net and SOTA Model on NTU RGB+D 120 Dataset.

Type	Method	X-Sub (%)	X-View (%)
CNN	DSTA-Net [18]	86.6	89.0
	Ta-CNN+ [14]	85.7	87.3
	PoseConv3D [11]	86.0	89.6
GCN	Shift-GCN [24]	85.9	87.6
	InfoGCN [25]	85.1	86.3
	HD-GCN+ [20]	85.7	87.3
ours	TCTE-Net	**86.6**	**89.9**

On the challenging NTU RGB+D 120 dataset, our model achieves excellent performance, as shown in Table 3. Compared with the state-of-the-art single model, TCTE-Net achieves 0.6% and 0.3% performance improvements on the X-Sub and X-View benchmarks respectively, achieving comparable performance to the state-of-the-art multi-stream fusion methods, and outperforming the previous state-of-the-art CNN method. Compared with the most advanced GCN method, TCTE-Net achieves 0.7% and 2.3% performance improvements on the X-Sub and X-View benchmarks respectively, which verifies the effectiveness of the proposed method in the skeleton action recognition task.

Table 4. Comparative Experiment of TCTE-Net and SOTA Model on FineGYM Dataset.

Type	Mean Top-1 Accuracy (%)
MS-G3D+ [3]	92.6
PoseConv3D [11]	93.2
TCTE-Net	**93.8**

Furthermore, we evaluate TCTE-Net on the FineGYM dataset. The results shown in Table 4 demonstrate that our model achieves state-of-the-art performance on the FineGYM dataset. Our model obtains an accuracy of 93.8%, which outperforms the state-of-the-art GCN-based method by 1.2%. Notably, the GCN-based methods are weak in modeling non-connected joint relationships, while our model is able to capture long-range correlations of non-directly connected joints in the skeleton. Therefore, for FineGYM with large movement and deformation, TCTE-Net achieves higher performance than the GCN-based methods.

4 CONCLUSION

This paper proposes TCTE-Net, a novel framework for skeleton-based action recognition that addresses the limitations of CNN in modeling the irregular topology of the skeletal data. Through the proposed Temporal-Channel Focus module and Dynamic Channel Topology Attention module, we enhance the ability of TCTE-Net to identify critical joint nodes and model the correlation between joints under different motion features. Experiments on three benchmark datasets show that TCTE-Net outperforms the previous state-of-the-art models. Our work contributes to exploring the potential of CNNs for modeling skeletal data, and we hope that this will inspire further investigations in this direction.

Acknowledgements. This work was supported by National Key R&D Program of China under Grant No. 2021YFE0205700, National Natural Science Foundation of China (No.62276260, 62076235, 62206283, 61976210, 62176254, 62002356, 62006230, 62002357, 62206290), sponsored by Zhejiang Lab (No.2021KH0AB07).

References

1. Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu.: Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. In: ICCV (2021)
2. K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu.: Decoupling GCN with DropGraph Module for Skeleton-Based Action Recognition. In: ECCV (2020)
3. Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang.: Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In: CVPR (2020)
4. L. Shi, Y. Zhang, J. Cheng, and H. Lu.: Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In: CVPR (2019)
5. S. Yan, Y. Xiong, and D. Lin.: Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In: AAAI (2018)
6. B. Li, X. Li, Z. Zhang, and F. Wu.: Spatio-Temporal Graph Routing for Skeleton-Based Action Recognition. In: AAAI (2019)
7. S. Wang, et al.: Skeleton-based Action Recognition via Temporal-Channel Aggregation. In: arXiv preprint arXiv: 2205.15936 (2022)
8. S. Das, S. Sharma, R. Dai, F. Brémond, and M. Thonnat.: VPN: Learning Video-Pose Embedding for Activities of Daily Living. In: ECCV (2020)

9. A. Vaswani et al.: Attention Is All You Need. In: NIPS (2017)
10. J. Fu et al.: Dual Attention Network for Scene Segmentation. In: CVPR (2019)
11. H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai.: Revisiting Skeleton-based Action Recognition. In: CVPR (2022)
12. C. Caetano, J. Sena, F. Brémond, J. A. dos Santos, and W. R. Schwartz.: SkeleMotion: A New Representation of Skeleton Joint Sequences Based on Motion Information for 3D Action Recognition. In: AVSS (2019)
13. C. Feichtenhofer, H. Fan, J. Malik, and K. He.: SlowFast Networks for Video Recognition. In: ICCV (2019)
14. K. Xu, F. Ye, Q. Zhong, and D. Xie.: Topology-Aware Convolutional Neural Network for Efficient Skeleton-Based Action Recognition. In: AAAI (2022)
15. A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang.: NTU RGB+D: A Large-Scale Dataset for 3D Human Activity Analysis. In: CVPR (2016)
16. J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot.: NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. In: TPAMI (2019)
17. D. Shao, Y. Zhao, B. Dai, and D. Lin.: FineGym: A Hierarchical Video Dataset for Fine-grained Action Understanding. In: CVPR (2020)
18. L. Shi, Y. Zhang, J. Cheng, and H. Lu.: Decoupled Spatial-Temporal Attention Network for Skeleton-Based Action Recognition. In: ACCV (2020)
19. L. Ke, K.-C. Peng, and S. Lyu.: Towards To-a-T Spatio-Temporal Focus for Skeleton-Based Action Recognition. In: AAAI (2022)
20. J. Lee, M. Lee, D. Lee, and S. Lee.: Hierarchically Decomposed Graph Convolutional Networks for Skeleton-Based Action Recognition. In: arXiv preprint arXiv:2208.10741 (2022)
21. I. Loshchilov and F. Hutter.: SGDR: Stochastic Gradient De-scent with Warm Restarts. In: arXiv preprint arXiv:1608.03983 (2016)
22. L. Shi, Y. Zhang, J. Cheng, and H. Lu.: Skeleton-Based Ac-tion Recognition with Multi-Stream Adaptive Graph Convolutional Networks. In: TIP (2020)
23. Y.-F. Song, Z. Zhang, C. Shan, and L. Wang.: Constructing Stronger and Faster Baselines for Skeleton-based Action Recognition. In: TPAMI (2022)
24. K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu.: Skeleton-Based Action Recognition with Shift Graph Convolutional Network. In: CVPR (2020)
25. H. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani.: InfoGCN: Representation Learning for Human Skeleton-Based Action Recognition. In: CVPR (2022)