STEP CATFormer: Spatial-Temporal Effective Body-Part Cross Attention Transformer for Skeleton-based Action Recognition

Bao Long Nguyen Huu
tp22010-0251@sti.chubu.ac.jp
Tohgoroh Matsui
matsui@isc.chubu.ac.jp

Department of Information Engineering
Chubu University
Kasugai, Aichi, Japan

Abstract

Graph convolutional networks (GCNs) have been widely used and achieved remarkable results in skeleton-based action recognition. We think the key to skeleton-based action recognition is a skeleton hanging in frames, so we focus on how the Graph Convolutional Convolution networks learn different topologies and effectively aggregate joint features in the global temporal and local temporal. In this work, we propose three Channel-wise Tolopogy Graph Convolution based on Channel-wise Topology Refinement Graph Convolution (CTR-GCN). Combining CTR-GCN with two joint cross-attention modules can capture the upper-lower body part and hand-foot relationship skeleton features. After that, to capture features of human skeletons changing in frames we design the Temporal Attention Transformers to extract skeletons effectively. The Temporal Attention Transformers can learn the temporal features of human skeleton sequences. Finally, we fuse the temporal features output scale with MLP and classification. We develop a powerful graph convolutional network named Spatial Temporal Effective Body-part Cross Attention Transformer which notably high-performance on the NTU RGB+D, NTU RGB+D 120 datasets. Our code and models are available at https://github.com/maclong01/STEP-CATFormer

1 Introduction

Computer vision is a field that has widespread applications in various aspects of life, such as object recognition, image segmentation, and human action recognition. In recent years, human action recognition has received significant attention due to advancements in deep learning and computer vision. Applications of human action recognition include games play, eldercare, healthcare assistance, and video surveillance. With the development of high-performance sensors and advanced algorithms for human pose estimation [11], it is now possible to acquire accurate 3D skeletal data. Recent advances in 3D depth cameras such as Microsoft Kinect camera [31] was an attempt to broaden the 3D gaming experience of the Xbox 360's audience and advanced human pose estimation algorithms [3] enable quick

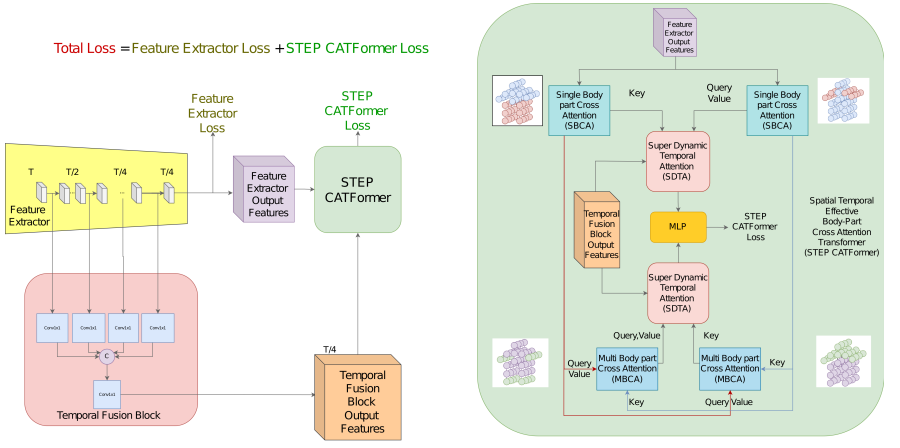

Figure 1: **STEP CAT model** has 3 branches Feature extractor branch, temporal fusion branch, and STEP CATFormer branch. In STEP CATFormer branch has 4 small branches Single Body-part Cross Attention (**SBCA**), Multi Body-part Cross Attention (**MBCA**), Super Dynamic Temporal Attention (**SDTA**), and MLP with classification.

and accurate adjustment of 3D skeletons using inexpensive devices. In computer vision for games, Kinect sensor real-time skeletal tracking recognizes the actions of the human body is the key technology behind Kinect is humanbody language understanding, which means that the computer first recognizes and understands what the user is doing, before responding exactly what they need to improve the quality of the game play. But skeleton-based action recognition still faces several challenges, including body size, viewpoint, and motion speed [1, 17, 26].

However, Graph Neural Networks (GNNs) [32], and specifically, Graph Convolutional Networks (GCNs)[8], can effectively capture spatial and temporal information to solve various problems. Yan et al. [26] first applied GCNs to the field of skeleton-based action recognition and showed that the joints of the skeletal spatiotemporal graph of human action are correlated. They used GCNs and temporal convolution [2] to extract motion features and graph structures to model the correlation between human joints in space-time. Various approaches to skeleton-based human action recognition have been proposed based on this idea, including the use of second-order information from skeletal data, multi-stream networks, and attentional mechanisms [6, 15, 16, 19, 30].

However, these approaches have limitations such as limiting the representational power of the model in channel topologies, having unnaturally connected joint relationships, and ignoring the variability of different channel topologies. To address these limitations, Chen et al. [4] proposed Channel Topology Refinement Graph Convolution (CTR-GCN), an approach that learns topologies and aggregates features in different channel dimensions dynamically and now many method [5, 6, 27] base on it. However, it mostly favors modeling in the spatial dimension and does not emphasize temporal dimensions.

To improve temporal features and consider the features importance in different body parts and joints, we propose a method called Spatial Temporal Effective Body-part Cross Attention Transformer. This method can dynamically learn the relationship between body parts and joints in body parts features on spatiotemporal dimensions. Specifically, we use CTR-

GCN's dynamic channels topologies temporal feature representation in high-dimensional space extracted in the last layer to capture features in body parts with intra-body joints in body parts relationships. We also propose a powerful temporal attention mechanism to efficiently extract temporal features using low-dimensional space extracted in the previous layer of CTR-GCN and fuse them with embedding body parts features. The body parts and intra-body part attention method combined with the powerful temporal attention transformers completes the modeling.

2 Related work

With the rise of vision transformers, transformer-based methods have been applied to skeleton data analysis, such as [12, 13, 15, 20, 24, 25]. Recent studies have extended the Transformer model to the recognition of actions based on skeleton data in both spatial and temporal dimensions [15, 20, 24]. The IIP-Transformer model [24] was the first to use self-attention to understand the relationships between joints, while some datasets employ a combination of spatial transformer and temporal transformer. LST [25] uses a hybrid architecture that combines GCN and Transformer in a body part, where each body-part have using contrastive learning before spatial in place of GCN and temporal convolution. These methods effectively capture spatiotemporal information about the skeleton and show promise in skeletal action recognition. Our work in this paper we combine graph convolution and transformer method hybrid architecture model training for skeleton-based action recognition.

3 Proposed Method

3.1 Spatial-Temporal Effective Body-part Cross Attention Transformer Overview

An overview of the proposed STEP-CATFormer network is shown in Figure 1. It consists of three types of blocks (Feature Extractor, Temporal Fusion, and Spatial-Temporal Effective Body-part Cross Attention Transformer), in STEP-CATFormers which our three primary components with spatial dimension are Single Body-part Cross Attention Block and Multi Body-part Cross Attention Block, and Temporal Dimension Super Dynamic Temporal Attention. STEP-CATFormers focus on modeling both the discriminative relationships between joints and body part skeletons in spatio-temporal motion patterns for recognition. First, given a skeleton sequence from the feature extractor $X_{in} \in \mathbb{R}^{N\times T\times C_0}$, a linear layer is applied to the STEP-CATFormer to project it to the position embedding, generating the feature $X_1 \in \mathbb{R}^{N\times T\times C_1}$. Then, X_1 is split and passed into two Single Body-part Cross Attention (SBCA) branches. One branch adaptively discriminates the relationship in the hand joints and another joint, producing features $\{X_2^H \in \mathbb{R}^{H\times T\times C}\}$, where H is the number of hand joints. And, the other branch partitions discriminate the relationship in the leg-foot joints and another joint, generating the feature $\mathrm{X}_2^F \in \mathbb{R}^{F\times T\times C}$, where F is the number of foot and leg joints. X_2^H and X_2^F are then passed through Super Dynamic Temporal Attention (SDTA) and Multi Body-part Cross Attention (MBCA). In one of the ways, the inputs X_2^H and X_2^F are passed into two MBCA branches. One branch adaptively discriminates the relationship in the wrist-ankle joints and another joint part, producing features $X_3^{WA} \in \mathbb{R}^{WA\times T\times C}$, where WA is the number of wrist-ankle joints. Other branch partitions discriminate the relationship

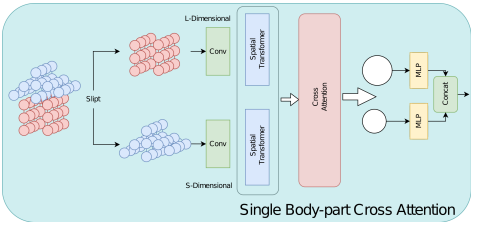

Figure 2: **Single Body-part Cross Attention** (SBCA) had input features from feature extractor to two Spatial Attention with one L Transformer proposed Large channels dimension Transformer and S Transformer

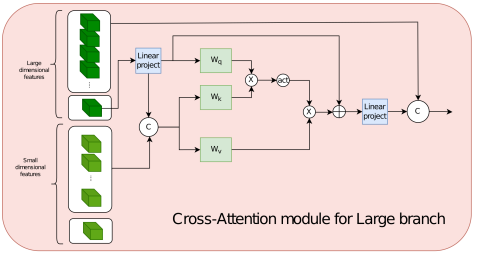

Figure 3: **Cross-attention module for Large branch**. The joint tokens of the large-dimension transformer serve as a query token to interact with the joint tokens from the small-dimension transformer. Linear Projects are projections to align dimensions.

in the up and down joints and another joint generates the feature $\mathrm{X}_3^{UD} \in \mathbb{R}^{UD\times T\times C}$, where UD is the number of up and down joints. In particular, the SBCA, MBCA block consists of a spatial transformer sub-block and a joint-part cross-attention sub-block to sufficiently model the spatial interaction information of actions. In the Super Dynamic Temporal Attention Transformer with the input X_2^H and X_2^F and temporal fusion features to give output $\{X_n^{T_1} \in \mathbb{R}^{T\times C}\}_{n=1}^N$. For other input features X_3^{WA} and X_3^{UD}, the SDTA gives the output features at $\{X_n^{T_2} \in \mathbb{R}^{T\times C}\}_{n=1}^N$. Finally, using two outputs of the SDTA to do element-wise addition, then processing out with MLP to produce features and applied Global Average Pooling (GAP) $X_{out} \in \mathbb{R}^{1\times 1\times C_{out}}$ to classification using fully connected layer and Softmax classifier.

3.2 Single Body-part Cross Attention

To allow information to diffuse across the joints and body parts, we develop a Single Body-part Cross Attention (SBCA) block with a cross-attention sub-block, shown in Figure 3.2. Cross-attention uses multi-head cross-attention to interact and diffuse the features of the two branches. We describe the cross-attention module for the large branch (L-branch) with the input as $\{x_2^F \in \mathbb{R}^{F\times T\times C}\}$, where F is the number of leg and foot joints extracted by 1×1 convolution and spatial transformation in large dimension, and the same procedure is performed for the small branch (S-branch) with the input as $\{x_2^O \in \mathbb{R}^{O\times T\times C}\}$, where O is the number of the remaining joints, excluding the foot and leg joints extracted by 1×1 convolution and spatial transformer in small dimension and by simply swapping the index l and s. The cross-attention module for the large branch is shown in Figure 3.2. For branch l, it first collects the tokens from the S-Branch and concatenates them, as shown in

$$(x^F)'^l_{cls} = f^l(x^F)^l_{cls}, \quad (x^F)'^l = [(x^F)'^l_{cls} \| (\mathrm{x}^{\mathrm{O}})^s_{d^s-cls}], \tag{1}$$

where $f^l(\cdot)$ is the projection function for dimension alignment and d^s is the s-branch dimension of x^O, $\{(x^F)'^l_{cls} \in \mathbb{R}^{F\times T\times 1}\}$, and $\{(x^O)^s_{d^s-cls} \in \mathbb{R}^{F\times T\times C-1}\}$. Mathematically, the

cross-attention can be expressed as

$$q = (x^F)'^l_{cls} W_q, k = (x^F)'^l W_k, v = (x^F)'^l W_v, A = \text{softmax}(qk^T/\sqrt{C/h}), CA((x^F)'^l) = Av, \tag{2}$$

where $W_q, W_k, W_v \in \mathbb{R}^{C\times(C/h)}$ are the learnable parameters, and C and h are the embedding dimensions and the number of heads, respectively. Since we use *cls* in the channel queries, the generation of the attention map A in cross-attention is linear as in all-attention. In self-attention, we also use multiple heads in the cross-attention and represent it as MCA. The output $(y^F)^l_2$ of the branch cross-attention module with a large channel dimension is defined as follows:

$$(y^F)^l_{cls} = f^l((x^F)^l_{cls}) + MCA(LN([f^l(x^F)^l_{cls} \| (x^O)^s_{d^s-cls}]) \tag{3}$$

$$(y^F)^l_2 = [g^l((y^F)^l_{cls}) \| (x^F)^l_{d^l-cls}], \tag{4}$$

where $\{(x^F)^l_{d^l-cls} \in \mathbb{R}^{F\times T\times C-1}\}$, and $f^l(\cdot)$ and $g^l(\cdot)$ are the projection and back-projection functions for dimensionaln alignment, respectively. Finally, after MLP alignment and concatenation, FFN contains a two-layer multilayer perceptron with expansion ratio r at the hidden layer and a GELU non-linearity is applied after the first linear layer. The layer normalization (LN) is applied before every block, and residual shortcuts after every block can be expressed as

$$z^F_2 = f_l[(y^F)^l_2] \| f_s[(y^O)^s_2], \quad X^F_2 = z^F_2 + FFN(z^F_2), \tag{5}$$

where $(y^O)^s_2$ is the output from the cross-attention small branch, f_l and f_s are the MLP alignments for the small and large channel dimensions. The output is X^F_2 and the out of X^H_2 has the same single body-part cross-attention construction module with X^F_2 with non-shared parameters.

3.3 Multi Body-part Cross Attention

Second, to defuse more detaild information across the joints and body parts based on the features extracted by SBCA, we develop a Multi Body-part Cross Attention (MBCA) block with cross-attention sub-block, shown in Figure 4.

Like SBCA, the cross-attention uses multi-head cross-attention to interact and diffuse features of the two branches. The cross-attention module for the large branch (L-branch) with the input as $\{(x)'^{UD}_3 \in \mathbb{R}^{UD\times T\times C}\}$, where UD is the number of up and down joints extracted by 2 1×1 convolution and spatial transformer block with 2 features input. One of them is extracted by hand joints cross-attention block $\{x^U_3 \in \mathbb{R}^{U\times T\times C}\}$, where U is the number of up joints that comes to q, v gated in spatial transformer by large dimensional convolution. And the other one as extracted in foot-leg joint cross-attention block $\{x^D_3 \in \mathbb{R}^{D\times T\times C}\}$, where D is the number of down joints that comes to k gated in spatial transformer by large dimensional convolution. Both the hand joint cross-attention block and the foot-leg joint cross-attention block module constructed by the SBCA module we introduced before.

And the same procedure is performed for the small branch (S-branch) with the input as $\{x^{sUD}_2 \in \mathbb{R}^{UD\times T\times C}\}$. One of them is extracted by hand joint cross-attention block $\{x^U_3 \in \mathbb{R}^{U\times T\times C}\}$, where U is the number of up joints comes to k gated in spatial transformer by small dimensional convolution. And the other one is extracted in foot-leg joint cross-attention block $\{x^D_3 \in \mathbb{R}^{D\times T\times C}\}$, where D is the number of down joints come to q, v gated

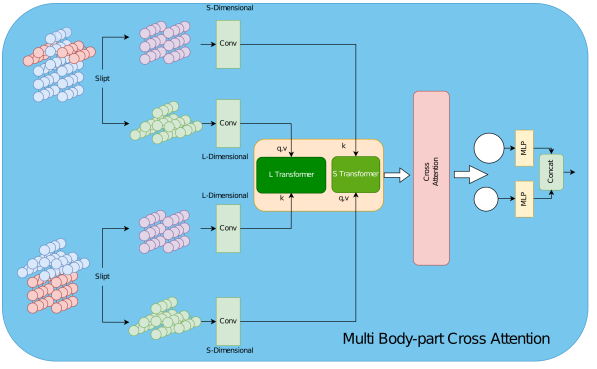

Figure 4: **Multi Body-part Cross Attention** (MBCA) had input with different attention body-part features to two Spatial Attention with one L Transformer proposed Large channels dimension Transformer and S Transformer represented Small channels dimension Transformer then with two input we use the cross attention method to give the relationship in L and S Transformer features output.

in spatial transformer by small dimensional convolution. Similar to the SBCA, the cross-attention module for the large branch is shown in Figure 3.2 with different alignment convolution and spatial transformer construction. Finally, the output is the X_3^{UD} and the out of X_3^{WA} has the same multi-body-part cross-attention construction module with X_3^{UD}.

3.4 Super Dynamic Temporal Attention Transformer

From a layer normalized tensor $Y \in \mathbb{R}^{N\times T\times C}$, our MHA first generates query (Q), key (K), and value (V) projections, and we perform average pooling to V temporal fusion features $X_1,...,X_V|X_i \in \mathbb{R}^{C\times T}$ to obtain the temporal fusion skeleton-level feature P. Then we used it P concatenation with $\widehat{Q} = Q\|P$; $\widehat{K} = K\|P$; $\widehat{V} = V\|P$ to capture the temporal fusion skeleton features together in the attention method. This is achieved by applying 1×1 convolutions to aggregate the pixel-wise cross-channel context followed by 3×3 depth-wise convolutions to encode channel-wise spatial context, yielding $Q = W_d^Q W_p^Q$, $K = W_d^K W_p^K$, where $W_p^{(\cdot)}$ is the 1×1 point-wise convolution and $W_d^{(\cdot)}$ is the 3×3 depth-wise convolution. We use bias-free convolutional layers in the network. Whereas the value (V), different from existing works, contains four blocks, each containing a 1×1 convolution to reduce the channel dimension. The first block just reduces the channel dimension, the second and third blocks contain two dilated temporal convolutions with kernel size 7 and different dilation rates d_t to obtain it, and a MaxPool following 1×1 convolution. The results of the four blocks are concatenated into the output denoted by

$$V = TCN_{d_2}(W_pY)\|TCN_{d_3}(W_pY)\|Max(W_pY)\|W_pY + Y. \tag{6}$$

Next, we reshape the query and key projections such that their dot-product interaction generates a transposed-attention map A of size $\mathbb{R}^{\widehat{C}\times\widehat{C}}$, instead of the attention map $\mathbb{R}^{\widehat{N}\widehat{T}\times\widehat{N}\widehat{T}}$ [9, 23]. Overall, the MHA process is defined as:

$$Attention(\widehat{Q},\widehat{K},\widehat{V}) = \widehat{V}\cdot \text{softmax}(\widehat{K}\cdot\widehat{Q}/\alpha), \tag{7}$$

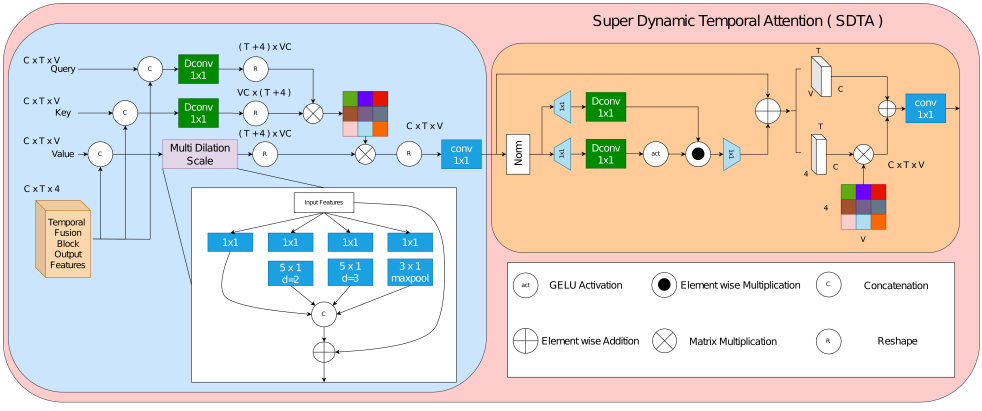

Figure 5: **Super Dynamic Temporal Attention** (SDTA) has blue area MHA is proposed Multi Head Attention and orange area represented FFN(Feed-Forward Network) have designed base in Gated-Dconv FFN[29] and Dynamic Temporal GCN method [10]

where X and X^T are the input and output feature maps; $\widehat{Q}, \widehat{V} \in \mathbb{R}^{\widehat{N}\widehat{T}\times\widehat{C}}$, and $\widehat{K} \in \mathbb{R}^{\widehat{C}\times\widehat{N}\widehat{T}}$ are matrices obtained after reshaping tensors from the original size $\mathbb{R}^{\widehat{N}\times\widehat{T}\times\widehat{C}}$. Here, α is a learnable scaling parameter to the magnitude of the dot product of $\widehat{K}$ and $\widehat{Q}$ before the softmax function is applied. Similar to the conventional multi-head SA [9], we divide the number of channels into heads and learn separate attention maps in parallel. After that, we use GDFN [29] of two fundamental modifications gating mechanism, and depthwise convolutions to improve representation learning.

Finally, we use the average pooling to joint-level temporal fusion features $\widehat{P}_1, ..., \widehat{P}_V | \widehat{P}_i \in \mathbb{R}^{C\times T}$ to obtain the skeleton-level feature $\widehat{P}'$. The skeleton-level temporal fusion feature takes the skeleton-level stride temporal features from the feature extractor. Joint-level temporal fusion is then applied to merge $\widehat{P}'$ into each X'_i. Like DG-TCN, each instance of the skeleton-level temporal fusion feature contains a learned parameter $\varphi \in \mathbb{R}^{\mathbb{V}}$. After the joint-level adaptive element-wise addition with skeleton-level temporal fusion, the feature for joint i is $\widehat{X}_i + \varphi_i \widehat{P}$, which will be further processed with a 1×1 convolution to get the output.

4 Experiments

4.1 Dataset

NTU RGB+D [18] is a popular resource for recognizing human actions based on skeletons. It consists of 56,880 sequences of such actions. There are two evaluation benchmarks: Cross-Subject (X-Sub) and Cross-View (X-View). The training and test sets are drawn from two non-overlapping sets of 20 subjects each in X-Sub. In X-View, the training set is made up of 37,920 samples captured by cameras 2 and 3, while the test set includes 18,960 sequences captured by camera 1.

NTU RGB+D 120 [14] is an extended version of the NTU RGB+D dataset, containing an additional 57,367 skeleton sequences across 60 additional action classes, for a total of 120 action classes. The authors propose two benchmark evaluations: Cross-Subject which

Table 1: Partition strategies on NTU RGB+D 120(X-sub).

Partition Strategy	FLOPS	Acc(%)
DSTA [21]	64.7G	86.6
EfficientGCN-B4 [22]	15.2G	88.3
Hyperformer (Joint Only) [33]	14.8G	**86.6**
Hands	4.1G	86.3
Legs	4.1G	86.0
Upper, Lower	4.3G	86.4
Wrist, Ankle	4.7G	86.5
Hands Legs, Upper, Lower, Wrist Ankle	18.3G	**89.6**

in NTU RGB+D, requires differentiation between two groups of subjects, and each group consists of 53 volunteers and Cross-Setup (X-Setup) which data are acquired in different configurations.

4.2 Ablation Study

Figure 6: Effect of LST on Different Skeleton Encoders

Backbone	Acc (%) w./o.	Acc (%) w.
ST-GCN	82.6	**84.6 (↑ 2.0)**
CTR-baseline	83.7	**85.5 (↑ 1.8)**
CTR-GCN (single scale)	84.6	**86.0 (↑ 1.4)**
CTR-GCN (multi scale)	84.9	**86.0 (↑ 1.1)**
LST	85.5	**85.9 (↑ 0.4)**

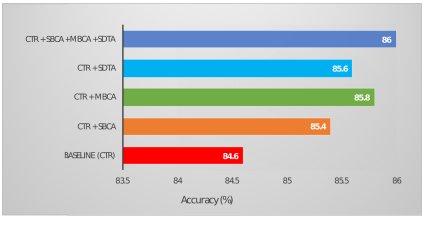

Figure 7: Impact of integrating our contributions in the baseline

In this section, we evaluate the performance of our Spatial Temporal Effective body-Part Cross Attention Transformer on the X-sub benchmark of the NTU RGB+D 120 dataset.

Partition Strategies

We test each body-part cross-attention partition strategy for STEP-CATFormer and the results are shown in Table 1. Hands, Legs, Wrist, and Ankle represents each hand, leg, wrist, and ankle joint with the remaining other joints cross-attention. And Upper and Lower represents upper joints with lower joints cross-attention. Finally, using more parts and multi-part cross-attention could steadily increase the performance, and it saturates at 86.0% when using 6 parts cross-attention, and it improves over the baseline by 1.4% in X-sub joints. And in 4 ensembles strategy (Joint, Joint-Motion, Bone, Bone-Motion) it saturates at 89.6% when using 6 parts cross-attention improves over the baseline by 0.7%.

Impact of the proposed contributions

This analysis evaluates the impact of our spatiotemporal enrichment module and query-class classifier. Our Spatiotemporal Cross-Attention module includes sub-modules SBCA, MBCA, and SDTA. Figure 7 compares the performance of our three contributions (Spatiotemporal Cross-Attention module and query-class classifier) with the baseline CTR-GCN.

Table 2: Performance on the NTU RGB+D and NTU RGB+D 120 dataset.

Methods	NTU-60		NTU-120	
	X-Sub (%)	X-View (%)	X-Sub (%)	X-Set (%)
GCN-based Methods				
ST-GCN[26]	81.5	88.3	70.7	73.2
CTR-GCN[4]	92.4	96.8	88.9	90.6
DG-STGCN[10]	**93.2**	**97.5**	89.6	**91.3**
LST[25]	92.9	97	**89.9**	91.1
Info-GCN[7]	93.0	97.1	89.8	91.2
Transformer-based Methods				
ST-TR[15]	89.9	96.1	82.7	84.7
STST[28]	91.9	96.8	-	-
IIP-Transformer[24]	92.3	96.4	88.4	89.7
FG-STFormer[12]	92.6	96.7	89.0	90.6
Hyperformer[33]	92.6	96.5	89.9	91.2
STEP CATFormer (CTR-GCN Feature Extractor)	93.0	96.9	89.6	90.8
STEP CATFormer (LST Feature Extractor)	**93.2**	**97.3**	**90.0**	**91.2**

The baseline CTR-GCN has an action classification accuracy of 84.6% (represented by the red bar). Integrating SBCA, which enriches the spatial context of single body-part cross-attention features before temporal modeling, results in 85.4% accuracy (represented by the orange bar). Similarly, the integration of MBCA, which enriches the spatial context of multi-body-part cross-attention features, results in a 1.2% gain (represented by the green bar). The integration of SDTA, which enriches the temporal context of frame-level features, further improves the accuracy to 85.6% (represented by the light blue bar). Finally, integrating the query-class classifier further enhances feature discriminability, resulting in an accuracy of 86.0% (represented by the blue bar). The final STEP-CATFormer framework achieves a 1.4% improvement over the baseline CTR-GCN (red bar). Our proposed STEP-CATFormer is decoupled from the network architecture and could be used to improve various skeleton encoders. Table 6 shows the experimental results of applying STEP-CATFormer to ST-GCN, CTR-baseline, CTR-GCN, and LST. STEP-CATFormer brings consistent improvements (0.4–2.0%) over the original models with no additional computational cost at inference, demonstrating the effectiveness and generalizability of STEP-CATFormer.

4.3 Comparison with the State-of-the-art

We compare our method with the state-of-the-art (SOTA) in Table 2. For a fair comparison, we use the 4 ensembles strategy (Joint, Joint-Motion, Bone, Bone-Motion) as it is adopted by most of the previous methods. We also show the ensemble results as well as the individual results. The results show that our proposed STEP-CATFormer method consistently high-performance one the state-of-the-art. Our method outperforms all existing transformer-based methods under nearly all evaluation benchmarks on NTU-60 and NTU-120, including the latest method Hyperformer [33]. Besides, our method outperforms CTR-GCN by 0.8% on NTU-60 X-Sub, and 0.2% both on NTU-60 X-Sub and X-View compared to Info-GCN [7]. It also outperforms LST [25] by 0.3% both on them. Table 2 shows that STEP-CATFormer outperforms CTR-GCN on the largest dataset, NTU RGB+D 120, by a significant margin of 1.1% on cross-subject and 0.6% on cross-set. Although Info-GCN also performs well in 6

ensemble strategies on this dataset, STEP-CATFormer still higher with 0.1% and 0.1% in 4 ensemble strategies, respectively.

5 Discussion and Conclusion

Our research paper is dedicated to the optimization of STEP-CATFormer to improve skeleton-based action recognition using Kinect camera, with a particular focus on the caracteristics of two specific skeleton datasets. While our proposed model may not necessarily be the most effective for other tasks, our approach of taking inherent relationships into account has the potential to enhance the use of Transformers in various applications.

The paper is focused on improve recognizes the action performed on the Kinect sensor real-time skeletal tracking to understands what the user is doing, then responding exactly what they need to improve the quality of the game play.

We present a novel spatio-temporal effective body-part cross-attention transformer network (STEP-CATFormer) for skeleton-based action recognition. In the spatial dimension, it learns joint and body-part correlations for adaptively sampled joint body-part relationships, which captures the discriminative and comprehensive spatial dependencies. In the temporal dimension, it explicitly learns dynamic temporal relations with a dilation convolution transformer, enabling the network to capture rich motion patterns effectively. STEP-CATFormer high-performed on state-of-the-art on NTU RGB+D, and NTU RGB+D 120 benchmarks.

References

[1] Ryoo Aggarwal. Human activity analysis: A review. 2011.

[2] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018.

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2018.

[4] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. July 2021.

[5] Zhan Chen, Sicheng Li, Bing Yang, Qinghan Li, and Hong Liu. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition, 2022.

[6] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[7] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20186–20196, June 2022.

[8] Tomasz Danel, Przemysław Spurek, Jacek Tabor, Marek Śmieja, Łukasz Struski, Agnieszka Słowik, and Łukasz Maziarka. Spatial graph convolutional networks, 2019.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[10] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. Dg-stgcn: Dynamic spatial-temporal modeling for skeleton-based action recognition, 2022.

[11] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[12] Zhimin Gao, Peitao Wang, Pei Lv, Xiaoheng Jiang, Qidong Liu, Pichao Wang, Mingliang Xu, and Wanqing Li. Focal and global Spatial-Temporal transformer for skeleton-based action recognition. October 2022.

[13] Boeun Kim, Hyung Jin Chang, Jungho Kim, and Jin Young Choi. Global-local motion transformer for unsupervised skeleton-based action learning. July 2022.

[14] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. May 2019.

[15] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. August 2020.

[16] Zhenyue Qin, Yang Liu, Pan Ji, Dongwoo Kim, Lei Wang, Bob McKay, Saeed Anwar, and Tom Gedeon. Fusing higher-order features in graph neural networks for skeleton-based action recognition. May 2021.

[17] Bin Ren, Mengyuan Liu, Runwei Ding, and Hong Liu. A survey on 3D skeleton-based action recognition using learning method. February 2020.

[18] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang 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 (CVPR)*, June 2016.

[19] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. May 2018.

[20] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Decoupled spatial-temporal attention network for skeleton-based action recognition. July 2020.

[21] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition, 2020.

[22] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. doi: 10.1109/TPAMI.2022.3157033. URL https://doi.org/10.1109/TPAMI.2022.3157033.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[24] Qingtian Wang, Jianlin Peng, Shuze Shi, Tingxi Liu, Jiabin He, and Renliang Weng. IIP-Transformer: Intra-Inter-Part transformer for Skeleton-Based action recognition. October 2021.

[25] Wangmeng Xiang, Chao Li, Yuxuan Zhou, Biao Wang, and Lei Zhang. Language supervised training for skeleton-based action recognition, 2022.

[26] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. January 2018.

[27] Fanfan Ye, Shiliang Pu, Qiaoyong Zhong, Chao Li, Di Xie, and Huiming Tang. Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition, 2020.

[28] Wen Li Yuhan Zhang, Bo Wu. Spatial-temporal specialized transformer for skeleton-based action recognition. In *Proc. ACM MM*, 2021.

[29] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration, 2021.

[30] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. April 2019.

[31] Zhengyou Zhang. Microsoft kinect sensor and its effect. In *IEEE multimedia 19(2)*, 2012.

[32] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications, 2018.

[33] Yuxuan Zhou, Chao Li, Zhi-Qi Cheng, Yifeng Geng, Xuansong Xie, and Margret Keuper. Hypergraph transformer for skeleton-based action recognition, 2022.