Focal and Global Spatial-Temporal Transformer for Skeleton-Based Action Recognition

Zhimin Gao[1], Peitao Wang[1], Pei Lv[1], Xiaoheng Jiang[1], Qidong Liu[1],
Pichao Wang[2], Mingliang Xu[1](✉), and Wanqing Li[3]

[1] Zhengzhou University, Zhengzhou, China
{iegaozhimin,ielvpei,jiangxiaoheng,ieqdliu,iexumingliang}@zzu.edu.cn
[2] DAMO Academy, Alibaba Group (U.S.) Inc., Bellevue, USA
[3] AMRL, University of Wollongong, Wollongong, Australia
wanqing@uow.edu.au

Abstract. Despite great progress achieved by transformer in various vision tasks, it is still underexplored for skeleton-based action recognition with only a few attempts. Besides, these methods directly calculate the pair-wise global self-attention equally for all the joints in both the spatial and temporal dimensions, undervaluing the effect of discriminative local joints and the short-range temporal dynamics. In this work, we propose a novel **F**ocal and **G**lobal **S**patial-**T**emporal Trans**former** network (FG-STFormer), that is equipped with two key components: (1) FG-SFormer: focal joints and global parts coupling spatial transformer. It 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. Meanwhile, the interactions between the focal joints and body parts are incorporated to enhance the spatial dependencies via mutual cross-attention. (2) FG-TFormer: focal and global temporal transformer. Dilated temporal convolution is integrated into the global self-attention mechanism to explicitly capture the local temporal motion patterns of joints or body parts, which is found to be vital important to make temporal transformer work. Extensive experimental results on three benchmarks, namely NTU-60, NTU-120 and NW-UCLA, show our FG-STFormer surpasses all existing transformer-based methods, and compares favourably with state-of-the-art GCN-based methods.

Keywords: Action recognition · Skeleton · Spatial-temporal transformer · Focal joints · Motion patterns

1 Introduction

Human action recognition has long been a crucial and active research field in video understanding since it has a broad range of applications, such as human-computer interaction, intelligent video surveillance and robotics [4,34,44].

L. Wang et al. (Eds.): ACCV 2022, LNCS 13844, pp. 155–171, 2023.
https://doi.org/10.1007/978-3-031-26316-3_10

In recent years, skeleton-based action recognition has gained increasing attention with advent of cost-effective depth cameras like Microsoft Kinect [52] and advanced pose estimation techniques [2], which make skeleton data more accurate and accessible. By representing the action as a sequence of joint coordinates of human body, the highly abstracted skeleton data is compact and robust to illumination, human appearance changes and background noises.

Effectively modelling the spatial-temporal correlations and dynamics of joints is crucial for recognizing actions from skeleton sequences. The dominant solutions to it in recent years are the graph convolutional networks (GCNs) [46], as they can model the irregular topology of the human skeleton. Via designing advanced graph topology or traversal rules, the recognition performance is greatly improved by GCN-based methods [30,40]. Meanwhile, the recent success of Transformer [41] has gained significant interest and performance boost in various computer vision tasks [3,9,29,32]. For skeleton-based action recognition, one would expect that the self-attention mechanism in transformer shall naturally capture effective correlations of joints in both spatial and temporal dimensions for action categorization, without enforcing the articulating constrains of human body like GCN. However, there are only a few transformer-based attempts [33,38,51], and they devise hybrid model of GCN and transformer [33] or multi-task learning framework [51]. How to utilize self-attention to learn effective spatial-temporal relations of joints and representative motion features is still a thorny problem. Moreover, most of these Transformer based methods directly calculate the global one-to-one relations of joints for spatial and temporal dimensions respectively. Such strategy undervalues the spatial interactions of discriminative local joints and short-term temporal dynamics for identifying crucial action-related patterns. On the one hand, since not all joints are informative for recognizing actions [16,27], these methods suffer from the influence of irrelevant or noisy joints by accumulating the correlations with them via attention mechanism, which could harm the recognition. On the other hand, with the fact that the vanilla transformer lacks of inductive bias [29] to capture the locality of temporal structural data, it is difficult for these methods to directly model effective temporal relations of joints globally over long input sequence.

To tackle these issues, we propose a novel end-to-end **F**ocal and **G**lobal **S**patial-**T**emporal Trans**former** network, dubbed as FG-STFormer, to effectively capture relations of the crucial local joints and the global contextual information in both spatial and temporal dimensions for skeleton-based action recognition. It is composed of two components: FG-SFormer and FG-TFormer. Intuitively, each action can be distinguished by the co-movement of: (1) some critical local joints, (2) global body parts, and (or) (3) joint-part interaction. For example, as shown in Fig. 1, actions such as *taking a selfie* and *kicking* mainly involve important joints of hands, head and feet, as well as related body parts of arms and legs, while the actions like *sit down* primarily require understanding of body parts cooperation and dynamics. Based on the above observations, at the late stage of the network, we adaptively sample the informative spatial local joints (focal joints) for each action, and force the network to focus on modelling the correlations among them via multi-head self-attention without involving non-informative joints. Meanwhile, in order to compensate for the missing global co-movement and spatial structure information, we incorporate the dependencies among human body parts using self-attention. Furthermore, interactions between the body parts and the focal joints are explicitly modelled via mutual cross-attention to enhance their spatial collaboration. All of these are achieved by the proposed FG-SFormer.

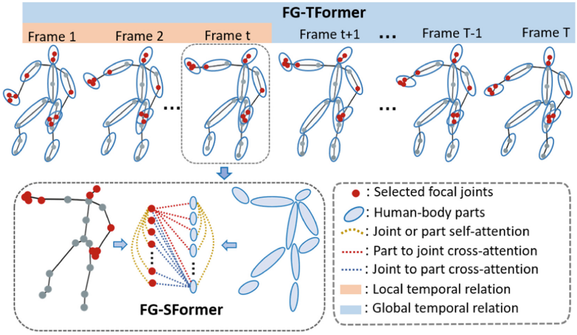
FG-TFormer
Frame 1
Frame 2
Frame t
Frame t+1
...
Frame T-1
Frame T
FG-SFormer
● : Selected focal joints
: Human-body parts
: Joint or part self-attention
: Part to joint cross-attention
: Joint to part cross-attention
: Local temporal relation
: Global temporal relation

Fig. 1. The proposed FG-SFormer (bottom) learns correlations for both adaptively selected focal joints and body parts, as well as the joint-part interactions via cross-attention. FG-TFormer (top) models the explicit local temporal relations of joints or parts, as well as the global temporal dynamics.

The FG-TFormer is designed to model the temporal dynamics of joints or body parts. It is found that straightforwardly using the vanilla temporal transformer leads to ineffective temporal relations and poor recognition performance. We found one of the key culprits lying in the absence of local bias, making it challenging for transformer to focus on effective temporal motion patterns in the long input. Taking these factors into consideration, we integrate the dilated temporal convolutions into multi-head self-attention mechanism to explicitly encode the short-term temporal motions of a joint or part from their neighbors respectively, which equips transformer with local inductive bias. The short-range feature representations of all the frames are further fused by the global self-attention weights to embrace the global contextual motion information into the representations. Thus, the designed strategy enables transformer to learn both important local and effective global temporal relations of the joints and human body parts in a unified structure, which is validated critical to make temporal transformer work.

To summarize, the contributions of this work lie in four aspects:

1. We propose a novel FG-STFormer network for skeleton-based action recognition, that can effectively capture the discriminative correlations of focal joints as well as the global contextual motion information in both the spatial and temporal dimensions.

2. We design a focal joints and global parts coupling spatial transformer, namely FG-SFormer, to model the correlations of adaptively selected focal joints and that of human body parts. The joint-part mutual cross-attention is integrated to enhance the spatial collaboration.
3. We introduce a FG-TFormer to explicitly capture both the short and long range temporal dependencies of the joints and body parts effectively.
4. The extensive experimental results on three datasets highlight the effectiveness of our method, that surpasses all existing transformer-based methods.

2 Related Work

Skeleton-Based Action Recognition. With great progress achieved in skeleton-based action recognition, existing works can be broadly divided into three groups, i.e., RNNs, CNNs, and GCNs based methods. RNNs concatenate the coordinates of all joints in one frame and treat the sequence as time series [10,19,24,49,53]. Some works design specialized network structure, like trees [26,42] to make RNN aware of spatial information. CNN based methods transform one skeleton sequence to a pseudo-image via hand-crafted manners [11,13,18,21,22,28,45], and then use popular networks to learn spatial and temporal dynamics in it.

The appearance of GCN based methods, like ST-GCN [46], enables more natural spatial topology representation of skeleton joints by organizing them as a non-Euclidean graph. The spatial correlation is modelled for bone-connected joints. As the fixed graph topology (or adjacency matrix) is not flexible to model the dependencies among spatially disconnected joints, many subsequent methods focus on designing high-order or multi-scale adjacency matrix [12,15,20,23,30], and dynamically adjusted graph topology [5,23,37,48,50]. Nevertheless, these manually devised joint traversal rules limit the flexibility to learn more effective spatial-temporal dynamics of joints for action recognition.

Transformer Based Methods. Several recent works extend Transformer [41] to spatial and temporal dimensions of skeleton-based action recognition. Among them, DSTA [38] is the first to use self-attention to learn joint relations, whereas in practice spatial transformer interleaved with temporal convolution is employed for some typical datasets. ST-TR [33] adopts a hybrid architecture of GCN and transformer in a two-stream network, with each stream replacing the GCN or temporal convolution with spatial or temporal self-attention. STST [51] introduces a transformer network that the spatial and temporal dimensions are parallelly separated. Besides, the network is trained together with multi-task self-supervised learning tasks.

3 Proposed Method

In this section, we first briefly review the basic spatial and temporal Transformer blocks (referred to as Basic-SFormer and Basic-TFormer blocks respectively) used by most existing skeleton-based action recognition methods [33,38], which is also the basics of our network. Then the proposed Focal and Global Spatial-Temporal Transformer (FG-STFormer) is introduced in detail.

3.1 Basic Spatial-Temporal Transformer on Skeleton Data

Vanilla Transformer (V-Former) Block. The vanilla transformer [41] block consists of two important modules: multi-head self-attention (MSA) and point-wise feed-forward network (FFN). Let an input composed of N elements and C-dimensional features be $X \in \mathbb{R}^{N \times C}$. For a MSA having H heads, X is first linearly projected to a set of queries Q, keys K and values V. Then, the scaled dot-product attention of head h is calculated as:

$$\text{Attention}(Q^h, K^h, V^h) = \text{softmax}(\frac{Q^h {K^h}^T}{\sqrt{d}})V^h = A^h V^h, \tag{1}$$

where Q^h, K^h, $V^h \in \mathbb{R}^{N \times d}$ with $d = C/H$ being the feature dimension of one head. $A^h \in \mathbb{R}^{N \times N}$ is the attention map.

MSA concatenates the output of all the heads and feeds into FFN module, that generally consists of a number of linear layers to transform the features.

Basic Spatial Transformer (Basic-SFormer) Block. For a skeleton sequence of T frames and N joints, let the input of C-dimension be $X = \left\{X_t \in \mathbb{R}^{N \times C}\right\}_{t=1}^{T}$. The Basic-SFormer block extends the V-Former block [41] to spatial dimension. It computes the inter-joint correlations for each frame X_t via Eq. (1) and generates an attention map $A_t^h \in \mathbb{R}^{N \times N}$, with each element $(A_t^h)_{ij}$ representing the spatial correlation score between joints i and j. Then, the features of each joint are updated as the weighted sum of values of all the joints. For the entire skeleton sequence, T spatial attention maps are produced.

Basic Temporal Transformer (Basic-TFormer) Block. By extending the V-Former to the temporal dimension, one Basic-TFormer learns global-range dynamics of a joint along the entire sequence. It rearranges the input as $X = \left\{X_n \in \mathbb{R}^{T \times C}\right\}_{n=1}^{N}$ to tackle temporal dimension. With Eq. (1), one of the N attention map $A_n^h \in \mathbb{R}^{T \times T}$ is computed for the nth joint. Each row in it stands for the dependencies of this joint across all the frames.

3.2 Focal and Global Spatial-Temporal Transformer Overview

The overview of the proposed FG-STFormer network is depicted in Fig. 2. It consists of two stages, in which our two primary components are FG-SFormer block and FG-TFormer block. The former is designed for the network late stage to model both the correlations of the sampled focal joints and the co-movement of human body parts globally in spatial dimension, as well as the interactions between the focal joints and body parts. The latter is devised to learn important local relations explicitly and global motion dynamics in temporal dimension, and is used in both stages. Therefore, the two stages are assigned specific responsibilities. That is, stage 1 aims to learn correlations for all joint pairs as generally done, so as to provide effective representations for stage 2 to mine reliable focal joints and part embeddings. Stage 2 targets at modelling both the discriminative

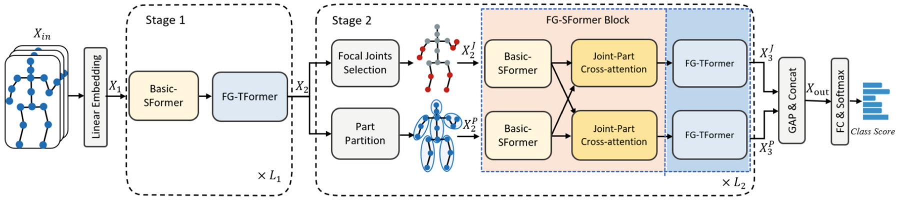

Fig. 2. Architecture of the proposed Focal and Global Spatial-Temporal Transformer (FG-STFormer). L_1 and L_2 are the number of layers in Stage 1 and Stage 2 respectively.

relations among focal joints and the global movement information of body parts. These two stages cooperate with each other to make the network learn discriminative and comprehensive spatial-temporal motion patterns for recognition.

Specifically, given a raw skeleton sequence $X_{\text{in}} \in \mathbb{R}^{N \times T \times C_0}$, a linear layer is first applied to project it from the 2D or 3D joint coordinates of C_0 to a higher dimension C_1, generating feature $X_1 \in \mathbb{R}^{N \times T \times C_1}$. Then, X_1 goes through the two successive stages of FG-STFormer. Stage 1 sequentially stacks L_1 layers with each consisting of a Basic-SFormer block and an our FG-TFormer block.

At the end of stage 1, we obtain the high-level feature representations $X_2 \in \mathbb{R}^{N \times T \times C_2}$. It is then passed into stage 2, where the network is split into two branches. One branch adaptively selects K focal joints for each frame of the sequence and discards the remaining non-informative ones, producing features $X_2^J \in \mathbb{R}^{K \times T \times C_2}$. Meanwhile, the other branch partitions the joints into P global-level human body parts and generates feature tensor $X_2^P \in \mathbb{R}^{P \times T \times C_2}$. X_2^J and X_2^P are then passed through L_2 layers that interleave FG-SFormer and FG-TFormer blocks. In particular, one FG-SFormer block consists of a Basic-SFormer sub-block and a joint-part cross-attention sub-block to sufficiently model the spatial interaction information of actions. Stage 2 then produces output features X_3^J and X_3^P from the two branches respectively. They are applied global average pooling (GAP), and then concatenated along feature channels producing features $X_{\text{out}} \in \mathbb{R}^{1 \times 1 \times C_{\text{out}}}$. With which, FG-STFormer finally performs classification using two fully connected layers and a Softmax classier.

3.3 Focal and Global Spatial Transformer (FG-SFormer)

The proposed FG-SFormer block designed for network stage 2 learns critical and comprehensive spatial structure and motion patterns based on facts in two aspects. For one aspect, there is often a subset of key joints that play a vital role in action categorization [16,27], while the other joints are irrelevant or even noisy for action analysis. Especially, for transformer-based methods, the features of one joint could be influenced by those non-informative ones when integrating features of all the joints. Therefore, it is beneficial to identify the focal joints and concentrate on them at the deep layers of the network after the shallow layers have sufficiently learned the relationships among all the joints.

For the other aspect, it is not enough to just focus on the movement of focal joints. The movement of human body parts carry crucial global contextual motion information for recognizing an action [10,14]. Meanwhile, the interactions between joints and parts convey rich kinematic information, that could be exploited to fully mine action-related patterns.

Therefore, we propose to learn relations for adaptively identified focal joints and for human body parts, as well as their interactions in spatial dimension. Three modules to achieve this are designed: (i) Focal joints selection; (ii) Global-level part partition encoding; and (iii) Joint-part cross-attention.

Focal Joints Selection. In the joint branch of stage 2, we design a ranking based strategy to adaptively sample the focal joints subset for each frame in an action sequence with the input $X_2 \in \mathbb{R}^{N \times T \times C_2}$, and discard the non-informative ones. To achieve this, we leverage a trainable projection vector $W_p \in \mathbb{R}^{C_2 \times 1}$ and sigmoid function to predict the informativeness scores $S \in \mathbb{R}^{N \times T}$ for all the joints in individual frame as:

$$S = \text{sigmoid}(X_2 W_p / ||W_p||), \tag{2}$$

Each element S_{ij} represents the informativeness score of i^{th} joint in j^{th} frame. The larger the score is, the more informative the joint is. We sort the scores of all the joints for each frame and take the features corresponding to the top K joints having the largest scores to form the features $X_2^J \in \mathbb{R}^{K \times T \times C_2}$ of the focal joints subset as:

$$\begin{aligned} \text{idx} &= \text{sort}(S, K), \\ X_2^J &= X_2(\text{idx}, :, :), \end{aligned} \tag{3}$$

where idx is the indices of the selected joints with largest scores.

X_2^J is then fed into the Basic-SFormer sub-block introduced in Sect. 3.1 to calculate the correlations only for those focal joints and update their feature embeddings. The Basic-SFormer block/sub-block used in both stages 1 and 2 is depicted in Fig. 3 (a). It uses the sine and cosine position encoding [41] to encode the joint type information. In the MSA module with H heads, the spatial attention map A_t is calculated for each frame. As in [38], we add a global regularization attention map A_g shared by all the sequences. The FFN module consists of a linear layer followed by applying activation function of Leaky ReLU [31].

Global-Level Part Partition Encoding. We explicitly model the correlations between global-level body parts in the other branch of stage 2 in our FG-STFormer. The joints are partitioned into P parts based on the physical skeleton structure and human prior. To obtain feature embeddings of the P parts with X_2, we concatenate the features of joints belonging to the same body part and then transform them into one part-level feature embedding via a linear layer shared by all parts. This generates the part embedding $X_2^P \in \mathbb{R}^{P \times T \times C_2}$, which is then passed into the Basic-SFormer sub-block shown in Fig. 3 (a) to model the one-to-one part relations and update features correspondingly.

Joint-Part Cross-Attention. To enable information diffusion across the focal joints and body parts to model their co-movement, we devise a joint-part cross-attention sub-block, termed as JP-CA. It uses multi-head cross-attention to

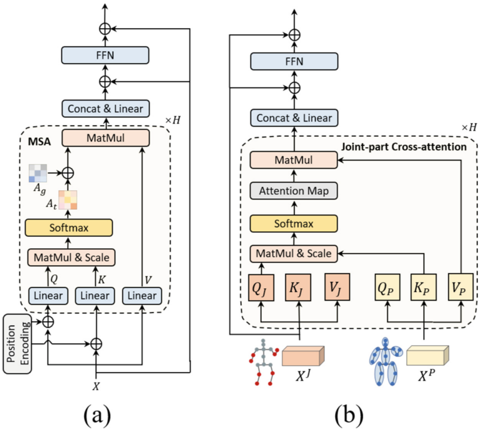

Fig. 3. (a) The Basic-SFormer block/sub-block used in Stages 1 and 2. (b) The Joint-part cross-attention (JP-CA) sub-block used in FG-SFormer block.

interact and diffuse features of the two branches. Here, we present JP-CA from the part branch to the focal joint branch as an example, as shown in Fig. 3 (b). For notational convenience, we omit the subscripts of X_2^J and X_2^P. Let Q_J, K_J and V_J be the queries, keys and values mapped from the joint-branch features X^J, and Q_P, K_P and V_P be those from the part-branch features X^P respectively. The part-to-joint cross-attention takes the Q_J as queries, and K_P and V_P as keys and values, and is calculated as:

$$\text{Attention}(Q_J, K_P, V_P) = \text{softmax}(\frac{Q_J {K_P}^T}{\sqrt{d}})V_P = A_{jp} V_P, \tag{4}$$

where d is the feature dimension of one head.

The attention map $A_{jp} \in \mathbb{R}^{K \times P}$ models the joint-part correlations and is used to aggregate part features for each focal joint. Other operations in this sub-block is same as those in the adopted Basic-SFormer sub-block. Notably, JP-CA is adaptive to actions, which is flexible to capture distinct collaborative patterns for input actions. Analogously, the cross-attention from the joint-branch to part-branch can be defined in similar operations.

3.4 Focal and Global Temporal Transformer (FG-TFormer)

Though temporal transformer has been applied in skeleton-based action recognition in existing works [1,33,51], it is rarely effectively deployed solely with spatial transformer in a single-stream architecture or in pure transformer-based models, largely because: (i) it is difficult for the self-attention to directly model effective temporal relations globally for distinguishing actions over the long input sequence; and (ii) the lack of inductive biases of transformer.

To address these issues, we propose to assist transformer in focusing on both the important local and the global temporal relations of joints explicitly, and design the component of focal and global temporal self-attention (FG-TSA), as depicted in Fig. 4. It utilizes the dilated temporal convolution to generate the values V in MSA, that works in two aspects: (i) explicitly learning the short-term temporal motion representation of a joint from its neighboring frames; and (ii) introducing beneficial local inductive biases to transformer. Meanwhile, the attention map generated by MSA models the global temporal correlations. Therefore, the resulting fused joint representations integrate both local temporal relation and global contextual information. The same effect is also achieved for the body part representations when FG-TFormer is applied to the part-branch.

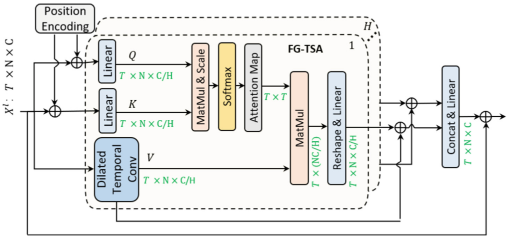

Fig. 4. The pipeline of the proposed FG-TFormer block.

Specifically, let the input feature tensor of the FG-TFormer block in layer l ($l = 1, 2, ..., L$) be $X^l \in \mathbb{R}^{N \times T \times C}$, where L is the total number of layers. First, the absolute position encoding is used to encode the temporal order information. And then, for the queries and keys Q^l_h, $K^l_h \in \mathbb{R}^{N \times T \times \frac{C}{H}}$ in head h among the H ones, they are generated via the usual linear projection in FG-TSA. Whereas for the V^l_h, different from existing works, we utilize dilated temporal convolution with kernel size $k_t \times 1$ and dilation rate d_t to obtain it, denoted as $V^l_h = \mathrm{TCN}_{\mathrm{dilate}}(X^l)_h \in \mathbb{R}^{N \times T \times \frac{C}{H}}$. Then, the global self-attention map is calculated via Eq. (1) and utilized to fuse the local feature representation V^l_h. Hence, each joint representation is injected with its global contextual dynamics. The features are then reshaped and linearly transformed with $W_h \in \mathbb{R}^{\frac{C}{H} \times \frac{C}{H}}$. This is followed by concatenating the output features of all the heads and conducting linear transform with $W_O \in \mathbb{R}^{C \times C}$ using activation function of Leaky ReLU. The output of the FG-TFormer block X^{l+1} is obtained by adding the shortcut from the input X^l. The whole process is formulated as:

$$\begin{aligned} X^{l+1} &= \mathrm{Concat}[\mathrm{head}(X^l)_1, ..., \mathrm{head}(X^l)_H] W_O + X^l, \\ \mathrm{head}(X^l)_h &= [\mathrm{Attention}(Q^l_h, K^l_h) \mathrm{TCN}_{\mathrm{dilate}}(X^l)_h] W_h + \mathrm{TCN}_{\mathrm{dilate}}(X^l)_h. \end{aligned} \tag{5}$$

Besides, we halve the temporal resolution of a sequence during generating V with convolution of stride 2 when the feature channels are doubled for a FG-TFormer block. This hierarchical structure reduces the computational cost.

4 Experiments

4.1 Dataset

NTU-RGB+D 60 (NTU-60) [35] contains 56,880 sequences in 60 classes. It is collected from 40 subjects and provides the 3D locations of 25 human body joints. It recommends two benchmarks for evaluation: (1) Cross-subject (X-Sub): training data is from 20 subjects and test data from the other 20 subjects. (2) Cross-view (X-View): sequences captured by camera views 2 and 3 are taken as training data, and those captured by camera view 1 as testing data.

NTU-RGB+D 120 (NTU-120) [25] has 120 classes and 113,945 samples captured from 32 camera setups and 106 subjects. It recommends two benchmarks: (1) Cross-subject (X-Sub): training data is from 53 subjects, and test data from the other 53 subjects. (2) Cross-setup (X-Set): samples with even setup IDs are used for training data, and samples with odd setup IDs for test.

Northwestern-UCLA (NW-UCLA) [43] is captured by three Kinect cameras from three viewpoins. It contains 1,494 sequences in 10 action categories, with each performed by 10 actors. The same evaluation protocol in [26] is used: training data from the first two cameras and test data from the other camera.

4.2 Implementation Details

Our FG-STFormer model consists of 8 layers in two stages. Stage 1 contains $L_1 = 6$ layers and stage 2 consists of $L_2 = 2$ layers. The channel dimensions of each layer are 64, 64, 128, 128, 256, 256, 256 and 256. The number of frames is halved at the third and fifth layers. The number of spatial attention heads for the Basic-SFormer and FG-SFormer blocks is set to be 3. Each FG-TFormer block uses 2 attention heads, which adopt temporal kernel size of $k_t = 7$, and dilation rates of $d_t = 1$ and $d_t = 2$ respectively. The numbers of focal joints and body parts in the two branches of stage 2 are $K = 15$ and $P = 10$ respectively.

All experiments are conducted on one RTX 3090 GPU with PyTorch framework. We use SGD with Nesterov momentum of 0.9 and weight decay of 0.0005 to train our model for 80 epochs. Warm up strategy is used for the first 5 epochs. The initial learning rate is set to 0.01 and decays by a factor of 10 at the 50th and 70th epochs. For NTU-60 and NTU-120, the batch size is 32. The sequences are sampled to 128 frames, and we use the data pre-processing method in [5]. For NW-UCLA, the batch size is 32, we use the same data pre-processing in [43].

4.3 Ablation Study

In this section, the effectiveness of individual component of FG-STFormer is evaluated under X-Sub protocol of NTU-60 dataset, using only the joint stream.

Effectiveness of FG-SFormer Block. To examine the effectiveness of the proposed FG-SFormer, we evaluate the important components in it, i.e., focal joints selection, part branch and joint-part cross-attention (JP-CA). We employ the Basic-SFormer as baseline, which calculates the correlations for all the joints at every layer of the network without using part branch and JP-CA. For temporal modelling, our FG-TFormer is used. We gradually replace the baseline by adding our designs one-by-one. The experimental results are shown in Table 1.

Table 1. Ablation study of different components in FG-SFormer block.

Methods	Focal Joints Selection	Part Branch	JP-CA	Acc (%)
Basic-SFormer	–	–	–	87.8
A	√	–	–	88.3
B	√	√	–	89.1
C	√	√	√	**89.5**

As seen, model A selects the focal joints at stage 2 of the network and improves the performance of Basic-SFormer by 0.5%. This indicates that it is beneficial to identify discriminative joints. Then, model B introduces the part branch to network stage 2. This provides performance improvement of 0.8% and reflects the spatial relations of intra-parts carry helpful global motion patterns. Finally, by adding the JP-CA into model C, the accuracy is further increased by 0.4%. This implies that the interactions between body parts and the selected focal joints are helpful for distinguishing actions.

Impact of Number of Focal Joints. To explore the effect of selecting different number of focal joints, we test the models using different K in FG-SFormer blocks at stage 2. Note that $K = 25$ means all the joints are used. As shown in Table 2, the accuracy gradually improves as K increases from 3 to 15, and then decreases when it further increases. This implies that the redundant or noisy joints indeed harm the recognition performance. In addition, too small number of focal joints are not enough to accurately identify the actions.

Table 2. Comparison of classification accuracy using different number of focal joints.

K	3	6	9	12	15	18	21	25
Acc (%)	88.6	88.9	89.0	89.2	**89.5**	89.2	89.3	89.1

Effectiveness of FG-TFormer Block. To evaluate the efficacy of FG-TFormer block, we build up experiments based on the complete network by modifying the FG-TFormer block only. The model using Basic-TFormer is taken as the baseline, which solely adopts the global MSA in temporal dimension. According to the results shown in Table 3, without the $\text{TCN}_{\text{dilate}}$ in MSA, the Basic-TFormer performs significantly worse than our FG-TFormer with a large margin of −6.7%. Besides, by replacing Basic-TFormer with $\text{TCN}_{\text{dilate}}$, the performance is greatly improved by 6.3%. Finally, our FG-TFormer further achieves improvement of 0.4% by integrating $\text{TCN}_{\text{dilate}}$ into self-attention mechanism.

Table 3. Ablation study for components in FG-TFormer block.

Methods	Global MSA	$\text{TCN}_{\text{dilate}}$	Acc (%)
Basic-TFormer	√	×	82.8
FG-TFormer	×	√	89.1
	√	√	**89.5**

Configuration Exploration. We explore different network configurations for stages 1 and 2 in our FG-STFormer by adjusting the number of layers L_1 and L_2. The total number of layers is fixed as 8. The results are shown in Table 4. Comparing models A, B and C, we can find that higher performance is obtained with more than 4 layers used in stage 1, and the best performance is achieved by $L_1 = 6$ and $L_2 = 2$. The accuracy drops down when stage 2 is assigned less layers in model D. These observations indicate that it is necessary for stage 1 to sufficiently learn the relations among all the joints, otherwise the performance could be harmed by focusing on unreliable focal joints and part collaborations.

Table 4. Comparison of different network configurations of our FG-STFormer.

Methods	Stage 1	Stage 2	Acc (%)	Methods	Stage 1	Stage 2	Acc (%)
	L_1	L_2			L_1	L_2	
A	4	4	88.8	C	6	2	**89.5**
B	5	3	89.0	D	7	1	89.0

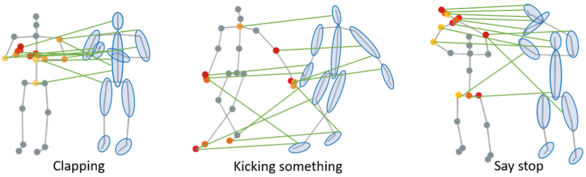

Fig. 5. The selected focal joints and learned joint-part interactions of actions. (Color figure online)

4.4 Visualization and Analysis

To validate what the focal joints are concentrated on at stage 2, we visualize the sampled 13 focal joints having largest scores for three actions in Fig. 5. These focal joints are depicted as coloured dots in the left skeleton of each action. The darker the dot is, the higher the informativeness score is of the joint. We can see that the actions *Clapping*, *Kicking something* and *Say stop* mainly select hands, shoulders, elbows and feet as the focal joints. Besides, Fig. 5 illustrates the learned attention weights from parts to focal joints of these actions. Attentions with large values are shown as green lines. As seen, the actions *Clapping* and *Say stop* mainly build interactions between focal joints and upper limbs, while action *Kicking something* interacts between focal joints and the whole body parts. These results verify that the spatial relations between the key joints and the global contextual movement information are captured by our FG-SFormer.

Moreover, we compare the performance of the Basic-TFormer with our FG-TFormer on action classes that the former has low accuracy. As shown in Fig. 6, our network improves the performance of those exhibited classes, which mainly involve the subtle and fine-grained motions of hands, feet and head. This concludes that our FG-TFormer can capture those subtle interaction patterns via explicitly embedding the neighboring relations into it.

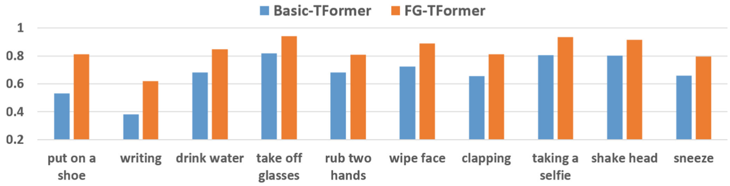

Fig. 6. Accuracy comparison between the Basic-TFormer and our FG-TFormer blocks.

Table 5. Comparison to state-of-the-arts on NW-UCLA dataset.

Methods	Year	NW-UCLA Top-1 (%)
HBRNN-L [10]	2015	78.5
Ensemble TS-LSTM [19]	2017	89.2
AGC-LSTM [39]	2019	93.3
Shift-GCN [8]	2020	94.6
DC-GCN+ADG [7]	2020	95.3
CTR-GCN [5]	2021	96.5
FG-STFormer (ours)	2022	**97.0**

4.5 Comparison with the State-of-the-Arts

We compare our FG-STFormer with existing state-of-the-art (SOTA) methods on three datasets: NW-UCLA, NTU-60 and NTU-120. Following the previous works [30,38,51], we fuse results of four modalities, i.e., joint, bone, joint motion, and bone motion. The results are shown in Tables 5 and 6. As seen, our method outperforms all existing transformer-based methods under nearly all evaluation benchmarks on NTU-60 and NTU-120, including the latest method STST [51] which uses not only the parallel spatial and temporal transformers but also multiple self-supervised learning tasks, and ST-TR [33] which adopts hybrid architecture of spatial-temporal transformer and GCN. Our method surpasses DSTA [38] by 2.4% and 1.6% on the two evaluation protocols of NTU-120.

Table 6. Performance comparisons against the SOTA methods on NTU- 60 and 120.

Methods	Year	NTU-60		NTU-120	
		X-Sub (%)	X-View (%)	X-Sub (%)	X-Set (%)
GCN-based methods					
ST-GCN [46]	2018	81.5	88.3	70.7	73.2
2s-AGCN [37]	2019	88.5	95.1	82.9	84.9
DGNN [36]	2019	89.9	96.1	–	–
Shift-GCN [8]	2020	90.7	96.5	85.9	87.6
Dynamic GCN [48]	2020	91.5	96.0	87.3	88.6
MS-G3D [30]	2020	91.5	96.2	86.9	88.4
MST-GCN [6]	2021	91.5	96.6	87.5	88.8
CTR-GCN [5]	2021	92.4	96.8	88.9	**90.6**
STF [17]	2022	92.5	**96.9**	88.9	90.0
Transformer-based methods					
DSTA [38]	2020	91.5	96.4	86.6	89.0
ST-TR [33]	2021	89.9	96.1	82.7	84.7
UNIK [47]	2021	86.8	94.4	80.8	86.5
STST [51]	2021	91.9	96.8	–	–
FG-STFormer (ours)	2022	**92.6**	96.7	**89.0**	**90.6**

Moreover, compared to GCN-based methods, the performance of our FG-STFormer is also at the top. It compares favourably with current state-of-the-art STF [17] and CTR-GCN [5] on NTU-60 and NTU-120, and even outperforms the latter on NW-UCLA by 0.5%, verifying the effectiveness of FG-STFormer.

5 Conclusion

In this work, we present a novel focal and global spatial-temporal transformer network (FG-STFormer) for skeleton-based action recognition. In spatial dimension, it learns intra- and inter- correlations for adaptively sampled focal joints and global body parts, which captures the discriminative and comprehensive spatial dependencies. In temporal dimension, it explicitly learns both the local and global temporal relations, enabling the network to capture rich motion patterns effectively. On three datasets, the proposed FG-STFormer achieves the state-of-the-art performance, demonstrating the effectiveness of our method.

Acknowledgement. The work of Zhimin Gao and Peitao Wang was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61906173. The work of Xiaoheng Jiang, Qidong Liu, and Mingliang Xu was supported by NSFC under U21B2037, 62276238 and 62036010 respectively.

References

1. Bai, R., et al.: GCST: graph convolutional skeleton transformer for action recognition. arXiv preprint arXiv:2109.02860 (2021)
2. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans. PAMI **43**(1), 172–186 (2019)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of CVPR, pp. 6299–6308 (2017)
5. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of ICCV, pp. 13359–13368 (2021)
6. Chen, Z., Li, S., Yang, B., Li, Q., Liu, H.: Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In: Proceedings of AAAI, vol. 35, pp. 1113–1122 (2021)
7. Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J., Lu, H.: Decoupling GCN with DropGraph module for skeleton-based action recognition. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12369, pp. 536–553. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58586-0_32
8. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-based action recognition with shift graph convolutional network. In: Proceedings of CVPR, pp. 183–192 (2020)
9. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. In: ICLR (2020)
10. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: CVPR, pp. 1110–1118 (2015)
11. Duan, H., Zhao, Y., Chen, K., Shao, D., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. arXiv preprint arXiv:2104.13586 (2021)
12. Gao, X., Hu, W., Tang, J., Liu, J., Guo, Z.: Optimized skeleton-based action recognition via sparsified graph regression. In: Proceedings of ACM MM, pp. 601–610 (2019)
13. Hou, Y., Li, Z., Wang, P., Li, W.: Skeleton optical spectra-based action recognition using convolutional neural networks. IEEE Trans. CSVT **28**(3), 807–811 (2018)
14. Huang, L., Huang, Y., Ouyang, W., Wang, L.: Part-level graph convolutional network for skeleton-based action recognition. In: Proceedings of AAAI, vol. 34, pp. 11045–11052 (2020)
15. Huang, Z., Shen, X., Tian, X., Li, H., Huang, J., Hua, X.S.: Spatio-temporal inception graph convolutional networks for skeleton-based action recognition. In: Proceedings of ACM MM, pp. 2122–2130 (2020)

16. Jiang, M., Kong, J., Bebis, G., Huo, H.: Informative joints based human action recognition using skeleton contexts. Signal Process. Image Commun. **33**, 29–40 (2015)
17. Ke, L., Peng, K.C., Lyu, S.: Towards to-at spatio-temporal focus for skeleton-based action recognition. In: Proceedings of AAAI (2022)
18. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3D action recognition. In: Proceedings of CVPR, pp. 3288–3297 (2017)
19. Lee, I., Kim, D., Kang, S., Lee, S.: Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks. In: Proceedings of ICCV, pp. 1012–1020 (2017)
20. Li, B., Li, X., Zhang, Z., Wu, F.: Spatio-temporal graph routing for skeleton-based action recognition. In: Proceedings of AAAI, vol. 33, pp. 8561–8568 (2019)
21. Li, C., Xie, C., Zhang, B., Han, J., Zhen, X., Chen, J.: Memory attention networks for skeleton-based action recognition. IEEE Trans. NNLS **33**(9), 4800–4814 (2021)
22. Li, C., Zhong, Q., Xie, D., Pu, S.: Skeleton-based action recognition with convolutional neural networks. In: ICMEW, pp. 597–600. IEEE (2017)
23. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of CVPR, pp. 3595–3603 (2019)
24. Li, S., Li, W., Cook, C., Zhu, C., Gao, Y.: Independently recurrent neural network (IndRNN): building a longer and deeper RNN. In: Proceedings of CVPR, pp. 5457–5466 (2018)
25. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding. IEEE Trans. PAMI **42**(10), 2684–2701 (2019)
26. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal LSTM with trust gates for 3D human action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 816–833. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_50
27. Liu, J., Wang, G., Hu, P., Duan, L.Y., Kot, A.C.: Global context-aware attention LSTM networks for 3D action recognition. In: Proceedings of CVPR (2017)
28. Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. Pattern Recogn. **68**, 346–362 (2017)
29. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of ICCV, pp. 10012–10022 (2021)
30. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of CVPR, pp. 143–152 (2020)
31. Maas, A.L., Hannun, A.Y., Ng, A.Y., et al.: Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of ICML, vol. 30, p. 3 (2013)
32. Neimark, D., Bar, O., Zohar, M., Asselmann, D.: Video transformer network. In: Proceedings of ICCV, pp. 3163–3172 (2021)
33. Plizzari, C., Cannici, M., Matteucci, M.: Skeleton-based action recognition via spatial and temporal transformer networks. Comput. Vis. Image Underst. **208**, 103219 (2021)
34. Poppe, R.: A survey on vision-based human action recognition. Image Vis. Comput. **28**(6), 976–990 (2010)
35. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+D: a large scale dataset for 3D human activity analysis. In: Proceedings of CVPR, pp. 1010–1019 (2016)

36. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with directed graph neural networks. In: Proceedings of CVPR, pp. 7912–7921 (2019)
37. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of CVPR, pp. 12026–12035 (2019)
38. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In: Proceedings of ACCV (2020)
39. Si, C., Chen, W., Wang, W., Wang, L., Tan, T.: An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In: Proceedings of CVPR, pp. 1227–1236 (2019)
40. Song, Y.F., Zhang, Z., Shan, C., Wang, L.: Constructing stronger and faster baselines for skeleton-based action recognition. IEEE Trans. PAMI **45**(2), 1474–1488 (2022)
41. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
42. Wang, H., Wang, L.: Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In: Proceedings of CVPR, pp. 499–508 (2017)
43. Wang, J., Nie, X., Xia, Y., Wu, Y., Zhu, S.C.: Cross-view action modeling, learning and recognition. In: Proceedings of CVPR, pp. 2649–2656 (2014)
44. Wang, P., Li, W., Ogunbona, P., Wan, J., Escalera, S.: RGB-D-based human motion recognition with deep learning: a survey. Comput. Vis. Image Underst. **171**, 118–139 (2018)
45. Wang, P., Li, Z., Hou, Y., Li, W.: Action recognition based on joint trajectory maps using convolutional neural networks. In: Proceedings of ACM MM, pp. 102–106 (2016)
46. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of AAAI (2018)
47. Yang, D., Wang, Y., Dantcheva, A., Garattoni, L., Francesca, G., Bremond, F.: Unik: a unified framework for real-world skeleton-based action recognition. In: Proceedings of BMVC (2021)
48. Ye, F., Pu, S., Zhong, Q., Li, C., Xie, D., Tang, H.: Dynamic GCN: context-enriched topology learning for skeleton-based action recognition. In: Proceedings of ACM MM, pp. 55–63 (2020)
49. Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N.: View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In: Proceedings of ICCV, pp. 2117–2126 (2017)
50. Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., Zheng, N.: Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of CVPR, pp. 1112–1121 (2020)
51. Zhang, Y., Wu, B., Li, W., Duan, L., Gan, C.: STST: spatial-temporal specialized transformer for skeleton-based action recognition. In: Proceedings of ACM MM, pp. 3229–3237 (2021)
52. Zhang, Z.: Microsoft Kinect sensor and its effect. IEEE Multimedia **19**(2), 4–10 (2012)
53. Zhu, W., et al.: Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In: Proceedings of AAAI, vol. 30 (2016)