IGFormer: Interaction Graph Transformer for Skeleton-based Human Interaction Recognition

Yunsheng Pang[1], Qiuhong Ke[1,2] *, Hossein Rahmani[3], James Bailey[1], and Jun Liu[4]

[1] The University of Melbourne, Australia
yunshengp@student.unimelb.edu.au, baileyj@unimelb.edu.au
[2] Monash University, Australia
Qiuhong.Ke@monash.edu
[3] Lancaster University, United Kingdom
h.rahmani@lancaster.ac.uk
[4] Singapore University of Technology and Design, Singapore
jun_liu@sutd.edu.sg

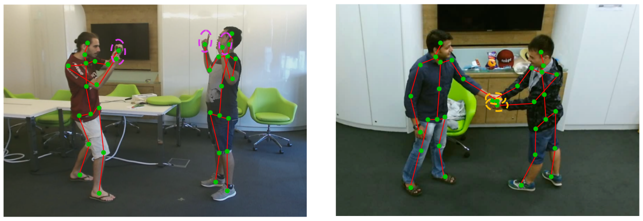
Abstract. Human interaction recognition is very important in many applications. One crucial cue in recognizing an interaction is the interactive body parts. In this work, we propose a novel Interaction Graph Transformer (IGFormer) network for skeleton-based interaction recognition via modeling the interactive body parts as graphs. More specifically, the proposed IGFormer constructs interaction graphs according to the semantic and distance correlations between the interactive body parts, and enhances the representation of each person by aggregating the information of the interactive body parts based on the learned graphs. Furthermore, we propose a Semantic Partition Module to transform each human skeleton sequence into a Body-Part-Time sequence to better capture the spatial and temporal information of the skeleton sequence for learning the graphs. Extensive experiments on three benchmark datasets demonstrate that our model outperforms the state-of-the-art with a significant margin.

Keywords: Transformer, Skeleton-based Human Action Recognition, Human Interaction Recognition.

1 Introduction

Human interaction recognition plays a significant role in a wide range of applications [1,26,36,31]. For example, it can be used in visual surveillance to detect dangerous events such as "kicking" and "punching". It can also be used for robot controlling for human-robot interaction. This paper addresses human interaction recognition from skeleton sequences [28,15]. Compared with RGB videos, skeleton sequences provide only 3D coordinates of human joints, which are more robust to unconventional and variable conditions, such as unusual viewpoints and cluttered backgrounds.

* Corresponding author



(a) Taking a photo

(b) Shaking hands

Fig. 1. (a) In the interaction of “Taking a photo”, there is strong semantic correlation between the hands holding the camera of one person and the hands with “yeah” of the other person. (b) In the interaction of “Shaking hands”, the interactive body parts demonstrate both semantic correlation and distance evolution, i.e., the hands of two interactive persons correspond to each other and are gradually close to each other when they are shaking hands.

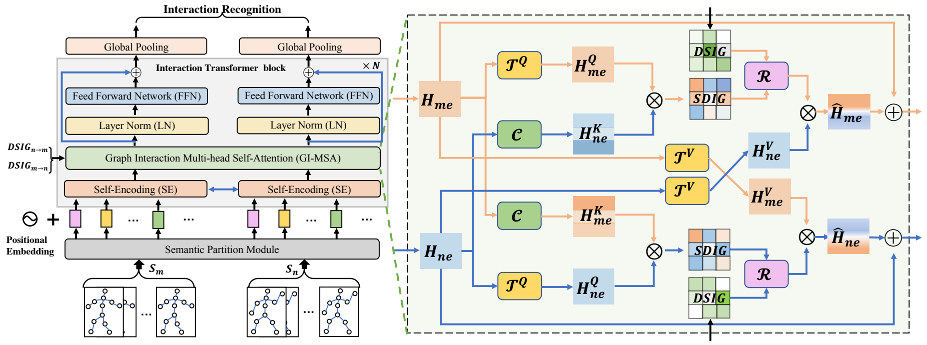
Compared with single-person action recognition, one additional crucial cue in recognizing a human interaction is the interactive body parts of the interactive persons. For example, the interactive hands of two persons are critical in understanding a “shaking hands” interaction. Generally, the interactive body parts in interactions demonstrate semantic correlations and correspondence. For example, in the interaction of “Taking a photo” shown in Fig. 1 (a), the hands holding the camera of one person and the hands with “yeah” of the other person demonstrate a strong correlation. Similarly, in “Shaking hands” shown in Fig. 1 (b), the interactive hands of the two persons correspond to each other. In these cases, exploring the semantic correlation between the interactive body parts is crucial for interaction understanding. In addition, for some interactions, the interactive body parts demonstrate distance evolution. For example, the hands of the two persons gradually approach each other when the two persons are “shaking hands”. Measuring the distance between body parts of the interactive persons can provide additional useful information to the semantic correlation for better interaction recognition.

Inspired by the above observation and the successful application of Transformer in many fields [4,5,41,37], we propose a novel Transformer-based model named Interaction Graph Transformer (IGFormer) for interaction recognition from skeleton sequences. In particular, the proposed IGFormer consists of a Graph Interaction Multi-head Self-Attention (GI-MSA) module, which aims at modeling the relationship of interactive persons from both semantic and distance levels to recognize actions. More specifically, the GI-MSA module learns a semantic-based graph and a distance-based interaction graph to represent the mutual relationship between body parts of the interactive persons. The semantic-based graph is learned by the attention mechanism in a data-driven manner to capture the semantic correlations of the interactive body parts. The distance-based graph is constructed by measuring the distance between pairs of body parts to excavate the distance information between interactive body parts. The two interaction graphs are combined to complement each other in a refinement way, making the model suitable for modeling different interactions.

To feed skeleton sequences to the IGFormer, one straightforward solution is to transform each skeleton sequence to a pseudo-image and divide the image into a sequence of patches, similar to the manner of ViT [5]. However, this may destroy the spatial relationship among the skeleton joints in each body part, which could hinder effective modeling of the interactive body parts for interaction recognition. To tackle this problem, we propose a Semantic Partition Module (SPM) to transform the skeleton sequence of each subject into a new format, i.e., a Body-Part-Time (BPT) sequence, each of which is the representation of one body part during a short period. The BPT sequence encodes semantic information and temporal dynamics of the body parts, enhancing the capability of the network for modeling interactive body parts for interaction recognition.

We summarize the contributions of this paper as follows:

- We introduce a Transformer-based model named IGFormer, which contains a novel GI-MSA module to learn the relationships of the interactive persons from both semantic and distance levels for skeleton-based human interaction recognition.
- We introduce a Semantic Partition Module (SPM) transforming each skeleton sequence into a BPT sequence to enhance the modeling of interactive body parts.
- We conduct extensive experiments on three challenging datasets and achieve state-of-the-art performance.



(a) The overall architecture of IGFormer

(b) A detailed version of GI-MSA

Fig. 2. (a) The overall architecture of the proposed IGFormer for skeleton-based human interaction recognition. Given the skeleton sequences of two subjects, they are first fed into the Semantic Partition Module (SPM) to generate two Body-part-time (BPT) sequences. The BPT sequences are then fed into the Interaction Transformer Block (ITB) for interactive learning. The ITB contains three main components: two self-encoding (SE) modules, the proposed GI-MSA module and two two-layer Feed-Forward Networks (FFN). Finally, a global average pooling followed by a softmax classifier is applied to the outputs of the last ITB to predict interaction labels. (b) The structure of the proposed Graph Interaction Multi-head Self-Attention (GI-MSA) module (DSIG: distance-based sparse interaction graph, SDIG: semantic-based dense interaction graph).

2 Related Work

2.1 Skeleton-based Action Recognition

Conventional deep learning-based methods model the human skeleton as a sequence of joint-coordinate vectors [18,28,7,30,35,13] or a pseudo-image [14,9,10,11,6], which is then fed into RNNs or CNNs to predict the actions. However, representing the skeleton data as a vector sequence or a 2D grid cannot fully express the dependency between correlated joints since the human skeleton is naturally structured as a graph. Recently, GCN-based methods [12,29,23] consider the human skeleton as a graph whose vertices are joints and edges are bones and apply graph convolutional networks (GCN) on the human graph to extract correlated features. These methods achieve better performance than RNN- and CNN-based methods, and become the mainstream methods in skeleton-based action recognition. However, these methods consider each person as an independent entity and cannot effectively capture human interaction. In this work, we focus on skeleton-based human interaction recognition and propose to model the interactive relationship of persons from both semantic and distance levels.

2.2 Human Interaction Recognition

Human interaction recognition [36,31,27] is a sub-field of action recognition. Compared with single-person action recognition, human interaction methods should not only be able to model the behavior of each individual but also capture the interaction between them. Yun et al. [34] evaluated several geometric relational body-pose features including joint features, plane features and velocity features for interaction modeling, and found out that joint features outperform others, whereas velocity features are sensitive to noise. Ji et al. [8] built poselets by grouping joints that belong to the same body part of each individual to describe the interaction of each body part. Recently, Perez et al. [24] proposed a two-stream LSTM-based interaction relation network called LSTM-IRN to model the intra relations of body joints from the same person and the inter relations of the joints from different persons. However, LSTM-IRN ignores the distance evolution of body parts, which is considered as an important prior knowledge for human interaction recognition. Different from the above-mentioned methods, we model the interaction relationship of interactive humans as two interaction graphs, which are constructed from the semantic and distance levels respectively to capture the semantic correlation and distance evolution between body parts.

2.3 Visual Transformer

Transformer was first proposed in [32] for machine translation task and since then has been widely adopted in various natural language processing (NLP) tasks. Inspired by the successful application in NLP, Transformer has been applied to the computer vision and demonstrated its scalability and effectiveness in many vision tasks. Vision Transformer (ViT) [5] was the first pure Transformer architecture for image recognition and obtained better performance and generalization than traditional convolutional neural networks (CNNs). After that, Transformer-based models with carefully designed and complicated architectures have been applied to various downstream vision tasks, such as object detection [40], semantic segmentation [38] and video classification [2]. In skeleton-based action recognition, Plizzari et al. [25] proposed ST-TR to model the dependencies between joints by substituting the graph convolution operator with the self-attention operator. Different from ST-TR, we focus on human interaction modeling and propose a novel self-attention-based GI-MSA module to model the correlations between body parts of interactive persons.

3 Interaction Graph Transformer

One important cue in recognizing human interaction is the interactive body parts. In this section, we introduce an Interaction Graph Transformer (IGFormer), which contains a Graph Interaction Multi-head Self-Attention (GI-MSA) module to model the interactive body parts at both semantic and distance levels for skeleton-based interaction recognition. The proposed IGFormer is also equipped with a Semantic Partition Module (SPM), which aims at retaining the semantic and temporal information of each body part within the input skeleton sequences for better learning of the interactive body parts.

The overall architecture of the proposed IGFormer is shown in Fig. 2 (a). Given the skeleton sequences of two interactive subjects $\mathbf{S}_m, \mathbf{S}_n \in \mathbb{R}^{T\times J\times C}$, where T and J represent the numbers of frames and joints in each frame, respectively, and $C = 3$ represents the dimension of the 3D coordinates of each joint, we first feed the two skeletons into the proposed SPM to generate two Body-Part-Time (BPT) sequences, $\mathbf{H}_m, \mathbf{H}_n$, which are then fed into a stack of *Interaction Transformer Blocks (ITBs)* for interaction modeling. Finally, a global average pooling followed by a softmax classifier is applied to the output of the last ITB to predict the interaction class.

More specifically, each ITB contains three components including two shared-weight self-encoding (SE) modules, the Graph Interaction Multi-head Self-Attention (GI-MSA) module, and two Feed-Forward Networks (FFN). Each SE module is a standard one-layer Transformer [5], which aims at modeling the interaction among the body parts within each individual skeleton. The two outputs of the SE are fed into the GI-MSA to model the interactive body parts and generate an enhanced representation for each interactive person. Finally, each output of the GI-MSA is fed to a Layer Normalization (LN) followed by a FFN. We add an addition operation between the output of GI-MSA and FFNs to improve the representation capability of the model. The ITB can be formulated as follows:

$$\begin{aligned}
&\mathbf{H}_{me}, \mathbf{H}_{ne} = \mathrm{SE}(\mathbf{H}_m), \mathrm{SE}(\mathbf{H}_n),\\
&\hat{\mathbf{H}}_{me}, \hat{\mathbf{H}}_{ne} = \text{GI-MSA}(\mathbf{H}_{me}, \mathbf{H}_{ne}),\\
&\hat{\mathbf{H}}_{mo} = \mathrm{FFN}(\mathrm{LN}(\hat{\mathbf{H}}_{me})) + \hat{\mathbf{H}}_{me},\\
&\hat{\mathbf{H}}_{no} = \mathrm{FFN}(\mathrm{LN}(\hat{\mathbf{H}}_{ne})) + \hat{\mathbf{H}}_{ne},
\end{aligned} \tag{1}$$

where $\mathbf{H}_{me}$ and $\mathbf{H}_{ne}$ denote the outputs of the SE, $\hat{\mathbf{H}}_{me}$ and $\hat{\mathbf{H}}_{ne}$ denote the outputs of the GI-MSA module, and $\hat{\mathbf{H}}_{mo}$ and $\hat{\mathbf{H}}_{no}$ are the outputs of the ITB.

The two SE modules in the first ITB take the Body-Part-Time (BPT) representations of two interactive subjects, i.e, $\mathbf{H}_m$ and $\mathbf{H}_n$, as input. The inputs of the SE in the following ITB are the outputs of the previous ITB. In the following subsections, we introduce the proposed SPM and GI-MSA in detail.

3.1 Semantic Partition Module

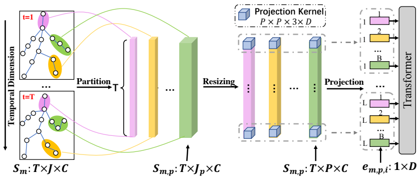


Fig. 3. The proposed Semantic Partition Module (SPM) performs three main operations (i.e., Partitioning, Resizing, and Projection) on the input skeleton sequence to generate its Body-Part-Time (BPT) sequence.

Different from natural 2D images that can be directly divided into a sequence of patches to feed to the Transformer [5], human skeleton sequences are represented as a set of 3D joints. Transforming the 3D skeleton sequences to 2D pseudo-images and passing them through a vision Transformer such as ViT [5] may result in losing the temporal dependency between frames as well as the correlation between joints. To better retain both spatial and temporal information of the skeleton sequences, we propose SPM to transform the skeleton sequence of each subject into a sequence of BPT. Each element in the BPT is the representation of one body part during a short temporal period. The overall architecture of the proposed SPM is shown in Fig. 3. There are three main steps in the SPM, i.e., partitioning, resizing, and projection, which are explained below.

Partitioning. Given the skeleton sequences of the interactive persons $\mathbf{S}_m, \mathbf{S}_n \in \mathbb{R}^{T \times J \times C}$, we first divide each skeleton sequence into B=5 body parts, i.e., *left arm, right arm, left leg, right leg and torso*, according to the natural structure of the human body. After the partitioning operation, each body part of each subject is represented as $\mathbf{S}_{m,p}, \mathbf{S}_{n,p} \in \mathbb{R}^{T \times J_p \times C}$, where $p \in B$ and J_p is the number of joints of body part p.

Resizing. Different body parts may have different numbers of joints. In order to adapt these body parts to the input of the Transformer, we adopt the linear interpolation to resize the spatial dimension J_p of all body parts to the same dimension P, i.e., $\mathbf{S}_{m,p}, \mathbf{S}_{n,p} \in \mathbb{R}^{T \times J_p \times C} \rightarrow \mathbf{S}_{m,p}, \mathbf{S}_{n,p} \in \mathbb{R}^{T \times P \times C}$, where $p \in B$. After the resizing operation, all B body parts have the same dimension.

Projection. The projection operation aims to transform the resized body parts of each person into a BPT sequence to feed to the Transformer. Specifically, we apply a 2D convolution with kernel size of $P \times P$ on $\mathbf{S}_{m,p}$ and $\mathbf{S}_{n,p}$ to generate 2D feature maps, respectively. The size of each output feature map is $L \times D$, where $L = \lceil (T + 2 \times \mathit{padding} - P + 1)/\mathit{stride} \rceil$ and D denotes the number of output channels. "*padding*" and "*stride*" denote the padding size and the stride of the convolutional filter. Each 2D feature map can then be split into a sequence of L steps, where each step is a feature vector of dimension D. The projection can be formulated as follows:

$$\begin{aligned}\mathbf{e}_{m,p,1}, \mathbf{e}_{m,p,2}, ..., \mathbf{e}_{m,p,L} &= \mathrm{Split}(\mathrm{Conv}(\mathbf{S}_{m,p})),\\ \mathbf{e}_{n,p,1}, \mathbf{e}_{n,p,2}, ..., \mathbf{e}_{n,p,L} &= \mathrm{Split}(\mathrm{Conv}(\mathbf{S}_{n,p})),\end{aligned} \tag{2}$$

where $\mathbf{e}_{m,p,j}, \mathbf{e}_{n,p,j} \in \mathbb{R}^D$ denote the embedding of the body part p at temporal step j for interactive person m and n, respectively. $j \in [1, \cdots, L]$, and D is the dimension of the embedding. L is the number of time steps of each body part. After projection, we concatenate the embedding of all the B body parts step by step for all the L time steps to generate a sequence with $M = B * L$ time steps. The sequence is referred to as the BPT sequence. As shown in Fig. 3, the BPT sequence can be considered as a combination of L sub-sequences, each of which is formed by the features of the B body part. We denote the BPT sequences generated from the skeleton sequences of the two interactive persons as $\mathbf{H}_m, \mathbf{H}_n \in \mathbb{R}^{M\times D}$. A learnable positional encoding [5] is added to $\mathbf{H}_m$ and $\mathbf{H}_n$ to form the inputs of two shared-weight Self Encoding (SE) modules, which are standard one-layer Transformers [5]. The output sequences of SE are denoted as $\mathbf{H}_{me}, \mathbf{H}_{ne} \in \mathbb{R}^{M\times D}$, which are then fed to the Graph Interaction Multi-head Self-Attention (GI-MSA) module to model the interactive body parts and generate an enhanced representation for each interactive subject.

3.2 Graph Interaction Multi-head Self-Attention

To accurately recognize human interaction, one critical cue is the interactive body parts. Considering the semantic correspondence and the distance characteristics that may exist in the interactive body parts, we propose a Graph Interaction Multi-head Self-attention (GI-MSA) module to model the interactive body parts as two interaction graphs as shown in Fig. 2 (b). Specifically, GI-MSA contains a Semantic-based Dense Interaction Graph (SDIG) and a Distance-based Sparse Interaction Graph (DSIG). The SDIG is learned by exploring the semantic correlations of the interactive body parts in a data-driven manner while the DSIG is constructed based on the prior knowledge that the physically close body parts of the interactive persons are generally interactive body parts and should be connected. With the SDIG and DSIG, the proposed GI-MSA models the interaction relationships of humans from both semantic and distance spaces to capture critical interactive information. Finally, the representation of each individual is enhanced by aggregating interactive features from the other person.

Semantic-based Dense Interaction Graph In order to capture the semantic correlations between interactive body parts of people (e.g., the hands holding the camera of one person and the hand with "yeah" of the other person in the action of "taking a photo"), we construct a Semantic-based Dense Interaction Graph (SDIG) for each interactive person. We take the learning of SDIG of person m (which is denoted as $\mathrm{SDIG}_{m\to n}$) as an example. As shown in Fig. 2 (b), given the representations of two interactive persons $\mathbf{H}_{me}, \mathbf{H}_{ne} \in \mathbb{R}^{M\times D}$, which are the outputs of the SE module, we first transform $\mathbf{H}_{me}$ into the latent space by a linear transformation function $\mathcal{T}^Q$:

$$\mathbf{H}_{me}^{Q} = \mathcal{T}^{Q}(\mathbf{H}_{me}) = \mathbf{H}_{me}\mathbf{W}^{Q}, \tag{3}$$

where $\mathbf{H}_{me}^{Q} \in \mathbb{R}^{M \times D}$ is the transformed query feature and $\mathbf{W}^{Q} \in \mathbb{R}^{D \times D}$ is the weight matrix. Then, we propose a context transformation function $\mathcal{C}$ to transform the representation of the other person $\mathbf{H}_{ne}$ into a high-level space as the key features,

$$\mathbf{H}_{ne}^{K} = \mathcal{C}(\mathbf{H}_{ne}) = (\mathbf{H}_{ne} + \mathbf{H}_{ne}^{tc} + \mathbf{H}_{ne}^{sc})\mathbf{W}^{K}, \tag{4}$$

where $\mathbf{H}_{ne}^{K} \in \mathbb{R}^{M \times D}$ is the key features. $\mathbf{W}^{K} \in \mathbb{R}^{D \times D}$ is the learned weight matrix. $\mathbf{H}_{ne}^{tc} \in$ and $\mathbf{H}_{ne}^{sc}$ are temporal and spatial contexts of $\mathbf{H}_{ne}$. To compute $\mathbf{H}_{ne}^{tc}$ and $\mathbf{H}_{ne}^{sc}$, we first compute $\mathbf{H}_{ne,p}^{tc}$ and $\mathbf{H}_{ne,t}^{sc}$ as follows, which denote the temporal context of each body part p and spatial context at time step t in $\mathbf{H}_{ne}$.

$$\begin{aligned} \mathbf{H}_{ne,p}^{tc} &= \frac{1}{L}\sum_{j=1}^{L}\mathbf{H}_{ne,p,j}, \\ \mathbf{H}_{ne,t}^{sc} &= \frac{1}{B}\sum_{i=1}^{B}\mathbf{H}_{ne,i,t}, \end{aligned} \tag{5}$$

where L denotes the temporal steps of each body part in $\mathbf{H}_{ne}$, and B is the number of body parts. $\mathbf{H}_{ne,p,j} \in \mathbb{R}^{D}$ and $\mathbf{H}_{ne,i,t} \in \mathbb{R}^{D}$ denote the feature encoding of body part p at time step j and the feature encoding of body part i at time step t in the sequence $\mathbf{H}_{ne}$, respectively. By stacking the temporal context of all B body parts and repeating L times, and repeating the spatial context of each time step B times and stacking the the repetition of all L time steps, respectively, we obtain $\mathbf{H}_{ne}^{tc}$, $\mathbf{H}_{ne}^{sc} \in \mathbb{R}^{M \times D}$. Finally, $\mathrm{SDIG}_{m \to n}$ can be obtained by performing the matrix multiplication operation between $\mathbf{H}_{me}^{Q}$ and $\mathbf{H}_{ne}^{K}$:

$$\mathrm{SDIG}_{m \to n} = \frac{\mathbf{H}_{me}^{Q}(\mathbf{H}_{ne}^{K})^{\top}}{\sqrt{D}}, \tag{6}$$

where $\top$ is the transpose operation, $\mathrm{SDIG}_{m \to n} \in \mathbb{R}^{M \times M}$. $\mathrm{SDIG}_{n \to m}$ can be obtained in a similar way, and the learnable weight matrices $\mathbf{W}^{Q}$ and $\mathbf{W}^{K}$ are shared for learning both $\mathrm{SDIG}_{m \to n}$ and $\mathrm{SDIG}_{n \to m}$.

Distance-based Sparse Interaction Graph In addition to modeling the interaction relationship from the semantic level, we also compute the distance correlation between body parts of the interactive persons. The DSIG is a predefined graph and could be constructed in the data pre-processing stage. The idea of DSIG is to leverage the distance between body parts to construct an adjacency matrix that contains the connection information between body parts of the interactive persons. More specifically, if the distance between two body parts of the interactive persons is small, then the two body parts are connected. Given the original skeleton sequences of two interactive humans $\mathbf{S}_m, \mathbf{S}_n \in \mathbb{R}^{T\times J\times C}$, we first divide the skeleton sequences into B body parts $\mathbf{S}_{m,p}, \mathbf{S}_{n,p} \in \mathbb{R}^{T\times J_p\times C}$ via the same Partitioning process in SPM. To estimate the distance between body parts, we first compute the representations of body parts by averaging the coordinates of joints within each body part:

$$\begin{aligned}\overline{\mathbf{S}}_{m,p} &= \frac{1}{J_p}\sum_{i=1}^{J_p}\mathbf{S}_{m,p}[i], i\in J_p, p\in B,\\ \overline{\mathbf{S}}_{n,p} &= \frac{1}{J_p}\sum_{j=1}^{J_p}\mathbf{S}_{n,p}[j], j\in J_p, p\in B,\end{aligned} \tag{7}$$

where $\overline{\mathbf{S}}_{m,p}, \overline{\mathbf{S}}_{n,p} \in \mathbb{R}^{T\times C}$ are the representations of body part p of two interactive persons respectively. $\mathbf{S}_{m,p}[i]$ and $\mathbf{S}_{n,p}[j]$ denote the i-th joint in $\mathbf{S}_{m,p}$ and the j-th joint in $\mathbf{S}_{n,p}$, respectively. J_p is the number of joints within body part p. We downsample the temporal dimension of $\overline{\mathbf{S}}_{m,p}$ and $\overline{\mathbf{S}}_{n,p}$ from T to L, i.e., $\overline{\mathbf{S}}_{m,p}, \overline{\mathbf{S}}_{n,p} \in \mathbb{R}^{T\times C} \rightarrow \overline{\mathbf{S}}_{m,p}, \overline{\mathbf{S}}_{n,p} \in \mathbb{R}^{L\times C}$. Then combining the representations of B body parts, we get the representations of two persons $\overline{\mathbf{S}}_m, \overline{\mathbf{S}}_n \in \mathbb{R}^{M\times C}$ in the distance space, where $M = L\times B$. $\overline{\mathbf{S}}_m, \overline{\mathbf{S}}_n$ can be treated as sequences with M time steps with dimension C. Each time step corresponds to a body part at a particular time step of the original sequence. For human m, we compute the Euclidean distance of each time step a in $\overline{\mathbf{S}}_m$ ($\overline{\mathbf{S}}_m[a]$) with each time step b in $\overline{\mathbf{S}}_n$ ($\overline{\mathbf{S}}_n[b]$):

$$\mathrm{A}_{m\rightarrow n}[a,b] = \sqrt{\sum_{c=1}^{C}(\overline{\mathbf{S}}_m[a] - \overline{\mathbf{S}}_n[b])^2}, \tag{8}$$

where $a, b \in [1, \cdots, M]$, and $\mathrm{A}_{m\rightarrow n} \in \mathbb{R}^{M\times M}$ records the distance between the body parts of two people. We finally connect each time step a in human m to the k nearest time step in human n to build the $\mathrm{DSIG}_{m\rightarrow n} \in \mathbb{R}^{M\times M}$ as below:

$$\mathrm{DSIG}_{m\rightarrow n}[a,b] = \begin{cases} 1, & \mathrm{A}_{m\rightarrow n}[a,b] <= \mathrm{A}^k_{m\rightarrow n}[a] \\ 0, & \mathrm{A}_{m\rightarrow n}[a,b] > \quad \mathrm{A}^k_{m\rightarrow n}[a] \end{cases} \tag{9}$$

where $\mathrm{A}^k_{m\rightarrow n}[a]$ is the k-th smallest value in a-th row of $\mathrm{A}_{m\rightarrow n}$. The $\mathrm{DSIG}_{n\rightarrow m} \in \mathbb{R}^{M\times M}$ is built in a similar way to encode the distance between each part of the interactive person n to all body parts of person m.

Interaction-based Feature Generation Given the semantic- and distance-based interaction graphs, we aggregate the interactive information of the graphs with the individual features of the interactive persons to generate an enhanced representation for better interaction recognition as shown in Fig. 2 (b). Specifically, we first transform the input individual representation $\mathbf{H}_{ne}$, which is the output of the SE module for person n, into the value features $\mathbf{H}_{ne}^{V}$:

$$\mathbf{H}_{ne}^{V} = \mathcal{T}^{V}(\mathbf{H}_{ne}) = \mathbf{H}_{ne}\mathbf{W}^{V}, \tag{10}$$

where $\mathbf{H}_{ne}^{V} \in \mathbb{R}^{M\times D}$, and $\mathbf{W}^{V}$ is the weight matrix. Then we perform the matrix multiplication operation on $\mathbf{H}_{ne}^{V}$ and the combination of $\mathrm{DSIG}_{m\to n}$ and $\mathrm{SDIG}_{m\to n}$, followed by an addition operation with $\mathbf{H}_{me}$ to obtain the interactive representation of person m:

$$\hat{\mathbf{H}}_{me} = \mathcal{R}(\mathrm{DSIG}_{m\to n}, \mathrm{SDIG}_{m\to n})\mathbf{H}_{ne}^{V} + \mathbf{H}_{me} \tag{11}$$

where $\hat{\mathbf{H}}_{me} \in \mathbb{R}^{M\times D}$, and $\mathcal{R}$ is the combination function:

$$\mathcal{R}(\mathrm{DSIG}, \mathrm{SDIG}) = \mathrm{Softmax}(\mathrm{DSIG}_{m\to n} + \alpha \cdot \mathrm{SDIG}_{m\to n}), \tag{12}$$

where α is a trainable scalar to adjust the intensity of each graph enabling the network to be adaptively adjustable between distance evolution and semantic correlation of body parts. Similarly, $\hat{\mathbf{H}}_{ne}$ can be obtained in the same way.

We define the above steps of generating $\hat{\mathbf{H}}_{me}$ and $\hat{\mathbf{H}}_{ne}$ from $\mathbf{H}_{me}$ and $\mathbf{H}_{ne}$ as Graph Interaction Self-Attention (GI-SA), which is formulated as:

$$\hat{\mathbf{H}}_{me}, \hat{\mathbf{H}}_{ne} = \textbf{GI-SA}(\mathbf{H}_{me}, \mathbf{H}_{ne}). \tag{13}$$

Finally, GI-MSA is defined by considering h attention "heads", i.e., h self-attention functions are applied to the input in parallel. Each head provides a sequence of size $M \times d$, where $d = D/h$. The outputs of the h self-attention functions are concatenated to form an $M \times D$ sequence to be fed to the a Layer Normalization (LN) followed by a FFN. The GI-MSA can be formulated as:

$$\begin{aligned} \text{GI-MSA}(\mathbf{H}_{me}, \mathbf{H}_{ne}) =& \mathrm{Concat}(\hat{\mathbf{H}}_{me,1}, ..., \hat{\mathbf{H}}_{me,h})\mathbf{W}^{m}, \\ & \mathrm{Concat}(\hat{\mathbf{H}}_{ne,1}, ..., \hat{\mathbf{H}}_{ne,h})\mathbf{W}^{n} \\ \hat{\mathbf{H}}_{me,i}, \hat{\mathbf{H}}_{ne,i} =& \textbf{GI-SA}(\mathbf{H}_{me,i}, \mathbf{H}_{ne,i}), \end{aligned} \tag{14}$$

where h is the number of heads, $\hat{\mathbf{H}}_{me,i}, \hat{\mathbf{H}}_{ne,i} \in \mathbb{R}^{M\times d}$ are output representations of i-th head of GI-SA, and $\mathbf{W}^{m}, \mathbf{W}^{n}$ are the weight matrices. $\mathbf{H}_{me,i}, \mathbf{H}_{ne,i} \in \mathbb{R}^{M\times d}$ are i-th head representations of $\mathbf{H}_{me}$ and $\mathbf{H}_{ne}$.

4 Experiments

The proposed IGFormer is evaluated on three benchmark datasets, i.e., SBU [34], NTU-RGB+D [28] and NTU-RGB+D120 [15], and is compared with state-of-the-art RNN-, CNN- and GCN-based human action and interaction recognition methods, including Co-LSTM [39], ST-LSTM [18], GCA-LSTM [21], 2s-GCA [20], FSNET [16], VA-LSTM [35], LSTM-IRN [24], ST-GCN [33], AS-GCN [12] and CTR-GCN [3]. Furthermore, to demonstrate the improvement of the proposed IGFormer over the standard Transformer model, we design a Transformer-based baseline named ViT-baseline, which is a ViT-base [5] model taking the pseudo-image representation of the skeleton sequence as input.

4.1 Datasets

SBU [34] is a two-person interaction dataset, which contains eight classes of human interactions including *approaching*, *departing*, *pushing*, *kicking*, *punching*, *exchanging objects*, *hugging*, and *shaking hands*. Seven participants (pairing up to 21 different permutations) performed all eight interactions. In total, the dataset contains 282 short videos. Each video contains 3D coordinates of 15 joints per person at each frame. Following [34] , we use the 5-fold cross validation protocol to evaluate our method.

NTU-RGB+D [28] is a large-scale action dataset containing 56,578 skeleton sequences from 60 action classes. Each action is captured by 3 cameras at the same height but from different horizontal angles. Each human skeleton contains 3D coordinates of 25 body joints. There are two standard evaluation protocols for this dataset including 1) Cross-Subject, where half of the subjects are used for training and the remaining ones are used for testing, and 2) Cross-View, where two cameras are used for training, and the third one is used for testing. This dataset contains 11 human interaction classes including *punch/slap*, *pat on the back*, *giving something*, *walking towards*, *kicking*, *point finger*, *touch pocket*, *walking apart*, *pushing*, *hugging* and *handshaking*. The maximum number of frames in each sample is 256.

NTU-RGB+D120 [15] extends NTU-RGB+D with an additional 57,367 samples from 60 extra action classes. In total, it contains 113,945 skeleton sequences from 120 action classes. There are two standard evaluation protocols for this dataset including 1) Cross-Subject, where half of the subjects are employed for training and the rest are left for testing, 2) Cross-Setup, where half of the setups are used for training, and the remaining ones are used for testing. In addition to the 11 interaction classes in the NTU-RGB+D, This dataset contains 15 additional human interaction classes including *hit with object*, *wield knife*, *knock over*, *grab stuff*, *shoot with gun*, *step on foot*, *high-five*, *cheers and drink*, *carry object*, *take a photo*, *follow*, *whisper*, *exchange things*, *support somebody* and *rock-paperscissors*, resulting a total of 26 interaction classes. In both NTU-RGB+D120 and NTU-RGB+D datasets, for samples with less than 256 frames, we repeat the sample until it reaches 256 frames.

4.2 Implementation Details

Transformer Architecture. We use a variant of ViT-Base [5] as the backbone of our proposed IGFormer model. The original ViT-base model contains 12 Transformer layers with the hidden size of 768 (D=768). The dimension of each MLP layer is four times the hidden size. However, due to the small number of samples in the human interaction recognition datasets, a lighter model is more suitable to avoid overfitting. Therefore, we reduce the number of Transformer layers to 3 (N=3) and initialize them with the pre-trained weights of the first three layers of the ViT-base model. We also remove the classification token (CLS) and adopt the average pooling operation to obtain the final representation from each sequence of patches. We set the patch size P in the Resizing step

Table 1. Performance comparison of different types of inputs and different lengths of the sequences on NTU-RGB+D and NTU-RGB+D 120. "PI" denotes "Pseudo-Image".

Input	**Length**	**NTU 60 (%)**		**NTU 120(%)**	
		X-Sub	**X-View**	**X-Sub**	**X-Set**
Sequence from PI	80	90.8	94.1	83.2	84.2
	125	91.8	95.2	83.7	85.0
	200	89.7	93.9	81.9	83.1
BPT Sequence	80	92.8	96.0	84.8	86.1
	125	**93.6**	**96.5**	**85.4**	**86.5**
	200	91.9	95.1	83.8	83.9

Table 2. Performance comparison of different interaction learning methods

Methods	**NTU 60 (%)**		**NTU 120(%)**	
	X-Sub	**X-View**	**X-Sub**	**X-Set**
Input Fusion	90.8	94.3	82.9	83.8
Late Fusion	91.2	94.8	83.0	84.1
IGFormer	**93.6**	**96.5**	**85.4**	**86.5**

of SPM to 16 and the stride of convolution in the Projection step to 10, which results in BPT sequences with M=125 for each person in all datasets. In each body part, L equals to 25. k in Eq. (9) is set to 15.

Training Details. The experiments are conducted on NVIDIA P100 GPU. We adopt SGD algorithm with Nesterov momentum of 0.9 as the optimizer. The initial learning rate is set to 0.01 and is divided by 10 at the 30^{th} and 40^{th} epochs. The training process is terminated at the 60^{th} epoch, batch size is 32.

4.3 Ablation Study

In this section, we conduct extensive ablation studies on both NTU RGB+D and NTU RGB+D 120 datasets to validate the effectiveness of the proposed SPM (Section 3.1) and GI-MSA (Section 3.2) modules.

Impacts of SPM. We compare two different representations of the skeleton sequences as the input of the proposed IGFormer to validate the effectiveness of the proposed SPM. The first one is **Pseudo-Image** representation, which have been widely used in CNN-based models [9,10,11] by transforming each 3D skeleton sequence to a 2D pseudo-image. We define the numbers of frames T and joints J of a skeleton sequence as the width and height of the image and then perform a linear projection on the image as ViT [5]. The second representation is the BPT sequence, which is generated by the proposed SPM. Moreover, skeletons are transformed into the different lengths by changing the stride of convolution projection in ViT and SPM to validate the robustness of the proposed SPM under different input configurations. The experimental results are shown in Table 1. We observe that the BPT representation outperforms Pseudo-Image representation at all three configurations, which validates the effectiveness of the proposed SPM. We also evaluate a baseline that models each skeleton joint as a token of the Transformer sequence and fuses features of two persons, but the performance drops by 2.2% compared with our SPM on X-Sub of NTU-RGB+D.

GI-MSA versus Input/Late fusion. We design two interaction learning baselines, i.e., **Input Fusion** and **Late Fusion**, to compare with our proposed GI-MSA module. The **Input Fusion** baseline merges the BPT sequences of two

Table 3. Performance comparison of different components of the proposed GI-MSA module. *sc* and *tc* represent spatial context and temporal context in Eq. (5).

Methods	**NTU 60 (%)**		**NTU 120(%)**	
	X-Sub	**X-View**	**X-Sub**	**X-Set**
baseline	90.2	93.3	82.1	83.6
DSIG	90.4	92.9	82.2	83.5
SDIG w/o sc	92.4	95.5	84.6	85.4
SDIG w/o tc	92.3	95.1	84.3	85.0
SDIG	92.8	95.7	84.8	85.5
SDIG + DSIG	**93.6**	**96.5**	**85.4**	**86.5**

subjects to form a single sequence and passes it through a standard Transformer to learn the interactions between two subjects. The **Late Fusion** baseline feeds the BPT sequences of two subjects individually through a Transformer model to extract their representations, which are then fused to model the interaction. As shown in Table 2, we observe that the performance of both input fusion and late fusion methods are worse than our proposed IGFormer on both datasets, demonstrating the efficacy of the proposed GI-MSA module for interactive learning.

Impacts of SDIG and DSIG. We evaluate the impacts of different components of the proposed GI-MSA, including SDIG, DSIG, the spatial and temporal context for learning SDIG. Here, we employ IGFormer without GI-MSA module as our baseline. Based on the results in Table 3, we draw three conclusions: (1) Both spatial and temporal context in Eq. (5) are important for learning key contextual features, i.e., the performance drops significantly by removing any of them. (2) The GI-MSA containing only SDIG can improve the performance of human interaction recognition, which validates the effectiveness of the proposed SDIG. (3) The DSIG, which serves as the prior knowledge of human interaction, does not perform well individually but provides extra information for interaction learning, leading to improved performance after being combined with SDIG.

Table 4. Performance comparison of number of ITB layers on NTU-RGB+D and NTU-RGB+D 120 datasets.

ITB	**NTU 60 (%)**		**NTU 120(%)**	
	X-Sub	**X-View**	**X-Sub**	**X-Set**
2	92.7	95.3	84.0	85.2
3	**93.6**	**96.5**	**85.4**	**86.5**
4	92.6	95.1	84.2	85.7
5	91.9	94.7	83.5	84.8

Impacts of Number of ITB layers. Our IGFormer is built by stacking several Interaction Transformer Blocks (ITBs) to enhance the capability of interaction modeling. Here, we evaluate the influence of different number of ITBs on the performance of IGFormer. As shown in Table 4, stacking 3 layers of ITB achieves the best results on both NTU-RGB+D and NTU-RGB+D 120. Increasing the number of ITBs degrades the accuracy due to over-fitting problem.

Impacts of the joint noise on human interaction. The skeletons in NTU-RGB+D are usually noisy, e.g., some joints are missing. We evaluate the performance of our IGFormer on X-Sub of NTU-RGB+D by adding zero-mean noise to the skeleton sequences. IGFormer achieves 93.6%, 93.1%, 92.0%, 90.4% accuracy when the standard deviation (σ) is set to 0cm, 1 cm, 2cm, 4cm, respectively, which demonstrates that IGFormer is robust against the input noise.

Table 5. Performance comparison on SBU, NTU-RGB+D and NTU-RGB+D 120.

Methods	**SBU(%)**	**NTU-RGB+D**		**NTU-RGB+D 120**	
		X-Sub (%)	**X-View (%)**	**X-Sub (%)**	**X-Set (%)**
Co-LSTM [39]	90.4	-	-	-	-
ST-LSTM [18]	93.3	83.0	87.3	63.0	66.6
GCA [22]	-	85.9	89.0	70.6	73.7
2s-GCA [19]	94.9	87.2	89.9	73.0	73.3
VA-LSTM [35]	97.2	-	-	-	-
FSNET [17]	-	74.0	80.5	61.2	69.7
LSTM-IRN [24]	98.2	90.5	93.5	77.7	79.6
ST-GCN [33]	-	83.3	87.1	78.9	76.1
AS-GCN [12]	-	89.3	93.0	82.9	83.7
CTR-GCN [3]	-	91.6	94.3	83.2	84.4
ViT-baseline	93.1	89.7	92.5	81.5	82.5
IGFormer	**98.4**	**93.6**	**96.5**	**85.4**	**86.5**

4.4 Comparison with State-of-the-arts

The experimental results on the interaction classes of SBU, NTU-RGB+D and NTU-RGB+D 120 datasets are shown in Table 5. The proposed IGFormer achieves state-of-the-art performance compared with other skeleton-based human interaction recognition methods. Benefiting from the proposed SPM and GI-MSA modules, IGFormer outperforms the CNN- and RNN-based methods by a large margin. IGFormer also outperforms state-of-the-art GCN-based method, CTR-GCN [3], by 2.0% and 2.2% on X-Sub and X-View of NTU-RGB+D, and 2.2% and 2.1% on X-Sub and X-set of NTU-RGB+D 120. Compared with the baseline Transformer-based method, ViT-baseline, our IGFormer achieves 3.4% and 3.2% gains on X-Sub and X-View of NTU-RGB+D, and 3.9% and 4.0% gains on X-Sub and X-Set of NTU-RGB+D 120.

5 Conclusion

In this work, we presented IGFomer, which consists of a GI-MSA module to model the interaction of persons as graphs. The GI-MSA learns an SDIG and DSIG to capture the semantic and distance correlations between body parts of interactive persons. We also presented a SPM to transform each human skeleton into a BPT sequence for retaining interactive information of body parts. The proposed IGFormer outperformed state-of-the-art methods on three datasets.

6 Acknowledgement

The research is partially supported by University of Melbourne Early Career Researcher Grant (No: 2022ECR008). This research is also partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215. This work is also partially supported by National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-100E-2020-065), SUTD Startup Research Grant and MOE Tier 1 Grant.

References

1. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. ACM Computing Surveys **43**(3), 16:1–16:43 (2011). https://doi.org/10.1145/1922649.1922653
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer (2021)
3. 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 the IEEE/CVF International Conference on Computer Vision. pp. 13359–13368 (2021)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (Jun 2019). https://doi.org/10.18653/v1/N19-1423, `https://aclanthology.org/N19-1423`
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Du, Y., Fu, Y., Wang, L.: Skeleton based action recognition with convolutional neural network. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). pp. 579–583. IEEE (2015)
7. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1110–1118 (2015)
8. Ji, Y., Ye, G., Cheng, H.: Interactive body part contrast mining for human interaction recognition. In: 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). pp. 1–6 (2014). https://doi.org/10.1109/ICMEW.2014.6890714
9. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3d action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3288–3297 (2017)
10. Kim, T.S., Reiter, A.: Interpretable 3d human action analysis with temporal convolutional networks. In: 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW). pp. 1623–1631. IEEE (2017)
11. Li, B., Dai, Y., Cheng, X., Chen, H., Lin, Y., He, M.: Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. In: 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 601–604. IEEE (2017)
12. 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 the IEEE/CVF conference on computer vision and pattern recognition. pp. 3595–3603 (2019)
13. Li, S., Li, W., Cook, C., Zhu, C., Gao, Y.: Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5457–5466 (2018)
14. Liu, H., Tu, J., Liu, M.: Two-stream 3d convolutional neural network for skeleton-based action recognition. arXiv preprint arXiv:1705.08106 (2017)

15. 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 transactions on pattern analysis and machine intelligence **42**(10), 2684–2701 (2019)
16. Liu, J., Shahroudy, A., Wang, G., Duan, L.Y., Kot, A.C.: Skeleton-based online action prediction using scale selection network (2019)
17. Liu, J., Shahroudy, A., Wang, G., Duan, L.Y., Kot, A.C.: Skeleton-based online action prediction using scale selection network. IEEE transactions on pattern analysis and machine intelligence **42**(6), 1453–1467 (2019)
18. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal lstm with trust gates for 3d human action recognition. In: European conference on computer vision. pp. 816–833. Springer (2016)
19. Liu, J., Wang, G., Duan, L.Y., Abdiyeva, K., Kot, A.C.: Skeleton-based human action recognition with global context-aware attention lstm networks. IEEE Transactions on Image Processing **27**(4), 1586–1599 (2017)
20. Liu, J., Wang, G., Duan, L.Y., Abdiyeva, K., Kot, A.C.: Skeleton-based human action recognition with global context-aware attention lstm networks. IEEE Transactions on Image Processing **27**(4), 1586–1599 (Apr 2018). https://doi.org/10.1109/tip.2017.2785279, `http://dx.doi.org/10.1109/TIP.2017.2785279`
21. Liu, J., Wang, G., Hu, P., Duan, L.Y., Kot, A.C.: Global context-aware attention lstm networks for 3d action recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3671–3680 (2017). https://doi.org/10.1109/CVPR.2017.391
22. 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 the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1647–1656 (2017)
23. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 143–152 (2020)
24. Perez, M., Liu, J., Kot, A.C.: Interaction relational network for mutual action recognition. IEEE Transactions on Multimedia (2021)
25. Plizzari, C., Cannici, M., Matteucci, M.: Skeleton-based action recognition via spatial and temporal transformer networks. Computer Vision and Image Understanding **208-209**, 103219 (2021). https://doi.org/https://doi.org/10.1016/j.cviu.2021.103219, `https://www.sciencedirect.com/science/article/pii/S1077314221000631`
26. Poppe, R.: A survey on vision-based human action recognition. Image and vision computing **28**(6), 976–990 (Jun 2010). https://doi.org/10.1016/j.imavis.2009.11.014, 10.1016/j.imavis.2009.11.014
27. Raptis, M., Sigal, L.: Poselet key-framing: A model for human activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2650–2657 (2013)
28. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: 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)
29. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12026–12035 (2019)

30. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Proceedings of the AAAI conference on artificial intelligence. vol. 31 (2017)
31. Vahdat, A., Gao, B., Ranjbar, M., Mori, G.: A discriminative key pose sequence model for recognizing human interactions. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). pp. 1729–1736. IEEE (2011)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
33. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence (2018)
34. Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., Samaras, D.: Two-person interaction detection using body-pose features and multiple instance learning. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 28–35 (2012). https://doi.org/10.1109/CVPRW.2012.6239234
35. 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 the IEEE International Conference on Computer Vision. pp. 2117–2126 (2017)
36. Zhang, Y., Liu, X., Chang, M.C., Ge, W., Chen, T.: Spatio-temporal phrases for activity recognition. In: European Conference on Computer Vision. pp. 707–721. Springer (2012)
37. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., Zhang, L.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR (2021)
38. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6881–6890 (2021)
39. Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., Xie, X.: Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 30 (2016)
40. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)
41. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable {detr}: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=gZ9hCDWe6ke