

S3GAAR: Segmented Spatiotemporal Skeleton Graph-Attention for Action Recognition

Musrea Ghaseb^{✉*}, Ahmed Elhayek^{✉†}, Fawaz Alsolami^{✉*}, and Abdullah Marish Ali^{✉*}

^{*}King Abdulaziz University, Faculty of Computing and Information Technology, Jeddah, Saudi Arabia

[†] University of Prince Mugrin, Faculty of Computer and Cyber Sciences, Madinah, Saudi Arabia.

[‡]Corresponding Author: Musrea Ghaseb (e-mail: gmusreaghaseb@stu.kau.edu.sa).

Abstract—Human motion recognition is extremely important for many practical applications in several disciplines, such as surveillance, medicine, sports, gait analysis, and computer graphics. Graph convolutional networks (GCNs) enhance the accuracy and performance of skeleton-based action recognition. However, this approach has difficulties in modeling long-term temporal dependencies. In Addition, the fixed topology of the skeleton graph is not sufficiently robust to extract features for skeleton motions. Although transformers that rely entirely on self-attention have demonstrated great success in modeling global correlations between inputs and outputs, they ignore the local correlations between joints. In this study, we propose a novel segmented spatiotemporal skeleton graph-attention network (S3GAAR) to effectively learn different human actions and concentrate on the most operative part of the human body for each action. The proposed S3GAAR models spatial-temporal features through spatiotemporal attention for each segment to capture short-term temporal dependencies. Owing to several human actions that focus on one or more body parts such as mutual actions, our novel method divides the human skeleton into three segments: superior, inferior, and extremity joints. Our proposed method is designed to extract the features of each segment individually because human actions focus on one or more segments. Moreover, our segmented spatiotemporal graph introduces additional edges between important distant joints in the same segment. The experimental results show that our novel method outperforms state-of-the-art methods up to 1.1% on two large-scale benchmark datasets, NTU-RGB+D 60 and NTU-RGB+D 120.

Index Terms—Self-attention, Transformers, Graph Attention, Human motion Recognition, Skeleton-based Action, Graph Convolutional Networks.

I. INTRODUCTION

HUMAN action recognition (HAR) involves the identification and classification of various actions or interactions that humans perform in a given situation. It involves analyzing the movements and poses of human bodies as well as the context in which the actions are taking place. Recently, several applications of human action recognition (HAR) have been developed for many disciplines, such as surveillance, medicine, sports, gait analysis, and computer graphics [1], [2]. Several RGB-based methods such as [3] and [4] are used for human action recognition, however, they are significantly affected by environmental factors such as brightness, intensity, ambiguity, and extreme self-occlusions. To avoid environmental noise, skeleton-based methods [5]–[10] use skeleton representations by extracting the main human

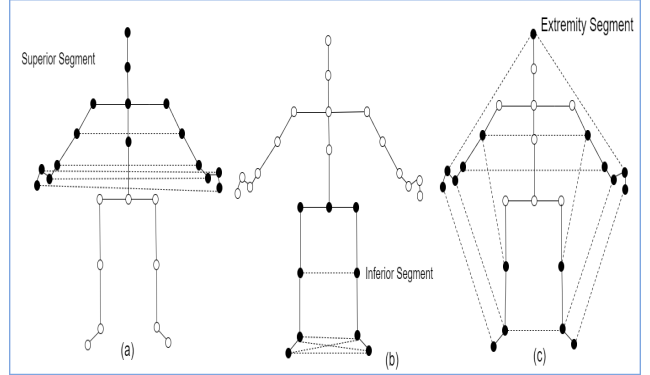


Fig. 1. Solid Edges indicate physical edges between joints and the dashed edges are additional edges to connect distant joints. The white joints represent the excluded joints in each segment. (a) The superior segment represents the top area of the human skeleton graph with additional edges connecting the left arm with the right one. (b) The inferior segment represents the bottom area of the body. (c) The extremity segment represents limb joints.

joint positions as 2D or 3D coordinates instead of using images or videos in the process of human action recognition.

Recent approaches such as Graph Convolutional Networks (GCN) [1], [10]–[12] and semantics-guided neural networks [13] use human-skeleton graphs into network convolutional layers. ST-GCN [11] proposed spatiotemporal graph convolutional networks for modeling both spatial and temporal features on non-Euclidean data; however, they have difficulties in modeling long-term temporal dependencies [14], [15]. Moreover, the fixed topology of the skeleton graph is not sufficiently robust to extract the features of skeleton motions [16].

Furthermore, transformer-based approaches have demonstrated great success across a wide range of tasks that rely entirely on self-attention to draw global correlations between inputs and outputs, which are effective solutions to problems of long-sequence data [5]–[9], [17]. However, self-attention ignores the local correlations among joints, which limits the improvement in recognition accuracy.

Most existing methods, such as DSTA-Net [7] use independent modules to extract spatial and temporal features. However, STGAT [6] recently proposed a spatiotemporal graph attention network for modeling short-term temporal movements by building a spatiotemporal graph from joints in neighboring frames.

An analysis of the movements and poses of the human body

shows that each body part has a different influence on human actions. For example, eating, drinking, or making phone calls depends only on the upper part of the body. However, the step of a foot action depends on the foot motion. In addition, interaction actions between two persons are performed by one part of the human body, such as shaking hands or kicking. For this purpose, we propose a segmented spatiotemporal graph attention with additional meaningful edges to detect the relationships between joint nodes in the same spatiotemporal segment. Our novel method divides the human skeleton into three segments: the superior, inferior, and extremity joints. Inspired by STGAT method [6], in this study, we consider the use of local spatiotemporal modeling for short-term dependencies in each segment instead of modeling spatial and temporal dependencies individually, which can further improve the performance of skeleton-based action recognition tasks.

The proposed method incorporates a self-attention mechanism into our segmented spatiotemporal graph that detects the relationships between meaningful adjacent and distant nodes in the same segment (Fig. 1). For instance, in the superior segment, we add edges between each node in the right hand and its counterpart in the left hand. The added edges provide strong relationships between the right and left hands for several human actions that are often performed by both hands such as typing on a keyboard or taking off glasses.

To concentrate on the most operative part of the human body for each action, our method learns the correlations between nodes in the same segment instead of learning correlations between nodes in the whole topology. However, besides the three segments, we add extra attention head for the whole body to capture actions that consider the entire body, such as swimming. To this end, we use multi-head attention to provide an attention aggregation for various heads. The code of the proposed method will be published publicly at <https://github.com/musrea/S3GAAR>

Our main contributions are summarized as follows:

- We propose a segmented spatiotemporal graph attention network to effectively learn different human actions and concentrate on the most operative segments of the human body.
- We propose a skeleton graph segmentation consisting of three segments: superior, inferior, and extremity joints.
- We introduce extra edges between important distant joints for each segment.
- A segmented cross-spacetime attention to capture short-term temporal dependencies.
- The experimental results demonstrate the effectiveness of our novel method and show that it outperforms state-of-the-art methods up to 1.1% on two large-scale benchmark datasets, NTU-RGB+D 60 and NTU-RGB+D 120.

The remainder of this paper is organized as follows. In Section II, related works on human action recognition using GCNs, graph attention, and self-attention are discussed. Section III demonstrates the proposed method. Section IV explains the proposed segmented spatiotemporal skeleton graph attention(S3GAAR) model in detail. Section V discusses the configurations and evaluation process of our model, including

datasets, training configurations, and experimental results. Finally, Section VI presents conclusions and suggestions for future research.

II. RELATED WORKS

In the last decade, some approaches, such as recurrent neural networks (RNN) [18] and long short-term memory (LSTM) networks [19], have been fundamental methods for skeleton-based action recognition that work effectively for sequential data tasks. However, they do not work as parallel methods and have difficulty in learning long-term patterns.

A. Action Recognition with GCNs

Graph Convolutional Networks (GCN) [1], [9]–[12], [20], [21], and semantics-guided neural networks [13] represent human skeleton data using a graph with joint nodes and focus on topology modeling. Yan et al. [12] proposed a spatiotemporal graph convolutional network (ST-GCN) for modeling both spatial and temporal features on non-Euclidean data. However, many GCN-based methods have difficulties with modeling long-term temporal dependencies [14]. Furthermore, the fixed topology of the skeleton graph is not sufficiently robust for extracting the features of skeleton motions.

B. Transformers for Action Recognition

Recently, the transformer-based approach has become the most common approach for Natural Language Processing (NLP). Its architecture has been extended to cover various aspects such as computer vision. The self-attention mechanism is the main block of transformer-based methods [5]–[8], [13], [17], [22]–[29]. Compared to RNN/CNN-based approaches, transformer-based approaches can effectively handle long-term temporal dependencies. However, due to the large attention maps, they often require high computational resources and longer training time.

Vision Transformers (ViT). In computer vision, transformer-based approaches have achieved remarkable success in several fields such as image classification, Image fusion [17], and object detection [30], [31]. However, image-related tasks are significantly affected by environmental factors such as brightness, intensity, ambiguity, and extreme self-occlusions. In the field of human action recognition and to avoid environmental noise and large-sized images, some studies use skeletal data instead of images or videos in the process of human action recognition [5]–[10].

The self-attention mechanism is used in skeleton-based action recognition to model complex intra-frame and inter-frame correlations. To capture the skeleton motion, spatial self-attention is used to model the relationships between different joints in the same frame, and temporal self-attention is used to model the relationships between the joint and its counterpart in adjacent frames.

Some models, such as [21], proposed a multi-scale temporal transformer (MTT) to address the problem of limited access to long-term temporal information. Pose Transformers (POTR) [22] approach consists of a transformer encoder that

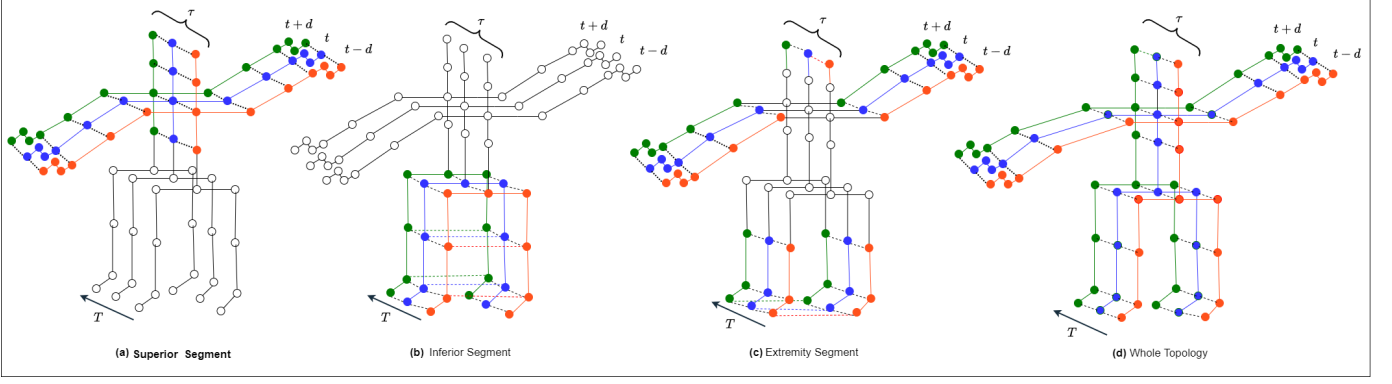


Fig. 2. An illustration for segmented spatiotemporal graphs which consist of τ frames at each timestamp t with the dilation d . (a) A spatiotemporal graph represents the superior segment of the skeleton. (b) A spatiotemporal graph for the inferior segment which represents the bottom area of the skeleton. (c) A spatiotemporal graph for the extremity segment which represents limb joints. (d) A spatiotemporal graph represents the standard topology of the human skeleton. Note that for simplicity, we exclude the additional edges from the graphs.

encodes the input sequence, and the decoder works in a non-autoregressive manner, generating the predictions of all poses in parallel. Shi et. al [24] propose a sparse transformer to reduce the computational complexity and memory usage. They proposed sparse attention for modeling spatial features and segmented linear attention for modeling temporal features.

Liu and Zhou [29] proposed a fully attentional network (FAN) consists of two modules, the spatial attention module is used to model the relationships between different joints in the same frame, and the temporal attention module is used to model the relationships between the joint and its counterpart in adjacent frames. However, DSTA-NET [7] proposed a decoupled spatial-temporal attention network that captures motion correlation by employing a spatial attention-based module for spatial modeling and a temporal attention-based module for temporal modeling. Furthermore, STTFormer [32] divided the skeleton sequence into several parts(tuples) to encode each tuple and passed a spatiotemporal tuple into a self-attention module to study the relationship between joints in each tuple.

C. Graph Attention for Action Recognition

KAAGTN [14] proposed a two-stream framework, an adaptive graph transformer network to learn spatial dependencies between joints from higher-order degrees, and a temporal attention block to model temporal features. However, STGAT [6] proposed a spatiotemporal modeling method using graph attention to capture short-term dependencies by building a local spatiotemporal graph containing nodes in a local cross-spacetime neighborhood. In contrast to previous methods, our method uses a self-attention mechanism with a segmented skeleton graph to study the most operative part of the human body for each action. We consider the use of local spatiotemporal modeling for short-term dependencies in each segment instead of modeling spatial and temporal dependencies individually, which can further improve the performance of skeleton-based action recognition tasks. Furthermore, we propose additional edges between important distant joints in the same segment to allow our model to effectively capture different human actions.

III. METHOD OVERVIEW

In this section, we define the method notations and formulate the segmented graph attention and multi-head attention.

Notations. The skeleton graph is denoted as $G = (V, E)$, where $V = \{v_1, \dots, v_N\}$ denotes a set of graph nodes(body joints), and E represents the edges(bones) connecting body joints. The edges are depicted using an adjacency matrix A of size $N \times N$, where $A_{i,j}$ indicates the relationship between node i and j , $\forall i, j \in N$. If there is no connection between node i and j , $A_{i,j}$ is set to zero. The input features come with tensor shape $X^{C \times T \times N}$ where each node v_n in the set of nodes N has a feature vector of dimension C across T frames(number of frames in a skeleton sequence).

We decompose the human skeleton graph into three segments: the superior, inferior, and extremity. Each segmented graph $G_s = (V_s, E_s)$ is a subset of G that focuses on a specific area in the human body. For each segmented graph, the adjacency matrix A_s denotes the relationships between nodes in a segment s as shown in Equation 1. For example, the upper-segmented graph consists of joints located in the superior segment of the body.

$$G_s \in G \text{ and } A_s \in A \quad (1)$$

where A is an adjacency matrix of the whole-skeleton graph, and A_s is an adjacency matrix for one of the segmented graphs G_s .

To enhance the capture of short-term dependencies, we build a spatiotemporal graph for each segmented graph which consists of τ frames at each timestamp t as shown in Fig. 2. The dilation d controls the interval between τ frames at each time step t (see Equations 2 and 3).

$$X_{\tau_s}^t = \{x_{t-d(\tau/2):t+d(\tau/2)} \in \mathbb{R}^{C \times \tau \times N} | t \in \mathbb{Z}, 0 \leq t \leq T\} \quad (2)$$

where τ is the number of input frames that are included in the local action sequence at a timestamp t , d is the dilation rate and $X_{\tau_s}^t$ is the local action sequence at a timestamp t . Each

frame x_t has an adjacent matrix $A_{\tau_s}^t$ which represents local spatiotemporal relationships for $X_{\tau_s}^t$ as following equation.

$$A_{\tau_s}^t = [A_{(1)_s}^t, \dots, A_{(\tau)_s}^t] \quad (3)$$

where $A_{\tau_s}^t$ denotes a segmented spatiotemporal graph at timestamp t , that models the relationships between nodes in the segment s and their local spatiotemporal neighborhood τ frames. For example, a node at the current timestamp t (corresponding to one column of $A_{\tau_s}^t$) is connected to $\tau \times N$ neighbors, comprising τ frames with N nodes at each timestamp.

In self-attention, each node v_i has three vectors, a query \mathbf{q}_i , a key \mathbf{k}_i and a value vector \mathbf{v}_i where $i=1, \dots, n$ and n is the total number of nodes. To calculate the correlation between nodes, the attention score for each node v_i is obtained by performing the scaled-dot product as shown in Equation (4). The scaling by d_k improves the stability of gradients during training.

$$a_{ij} = \text{softmax}\left(\frac{\mathbf{q}_i \cdot \mathbf{k}_j^\top}{\sqrt{d_k}}\right) \quad \forall \quad 1 \leq i, j \leq n \quad (4)$$

where \mathbf{q}_i is the query of node v_i , and \mathbf{k}_j^\top is the transposed key for node v_j . d_k is the channel dimensions of key vectors \mathbf{k} . The scaled-dot product attention is normalized and transformed using the Softmax function to ensure that the attention weights for each node v_i are scaled to a probability distribution (see Fig. 3).

The sum of attention weights for node v_i to all other nodes are obtained using $\text{Attention}(\mathbf{q}_i, \mathbf{k}_j, \mathbf{v}_j)$ for every $j = 1, 2, \dots, n$.

$$\mathbf{w}_i = \text{Attention}(\mathbf{q}_i, \mathbf{k}_j, \mathbf{v}_j) = \sum_j a_{ij} \cdot \mathbf{v}_j \quad (5)$$

where \mathbf{w}_i is the weighted sum for node v_i and a_{ij} is the attention score for node v_i which indicates how much node v_j attends to node v_i . \mathbf{v}_j represents the value vector of node v_j .

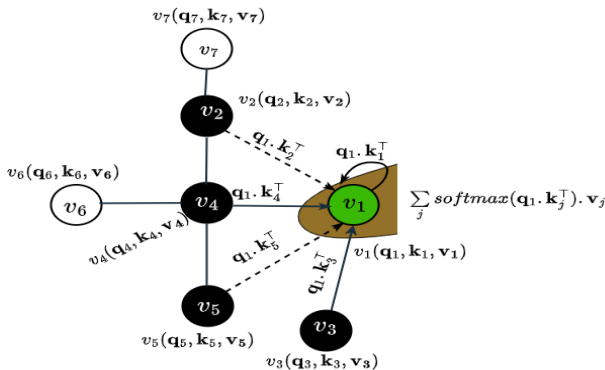


Fig. 3. The attention scores for node v_1 is obtained from the dot product between queries(\mathbf{q}_1) and keys(\mathbf{k}_j). Solid edges indicate physical edges and the dashed edges are additional edges to connect distant joints in the same segment.

In general, to obtain an attention head weight, the input sequence is transformed into three feature matrices, query(\mathbf{Q}), key(\mathbf{K}), and value(\mathbf{V}) using three linear projection layers.

Attention scores are obtained from the dot product between queries and keys and are formulated as :

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \quad (6)$$

where \mathbf{Q}, \mathbf{V} , and \mathbf{K}^\top represent the feature vectors of queries, values, and the transposed keys, respectively, and d_k denotes the channel dimensions of key vectors.

To compute the segmented attention weight \mathbf{W}_s , the input sequence for each segment s is transformed into three feature vectors, query(\mathbf{Q}_s), key(\mathbf{K}_s), and value(\mathbf{V}_s) using three linear projection layers. Then multiply query(\mathbf{Q}_s) with the transposed key(\mathbf{K}_s) as:

$$\mathbf{W}_s = \text{softmax}\left(\frac{\mathbf{Q}_s\mathbf{K}_s^\top}{\sqrt{d_k}}\right)\mathbf{V}_s \quad (7)$$

where \mathbf{Q}_s , \mathbf{V}_s , and \mathbf{K}_s^\top represent the feature matrices of queries, values, and the transposed keys for the segment s , respectively.

From Equation (8) the segmented spatiotemporal graph attention score for one head within the segment s is obtained as follows:

$$X_s^{\text{out}} = \sigma(\mathbf{W}_s X_{\tau_s}^{\text{in}} A_{\tau_s}) \quad (8)$$

where σ denotes the non-linear activation function, \mathbf{W}_s is the segmented weight matrix, $X_{\tau_s}^{\text{in}}$ is the segmented local input sequence and A_{τ_s} is the adjacency matrix of the segmented spatiotemporal graph as shown in Fig. 4 . We use a multi-head

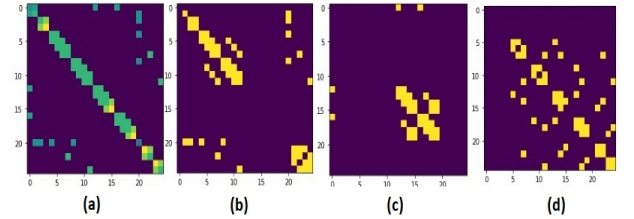


Fig. 4. (a)Adjacency matrix for whole-skeleton graph. (b),(c), and (d) represent adjacency matrices for the superior, inferior, and extremity graphs, respectively.

self-attention mechanism to learn various edge weights for all segmented spatiotemporal graphs. Therefore, for each segment s there are N_h heads. To obtain the final output, all heads for all segments are concatenated, which is formulated as.

$$X^{\text{out}} = \text{Concat}(\text{head}_1, \dots, \text{head}_{N_h})^{\times S} \quad (9)$$

where $\text{head}_h = \sigma(\mathbf{W}_s X_{\tau_s}^{\text{in}} A_{\tau_s})$ is a head attention for the segment s and X^{out} is the concatenated output of all heads with various segmented spatiotemporal graphs.

IV. METHODOLOGY

In this section, we elaborate on segmented spatiotemporal Graph Attention(S3GAAR) and analyze the importance of Segmented Graph Attention(SGAT). We also detail the Temporal Convolutional layer to extract global dependencies. Finally, we introduce the overall structure of the S3GAAR (see Fig. 5).

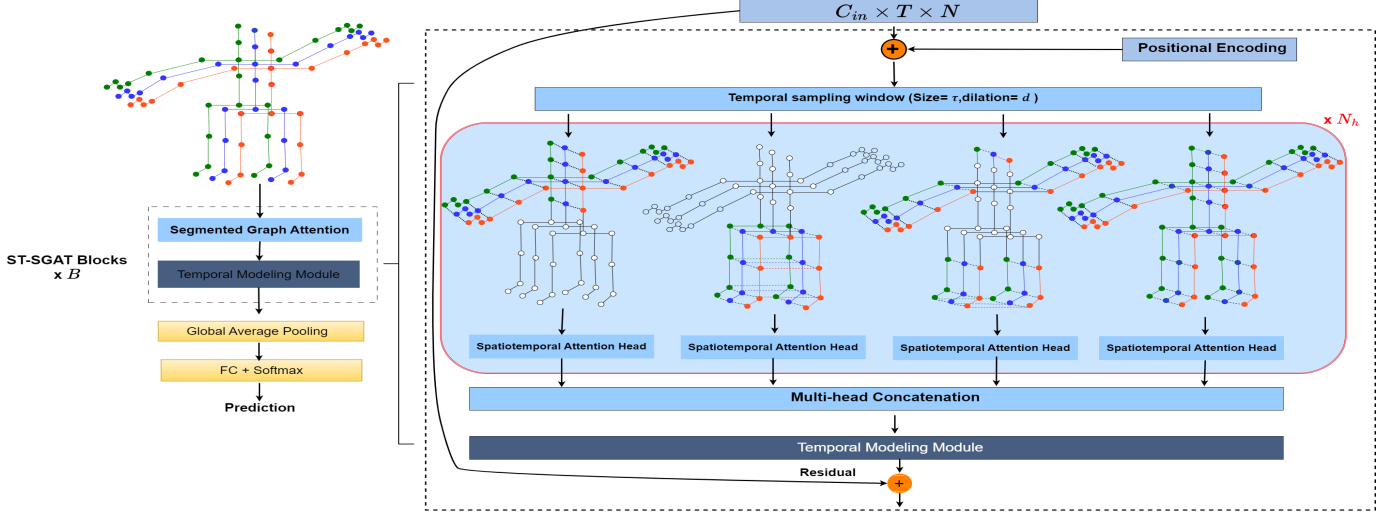


Fig. 5. Architecture Overview of S3GAAR: Our Proposed method consists of a Segmented Graph Attention block (SGAT) that extracts local spatiotemporal features. There are N_h heads for each segment concatenated together to obtain the final attention weight. The temporal convolution layer captures long-term temporal dependencies. Our network consists of B S3GAAR blocks with additional layers, global average pooling, fully connected layer, and Softmax function for the classification task.

A. Segmentation Methodology for Human Skeleton Graph

Our proposed segmentation methodology stands on classifying human actions based on the most operative part of the human body into four types: 1- Actions performed by the superior part of the body. 2- Actions performed by the inferior part of the body. 3- Actions Performed by extremities. 4- Actions performed by more than one part. Therefore, the first step is to divide the human skeleton graph into three segments: superior, inferior, and extremity segments because several actions depend on one of these segments. For example, eating, drinking, using a phone, typing on a keyboard, or shaking hands depends on the upper part of the body. In Addition, to perform actions such as kicking an object, the inferior part of the skeleton is more important than the other parts. In addition, actions such as putting on a shoe or taking it off depend more on the extremity joints than on the other joints.

Defining Additional Edges. In several actions, there are important correlations between joint nodes that are not physically connected. To model the correlations between distant joint nodes in the same segment, we define meaningful edges in each segment graph. For instance, in the superior segment, we add edges between each node in the right hand and its counterpart in the left hand. The added edges provide strong relationships between the right and left hands for several human actions that are often performed by both hands such as typing on a keyboard or taking off glasses. Furthermore, additional edges between the right and left foot nodes improve the recognition process for several human actions such as walking or standing up by capturing motion patterns of the lower part of the human body. In addition, several human actions are performed by extremities such as taking off a shoe. So, we also add extra edges between the upper and lower limbs as shown in Fig. 1 to enhance the recognition process for these types of human actions.

Although our segmented graphs have additional edges compared to conventional graphs, they notice it is difficult to capture the relationships between nodes in different segments. To overcome this limitation, we used a conventional graph as the fourth segment to extract the local features from the whole-skeleton graph.

B. Segmented spatiotemporal Graph Attention Network

Building Segmented Spatiotemporal Graphs. In this study, we provide spatial-only modules that are more capable of capturing local correlations, whereas temporal-only modules are responsible for global correlation modeling. Inspired by STGAT [6], we expanded spatial-only modules to build a local spatiotemporal graph to aggregate features between local spatiotemporal neighborhoods. We build a spatiotemporal graph for each segment consisting of nodes in a segmented spatial graph in a particular timestamp including nodes from its cross-spacetime neighborhood (see Equation. 3). The spatiotemporal graph consists of edges with their counterparts in neighboring timestamps beside their edges in the original segmented spatial graph. In this method, each node in the proposed segment is directly linked to other local spatiotemporal neighbors (see Fig. 2).

Implementing Segmented spatiotemporal Graph Attention.

Using Equation (8), we obtain the graph attention score using the segmented spatiotemporal $A_{s_\tau}^t$ for single head attention. We use a multi-head self-attention mechanism to learn various patterns for each segmented spatiotemporal graph. To obtain the final output, we concatenate different attention heads, as shown in Equation (9).

C. Temporal Modeling Module

The Temporal Modeling Module is a temporal convolutional layer that follows the segmented spatiotemporal module to

capture long-term temporal dependencies.

D. Architecture Overview of S3GAAR

In contrast to DSTA-Net [7], which proposes decoupled modules, a spatial attention module models spatial features, and a temporal attention module models long-term temporal dependencies. Our Proposed network consists of B segmented spatiotemporal graph attention blocks to effectively extract local spatiotemporal features, and a temporal module to model temporal features. The input data passed B sequences of the S3GAAR blocks including the temporal modeling module.

Our spatiotemporal module consists of N_h attention heads for each segment to improve the model's ability to learn various patterns. In addition, attention scores for different heads are concatenated to obtain the final scores. For classification tasks, we added the following layers: global average pooling, fully connected, and Softmax functions.

V. EXPERIMENTS AND RESULTS

A. Datasets

NTU-RGB+D 60. NTU-RGB+D 60 [33] is a skeleton-based dataset for human action recognition tasks. Skeleton keypoint data were collected from 56,880 video samples for 60 different action classes. All the Action samples are carried out by 40 different people. One or more people were involved in each video and was recorded using three cameras from Microsoft Kinect. The authors recommend two methods to split the data into training and testing sets. For the subjects, the Cross-Subject (X-Sub) benchmark uses half of the subjects for training, and the other half for testing. Concerning views, the Cross-View (X-View) benchmark uses three cameras with angles (45° , 0° , -45°), two of them are used for training, and the third camera is used for testing.

NTU-RGB+D 120. NTU-RGB+D 120 [34] is also a skeleton-based large-scale dataset for human action recognition tasks which is an extension of NTU-RGB+D 60 dataset. The skeleton keypoint data is collected from 114,480 video samples for 120 different action classes. All samples are carried out by 106 different persons in different views and backgrounds with 32 setups. Each video sample consisted of one or more people and was recorded using three cameras. In a protocol similar to the NTU-RGB+D 60 dataset, the authors recommend two methods to split the data into training and testing sets into two benchmarks: Cross-Subject (X-Sub) and Cross-Setup (X-setup) benchmarks. For the Cross-Subject (X-Sub) benchmark, half of the subjects are used for training, and the other half are used for testing. For setups, the Cross-Setup (X-Setup) benchmark, even setup numbers are used for training, and odd setup numbers are used for testing.

Northwestern-UCLA. Northwestern-UCLA dataset [35] was concurrently collected using three Kinect cameras from several viewpoints. It contains ten different action categories, including 1494 video samples collected from 10 different subjects. Each action sequence is represented by 20 key points at each timestamp. As the evaluation protocol in [35], we evaluated this dataset using three cameras with different views, two of them for training, and the third camera used for testing.

B. Implementation Details

For fair comparison to the base model, our experiments were conducted with network architecture consisting of eight stacked S3GAAR blocks. Each block has a certain number of output channels with the following numbers 64, 64, 128, 128, 128, 256, 256, and 256. Moreover, each block contains a residual connection for both the spatiotemporal graph attention module and the temporal modeling module.

Our experiments were conducted on a General-Purpose Graphics Processing Unit (GPGPU) node with an NVIDIA A100 Tensor Core GPU, 96 GB memory, and 4992 CUDA cores powered by Aziz supercomputer (King Abdulaziz University, Jeddah, Saudi Arabia). The models were built using PyTorch framework and trained twice with different learning rates (0.1 and 0.05) and decay factors (0.1 and 0.05). For the Optimization, we use SGD with a momentum (0.9), a weight decay of 0.0004, and cross-entropy as the loss function. For all our models, we set the batch size to 32 and the training epoch to 90.

C. ABLATION STUDY

In this section, we explore several experiments with different configurations and the effectiveness of segmented spatiotemporal graph attention on skeleton-based action recognition. Our experiments were conducted on X-sub benchmark NTU RGB + D dataset, which is a large-scale human skeleton dataset for action recognition. We use the same experimental setup as in previous studies, in which we evaluate the performance of the different models using the mean accuracy metric. We use CTR-GCN [10] as a baseline for the experimental settings related to dataset preprocessing and general model settings. In addition, we employ STGAT [6] as a baseline to build our segmented spatiotemporal graph.

TABLE I
COMPARISONS OF VARIOUS SETTINGS ON NTU RGB+D 60 XSUB INCLUDING SEGMENTS S , ADDITIONAL EDGES, HEADS ATTENTION N_h , TEMPORAL SAMPLING WINDOW τ AND DILATION d . MODELS A-H REPRESENT S3GAAR WITH DIFFERENT SETTINGS

Methods	S	Additional Edges	Whole-Skeleton	N_h	τ	d	Acc(%)
Baseline	×	×	✓	8	3	1	90.2
A	✓	✓	×	4	3	1	88.4
B	✓	✓	×	4	3	2	88.1
C	✓	✓	×	4	5	2	87.0
D	✓	✓	✓	4	3	2	89.3
E	✓	✓	✓	8	3	1	89.4
F	✓	✓	✓	8	5	1	89.6
G	✓	✓	✓	8	3	1	90.8
H	×	✓	✓	8	3	1	90.4

Segmented Graphs. As shown in Table I, we first replace the conventional graph which represents the whole-skeleton structure with three segmented graphs S (superior, inferior, and extremity) in models A, B, and C. In these models, each segmented graph contains not only the physically connected edges but also the proposed additional edges for each segment (shown in Figure 1). We observe the accuracies of A, B, and C models are not adequate. In models D-G, we extend the

segmented graphs with the whole-skeleton graph to study the importance of the whole-skeleton structure for actions that consider the motion of the whole-skeleton. We observe that the accuracy increases when whole-skeleton graph is added to Segmented Graphs because the model is capable of capturing human actions that are performed by the whole-skeleton structure. Finally, to validate the importance of segmented graphs as well as the additional edges, we conduct experiment H excluding segmented graphs. We notice that the accuracy is substantially dropped when all segmented graphs are excluded compared to G model. Therefore, the performance of G model where we include segmented and whole-skeleton graphs confirms the effectiveness of S3GAAR.

Segmented Spatiotemporal Graphs. To build segmented spatiotemporal graphs, we set the temporal sampling window $\tau=3,5$ and dilation $d=1,2$ for models A-H(see Table. I). These models aimed to investigate the impact of cross-spacetime motion from adjacency frames at timestamp t . We observe that models with $\tau=3$ with dilation $d=1$ perform better results compared to others.

Multi-heads attention . We explore various configurations including multi-head attention $N_h=4,8$. We observe that the accuracy increases regularly when the attention heads are increased. We also notice that the accuracy is considerably improved when the conventional graph is combined with segmented graphs, which demonstrates the robustness of S3GAAR.

TABLE II

COMPARISON FOR S3GAAR, STGAT [6] AND DSTANet [7] ON ACTION CATEGORIES: DAILY ACTIONS, TWO PERSON INTERACTIONS, AND MEDICAL CONDITIONS.

Classes	DSTANet	STGAT	S3GAAR
Average of all classes	89.1	88.2	90.8
Make a phone call	82.5	83.60	90.91 (+8.41,+7.31)
Reach into pocket.	81.8	80.30	87.59 (+5.79,+7.29)
Wear a shoe.	85	84.20	90.11 (+5.11,+5.91)
Take off jacket.	96.4	97.10	97.83 (+1.43,+0.73)
Touch other person's pocket.	89.8	93.80	96.36 (+6.56,+2.56)
Punching/slapping other person.	90.9	91.20	92.70 (+1.8,+1.5)
Pushing other person.	94.2	97.10	98.19 (+3.99,+1.09)
Hugging other person.	96.4	98.50	98.91(+2.51,+0.41)
Kicking other person.	92	94.90	94.93 (2.93,0.03)
Sneeze/Cough.	73.2	75	79.7 (+6.51,+4.71)
Backache.	93.1	90.2	95.65 (+2.55,+5.45)
Neckache.	80.4	86.5	86.23 (+5.83,-0.27)
Vomiting Condition.	84.7	85.1	88 (+3.3,+2.9)

Effect of classifying actions based on their categories

The majority of human actions in our lives include more than one person and most of them are performed by one of the body parts. Therefore, NTU datasets categorize actions into three main categories: daily actions, mutual actions(two-person interactions), and medical conditions [33]. Therefore, it is important to validate the effectiveness of S3GAAR models for classification based on the action category. The comparison between S3GAAR, STGAT [6] and DSTANet [7] in Table II shows that S3GAAR outperforms eight out of 11 mutual actions and seven out of nine medical conditions. These findings confirmed the efficiency of S3GAAR in the

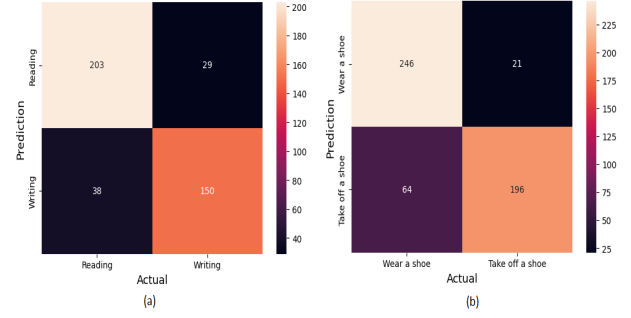


Fig. 6. (a) The confusion matrix for reading and writing and (b) The confusion matrix for putting on a shoe and taking off a shoe, respectively

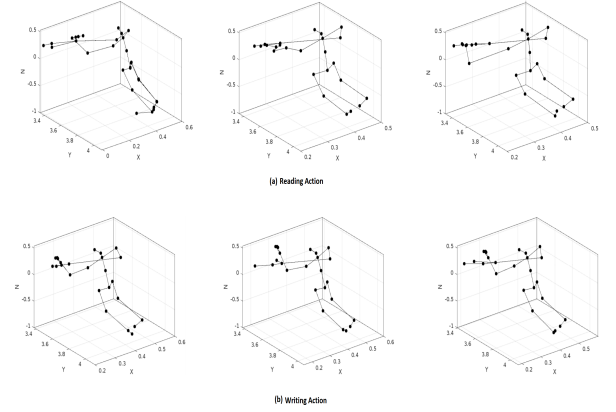


Fig. 7. (a) Skeleton motion for reading action. (b) Skeleton motion for writing action

classification of mutual actions and medical conditions as a result of segmented spatiotemporal modeling.

Our models achieve state-of-the-art results for the majority of the actions of NTU RGB+D 60 and NTU RDB+D 120 datasets (see Table II). However, as shown in confusion matrix 6, we notice that there are difficulties in differentiating between related actions such as writing and reading, putting on shoes, and taking off shoes. This is because of the identical skeleton motions in both actions (see Fig. 7 for reading compared to writing action. To measure the influence of confusion between putting on and taking off a shoe, we merge these actions into one action called putting on/taking off a shoe. Therefore, we compare the performance before and after merging by putting on a shoe/taking off (see Table. IV). We observe a remarkable improvement in the model performance.

D. Comparison with State-of-the-Arts Methods

Compared to state-of-the-art methods including GCN-based methods [11], [1], [12], [10], transformer-based methods [5], [7], [8], [9] and graph attention networks [14], [6], our method S3GAAR demonstrates superior performance across four different data streams, joint and bone streams, and two calculated motion streams for joint, and bone. Notably, our model achieves commendable results across NTU RGB+D 120, NTU RGB+D, and NW-UCLA datasets, as listed in Tables III and

TABLE III

DETAILED COMPARISONS OF S3GAAR AGAINST STATE-OF-THE-ART METHODS ON THE NTU RGB+D 60 AND NTU RGB+D 120 DATASETS IN THE FOUR DIFFERENT STREAMS(TWO STREAMS FOR JOINT AND BONE, AND TWO STREAMS FOR THEIR MOTIONS)

Methods	NTU RGB+D 60 (%)								NTU RGB+D 120 (%)							
	X-Sub				X-View				X-Sub				X-Set			
	J	B	J+B	4S	J	B	J+B	4S	J	B	J+B	4S	J	B	J+B	4S
Shift-GCN [36]	87.8	-	89.7	90.7	95.1	-	96	96.5	80.9	-	85.3	85.9	83.2	-	86.6	87.6
DC-GCN+ADG [37]	-	-	90.8	-	-	-	96.6	-	-	-	86.5	-	-	-	88.1	-
Dynamic GCN [38]	-	-	-	91.5	-	-	-	96	-	-	-	87.3	-	-	-	88.6
MS-G3D [39]	89.4	90.1	91.5	-	95	95.3	96.2	-	-	-	86.9	-	-	-	88.4	-
MST-GCN [40]	89	89.5	91.1	91.5	95.1	95.2	96.4	96.6	82.8	84.8	87	87.5	84.5	86.3	88.3	88.8
EfficientGCN-B4 [41]	-	-	-	91.7	-	-	-	95.7	-	-	-	88.3	-	-	-	89.1
2S-AGCN [42]	-	-	88.5	-	93.7	93.2	95.1	-	-	-	-	-	-	-	-	-
STGAT [6]	90.2	90.6	92.2	92.8	95.4	95.8	96.9	97.3	-	-	-	88.7	-	-	-	90.4
DSTA-Net [7]	-	-	-	91.5	-	-	-	96.4	-	-	-	86.6	-	-	-	89
KA-AGTN [14]	88.8	88.3	90.4	-	94.9	94.3	96.1	-	82.7	84.1	86.1	-	84.3	86.2	88	-
S3GAAR	90.8	90.1	91.9	92.9	95.7	94.6	96.4	96.8	86.2	85.5	88.3	88.8	87.6	87.2	89.9	90.4

TABLE IV

COMPARISON OF CLASSIFICATION ACCURACY OF DIFFERENT STATE-OF-THE-ART METHODS AGAINST S3GAAR ON THE NTU RGB+D 60 AND NTU RGB+D 120 DATASETS. S3GAAR* IS THE ACCURACIES WHEN PUTTING ON/TAKING OFF A SHOE ACTIONS WERE MERGED INTO ONE ACTION

Methods	NTU RGB+D 60 (%)		NTU RGB+D 120 (%)	
	X-Sub	X-View	X-Sub	X-Set
STGAT [6]	9.2	97.3	88.7	90.4
2s-SGR [5]	89.4	96.0	84.8	86.4
DSTA-Net [7]	91.5	96.4	86.6	89.0
KA-AGTN [14]	90.4	96.1	86.1	88.0
IIP-Transformer [8]	92.3	96.4	88.4	89.7
STTFormer [32]	92.3	96.5	88.3	89.2
PG-GCN [43]	91.8	95.8	88.4	88.8
4s MST-GCN [40]	91.5	96.6	87.5	88.8
CTR-GCN [10]	92.4	96.8	88.9	90.6
ST-DGAT [44]	91.1	96.4	86.5	88.2
S3GAAR	92.9	96.8	88.8	90.4
S3GAAR*	93.3	97.1	89.9	90.5

TABLE V

COMPARISON OF S3GAAR ACCURACY WITH DIFFERENT STATE-OF-THE-ART ON THE NW-UCLA DATASET

Model	Northwestern-UCLA Top-1 (%)
DC-GCN+ADG [45]	95.3
CTR-GCN [10]	96.5
HD-GCN [9]	97.0
TD-GCN [46]	97.4
S3GAAR	94.7

V, respectively. By harnessing our novel methodology, we are the first to model the features of each segmented part individually, which effectively learns different human actions and concentrates on the most operative part of the human body for each action. In terms of complexity, our S3GAAR involves additional parameters and computation cost compared with other attention mechanisms(See Table VI).

TABLE VI

COMPARISON OF S3GAAR COMPLEXITY WITH PREVIOUS METHODS ON THE NTU RGB+D 60 (X-SUB)

Methods	Param.	GFLOPs	Training Hours	Top-1 (%)
Baseline [6]	2.4M	14.6	10.1	90.2
DSTA-NET [7]	2.4M	10.2	6.0	88.2
MS-G3D [39]	1.4M	12.4	8.6	89.4
MS-AAGCN [47]	3.77M	-	-	90.0
S3GAAR	5.6M	25.2	14.2	90.8

VI. CONCLUSION AND FUTURE WORK

In this study, we introduce a novel segmented spatiotemporal skeleton graph-attention network (S3GAAR) for skeleton-based action recognition. The S3GAAR effectively learns segmented spatiotemporal features and concentrates on the most operative part of the human body for each action. The experimental results of our model achieved state-of-the-art results for the majority of actions on three large-scale datasets and made a remarkable improvement over the classification of mutual actions.

The findings of this study demonstrate the importance of segmented graph attention for focusing more on one part of the body than on other parts. However, most skeleton-based methods have difficulties in differentiating between related actions such as writing and reading, putting on shoes, and taking off shoes owing to identical skeleton motions (see Fig. 7). In the future, we will extend our work to involve techniques such as hand motion methods and object detection that may be involved in specific actions such as pen, knife, or phone.

ACKNOWLEDGMENTS

This work is supported by King Abdulaziz University, Jeddah, Saudi Arabia. The authors sincerely thank the KAU High-Performance Computing (HPC) Center, particularly for providing access to the Aziz Supercomputer and using their computational resources to conduct our experiments.

REFERENCES

- [1] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning graph convolutional network for skeleton-based human action recognition by neural searching," *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, 2020.
- [2] T. Si, F. He, Z. Zhang, and Y. Duan, "Hybrid contrastive learning for unsupervised person re-identification," *IEEE Transactions on Multimedia*, vol. 25, pp. 4323–4334, 2023.
- [3] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 20–36.
- [4] V. Veeriah, N. Zhuang, and G. Qi, "Differential recurrent neural networks for action recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, dec 2015, pp. 4041–4049. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.460>
- [5] X. Li, J. Zhang, S. Wang, and Q. Zhou, "Two-stream spatial graphormer networks for skeleton-based action recognition," *IEEE Access*, vol. 10, pp. 100 426–100 437, 2022.
- [6] L. Hu, S. Liu, and W. Feng, "Skeleton-based action recognition with local dynamic spatial-temporal aggregation," *Expert Systems with Applications*, vol. 232, p. 120683, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417423011855>
- [7] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition," in *ACCV*, 2020.
- [8] Q. Wang, J. Peng, S. Shi, T. Liu, J. He, and R. Weng, "Iip-transformer: Intra-inter-part transformer for skeleton-based action recognition," *CoRR*, vol. abs/2110.13385, 2021. [Online]. Available: <https://arxiv.org/abs/2110.13385>
- [9] J. Lee, M. Lee, D. Lee, and S. Lee, "Hierarchically decomposed graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 10 444–10 453.
- [10] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 13 339–13 348.
- [11] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, and N. M. Thalmann, "Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2272–2281.
- [12] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," 2018.
- [13] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [14] Y. Liu, H. Zhang, D. Xu, and K. He, "Graph transformer network with temporal kernel attention for skeleton-based action recognition," *Knowledge-Based Systems*, vol. 240, p. 108146, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0950705122000211>
- [15] H. Tian, X. Ma, X. Li, and Y. Li, "Skeleton-based action recognition with select-assemble-normalize graph convolutional networks," *IEEE Transactions on Multimedia*, vol. 25, pp. 8527–8538, 2023.
- [16] J. Shi, J. Zhong, and W. Cao, "Multi-semantics aggregation network based on the dynamic-attention mechanism for 3d human motion prediction," *IEEE Transactions on Multimedia*, pp. 1–13, 2023.
- [17] W. Tang, F. He, and Y. Liu, "Ydtr: Infrared and visible image fusion via y-shape dynamic transformer," *IEEE Transactions on Multimedia*, vol. 25, pp. 5413–5428, 2023.
- [18] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (indrn): Building a longer and deeper rnn," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5457–5466.
- [19] L. Sun, K. Jia, K. Chen, D. Y. Yeung, B. E. Shi, and S. Savarese, "Lattice long short-term memory for human action recognition," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2166–2175.
- [20] D. Feng, Z. Wu, J. Zhang, and T. Ren, "Multi-scale spatial temporal graph neural network for skeleton-based action recognition," *IEEE Access*, vol. 9, pp. 58 256–58 265, 2021.
- [21] H. Xia and X. Gao, "Multi-scale mixed dense graph convolution network for skeleton-based action recognition," *IEEE Access*, vol. 9, pp. 36 475–36 484, 2021.
- [22] A. Martínez-González, M. Villamizar, and J. Odobez, "Pose transformers (potr): Human motion prediction with non-autoregressive transformers," in *IEEE/CVF International Conference on Computer Vision - Workshops (ICCV)*, 2021.
- [23] Y. Zhang, B. Wu, W. Li, L. Duan, and C. Gan, "Stst: Spatial-temporal specialized transformer for skeleton-based action recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2021, p. 3229–3237. [Online]. Available: <https://doi.org/10.1145/3474085.3475473>
- [24] F. Shi, C. Lee, L. Qiu, Y. Zhao, T. Shen, S. Muralidhar, T. Han, S. Zhu, and V. Narayanan, "STAR: sparse transformer-based action recognition," *CoRR*, vol. abs/2107.07089, 2021. [Online]. Available: <https://arxiv.org/abs/2107.07089>
- [25] H. Qiu, B. Hou, B. Ren, and X. Zhang, "Spatio-temporal tuples transformer for skeleton-based action recognition," *ArXiv*, vol. abs/2201.02849, 2022.
- [26] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3d human pose estimation with spatial and temporal transformers," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.
- [27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- [28] C. Plizzari, M. Cannici, and M. Matteucci, "Spatial temporal transformer network for skeleton-based action recognition," in *Pattern Recognition. ICPR International Workshops and Challenges*, A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, and R. Vezzani, Eds. Cham: Springer International Publishing, 2021, pp. 694–701.
- [29] C. Liu and H. Zhou, "Fully attentional network for skeleton-based action recognition," *IEEE Access*, vol. 11, pp. 20 478–20 485, 2023.
- [30] J. Wu, F. Hao, W. Liang, and J. Xu, "Transformer fusion and pixel-level contrastive learning for rgb-d salient object detection," *IEEE Transactions on Multimedia*, vol. 26, pp. 1011–1026, 2024.
- [31] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 459–479.
- [32] H. Qiu, B. Hou, B. Ren, and X. Zhang, "Spatio-temporal tuples transformer for skeleton-based action recognition," *CoRR*, vol. abs/2201.02849, 2022. [Online]. Available: <https://arxiv.org/abs/2201.02849>
- [33] A. Shahrourdy, J. Liu, T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2016, pp. 1010–1019. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.115>
- [34] J. Liu, A. Shahrourdy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, p. 2684–2701, October 2020. [Online]. Available: <http://arxiv.org/pdf/1905.04757>
- [35] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning, and recognition," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2649–2656.
- [36] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 180–189.
- [37] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu, "Decoupling gcw with dropout module for skeleton-based action recognition," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Springer International Publishing, 2020.
- [38] F. Ye, S. Pu, Q. Zhong, C. Li, D. Xie, and H. Tang, "Dynamic gcw: Context-enriched topology learning for skeleton-based action recognition," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:220845919>

- [39] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 143–152.
- [40] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1113–1122.
- [41] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1474–1488, 2023.
- [42] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019.
- [43] H. Chen, Y. Jiang, and H. Ko, "Pose-guided graph convolutional networks for skeleton-based action recognition," *IEEE Access*, vol. 10, pp. 111 725–111 731, 2022.
- [44] M. Rahevar, A. Ganatra, T. Saba, A. Rehman, and S. A. Bahaj, "Spatial-temporal dynamic graph attention network for skeleton-based action recognition," *IEEE Access*, vol. 11, pp. 21 546–21 553, 2023.
- [45] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu, "Decoupling gcN with dropgraph module for skeleton-based action recognition," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Springer International Publishing, 2020.
- [46] J. Liu, X. Wang, C. Wang, Y. Gao, and M. Liu, "Temporal decoupling graph convolutional network for skeleton-based gesture recognition," *IEEE Transactions on Multimedia*, pp. 1–13, 2023.
- [47] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 9532–9545, 2020.



Musreah Abdo Ghaseb received a bachelor's degree in computer engineering from Taif University, Saudi Arabia in 2017. He is currently a dedicated and aspiring master's student in computer science at King Abdulaziz University, where he is pursuing advanced studies in artificial intelligence, computer vision, and software engineering. In his current pursuit of a Master's degree, Ghaseb is actively involved in promising research projects related to human motion analysis and skeleton-based action recognition using deep learning.



Ahmed Elhayek received his Master's degree from Saarland University in the year 2010 where he developed a novel simultaneous interpolation and deconvolution approach for 3D reconstruction of cell images. Then, he worked on implementing a deblurring algorithm for high-speed motion capture during his internship at the Max-Planck Institute for Informatics (Germany). He received his Ph.D. degree from the Max-Planck Institute and Saarland University in 2015. The research topic of his Ph.D. was "Human motion capture in general uncontrolled environments with sparse multi-camera setup". After his Ph.D., he worked in the Augmented Vision group at DFKI (German Research Centre for Artificial Intelligence) and the University of Kaiserslautern for three years. In 2018, he joined the faculty of Computer and Cyber Sciences at the University of Prince Mugrin. Currently, Dr. Elhayek is the head of the Artificial Intelligence Department at the same college.



Fawaz Alsolami received his M.A.Sc in Electrical and Computer Engineering from University of Waterloo, Canada, in 2008, and his Ph.D. degree in Computer Science from KAUST University, Thuwal, Saudi Arabia. He is an associate professor of computer science at King Abdulaziz University. His research interests are artificial Intelligence, machine learning and data Mining, and combinatorial optimization. He also published many articles and one monograph.



Abdullah Marish Ali received the B.Sc. degree in computer science from Al-Mustansiriya University, Baghdad, Iraq, the M.Sc. degree in computer science from King Abdulaziz University (KAU), Jeddah, Saudi Arabia, and the Ph.D. degree in computer science from Universiti Teknologi Malaysia (UTM), Malaysia. He is currently an Assistant Professor with the Department of Computer Science, Faculty of Computing and Information Technology, KAU. His research interests include cloud computing, software agents, data mining, information retrieval, machine learning, and cyber security.