IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 January 2023, accepted 16 February 2023, date of publication 22 February 2023, date of current version 7 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3247820

RESEARCH ARTICLE

Spatial–Temporal Dynamic Graph Attention Network for Skeleton-Based Action Recognition

MRUGENDRASINH RAHEVAR[1], AMIT GANATRA[2], (Member, IEEE), TANZILA SABA[3], (Senior Member, IEEE), AMJAD REHMAN[3], (Senior Member, IEEE), AND SAEED ALI BAHAJ[4,5], (Member, IEEE)

[1]Chandubhai S. Patel Institute of Technology, Charotar University of Science and Technology, Changa, Anand, Gujarat 388421, India
[2]Parul University, Wagodhiya, Gujarat 391760, India
[3]Artificial Intelligence and Data Analytics Laboratory (AIDA), College of Computer and Information Sciences (CCIS), Prince Sultan University, Riyadh 11586, Saudi Arabia
[4]MIS Department, College of Business Administration, Prince Sattam bin Abdulaziz University, Alkharj 11942, Saudi Arabia
[5]Department of Computer Engineering, College of Engineering and Petroleum, Hadhramaut University, Mukalla 50511, Yemen

Corresponding author: Saeed Ali Bahaj (saeedalibahaj@gmail.com)

This work was supported in part by the Artificial Intelligence and Data Analytics (AIDA) Laboratory, College of Computer and Information Sciences (CCIS), Prince Sultan University, Riyadh, Saudi Arabia.

ABSTRACT Human body skeleton, acting as a spatiotemporal graph, is increasing attentions of researchers to adopt graph convolutional networks (GCN) to mine the discriminative features from skeleton joints. However, one of GCN's flaws is its inability to handle long-distance reliance between joints. In this regard, graph attention network (GAT) was recently suggested, which combines graph convolutions with a self-attention mechanism to extract the most informative joint of a human skeleton and increase the model accuracy. However, GAT can compute only static attention: for each query node, the attention rank is same which severely hurts and limits the expressivity of an attention mechanism. In this work, we present a spatial-temporal dynamic graph attention network (ST-DGAT) to learn the spatial-temporal patterns of skeleton sequences. For dynamic graph attention, we tweak the order of weighted vector operations in GAT, our approach achieves a global approximate attention function, making it strictly superior to GAT. Experiments show that by fixing the order of internal operation of GAT the proposed model achieved better action classification results while maintaining the same computing cost as GAT. The proposed framework has been evaluated on well-known publicly available large-scale datasets NTU60, NTU120, and Kinetics-400, which notably outperforms state-of-the-art (SOTA) results with an accuracy of 96.4%, 88.2%, and 61.0%, respectively.

INDEX TERMS Skeleton, action recognition, graph attention network, multihead attention.

I. INTRODUCTION

The amount of multimedia content (e.g. video) uploaded through assorted nodes has skyrocketed in recent years. As a result, the autocratic need has arisen to automate human action analysis based on video data. In the past decade, human action recognition in extended video sequences, in conjunction with localizing actions both spatially and temporally [1], [2] is a prominent research area due to its extensive applications like sports analysis, human-machine interaction, and intelligent robotics systems [3], [4], [5], [6], [7] and so on. Compared with modalities like RGB-D and optical flow [8], [9], [10], the skeleton-based approaches are computationally fast and simplify storage data [11], [12], [13]. Moreover, skeleton data is lightweight and robust against irrelevant objects, body scale, and camera viewpoint and can be predicted efficiently. However, the human skeleton is indeed a graph structure, not a sequence making it difficult for proven neural network models like CNN to perform well. The generalized form of CNN is GCN, a powerful graph representation learning method capable of handling arbitrary graph structure data, which has received increasing attention among researchers.

Handcrafted features are often used in traditional skeleton-based action recognition techniques to describe the motif of co-occurrences from skeleton sequences [14], [15], [16], [17].

The associate editor coordinating the review of this manuscript and approving it for publication was Sathish Kumar.

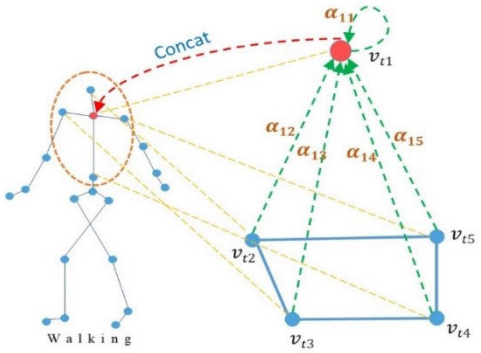

FIGURE 1. Illustration of attention mechanism by root node on its neighborhood in the spatiotemporal environment.

On the other hand, recurrent neural networks (RNNs) and convolutional neural networks (CNNs) possess a solid ability to model sequential features. These approaches showed an impressive result with Euclidean data, e.g. images and video. However, their direct implementation is not intuitive in the non-euclidean space since they lack the skeleton structure connection property, resulting in partial generalization and poor robustness.

Recent research [12], [18], [19] considers an entire graph for processing and shared weight vector with every node, which obscuring the importance of distinct nodes for different activities. This bleak outcome and effect on a model's overall performance on graph classification significantly increases computational cost and unintentionally presents noise. To tackle this, graph attention networks (GAT) [20] were led into skeleton-based action recognition. We study an attention model that accentuates the important nodes while quashing redundant nodes. The advantage of the attention model is that it handles variable-sized inputs effectively. The GAT model has the disadvantage of sharing the attention score with all nodes and not being bound by the query node. However, Brody et al. [21] proposed GATv2 to alleviate GAT's restriction by reordering the weighted vector operations and putting it ahead of GAT in terms of performance. To the best of our knowledge, GATv2 model has never been applied to the field of skeleton-based action recognition. This motivates us to develop a model that allows the extraction of skeleton properties in a spatiotemporal environment.

Based on ST-GCN [19] network, we present a spatial-temporal dynamic graph attention network (ST-DGAT) to classify human actions. As illustrated in Fig 1, first, we leverage the long-sequence skeleton data to construct a spatiotemporal graph that naturally unit the human body joints (node) and conjugates in time. Then we feed the generated spatiotemporal graph into our ST-DGAT network. Afterward, the model computes the dynamic attention coefficient by fixing the order of internal operation in GAT. Finally, we calculate the class rank by using adjacent node properties and dynamic attention score. We apply the multihead attention [22] to stiffen the learning process and also improve the performance of action recognition w.r.t accuracy. We ran extensive tests on three large-scale action recognition datasets NTU60, NTU120, and Kinetics, and found that our model ST-DGAT outperforms the baseline model and achieves SOTA performance.

The following is a summary of our contribution:

- We propose a novel dynamic graph attention network by changing the order of internal operation of GAT to model the skeletons features in a spatiotemporal environment.
- We further conducted extensive experiments to compare static attention and dynamic attention.

The proposed network surpasses the SOTA performance on three extensive datasets NTU RGB+D 60, NTU-RGB+D 120, and Kinetic without hyper-parameter tuning.

II. RELATED WORK

A. CNN, RNN AND LSTM BASED APPROACH FOR SKELETON-BASED ACTION RECOGNITION

Recently, deep learning techniques have achieved remarkable gains in vision tasks, and numerous methods are suggested for action recognition. A current model falls into two architectures. The first architecture, called convolutional neural network (CNN) based model, [23] transforms every generated clip's to long-term temporal skeleton sequences and applies the convolutional operator to the entire frame sequences parallelly to integrate spatial structural features for action recognition. In [24], created a view-independent sequence-based method for describing skeleton sequences as a series of color images and feeding them into the CNN model for classifying actions. Although CNN-based algorithms excel in the spatial information domain, they commonly neglect the temporal information domain. The author of [25] uses a 2D convolutional to learn the temporal information. Such a procedure is sluggish and emphasizes unneeded features, which has a negative impact on the performance of the model. The second architecture, called LSTM [26], [27], [28] has effectively modelled temporal dependency as compared to CNN. In [29] presented a spatial-temporal LSTM network with a gating mechanism to remove erratic input caused by occlusion and noise. In [30], The first tier records the skeleton pattern, while the second tier applies the attention model to enrich the global context to recognize human action.

The author of [31] presented that each skeleton part holds distinct LSTM cells to extract the features; then synthesis features are used in place of sharing a cell. Also, sequence base approaches are significantly increasing computational costs. As the human body is logically a graph data, graph-based techniques are more apparent than sequential methods.

B. GCN FOR SKELETON-BASED ACTION RECOGNITION

Structure graph data are more general than sequential data, which cannot be directly fed to conventional approaches such as CNNs and RNNs. The principle of applying GCN on the graph has two perspectives i) spectral perspective, where locality of graph convolutional are applied in

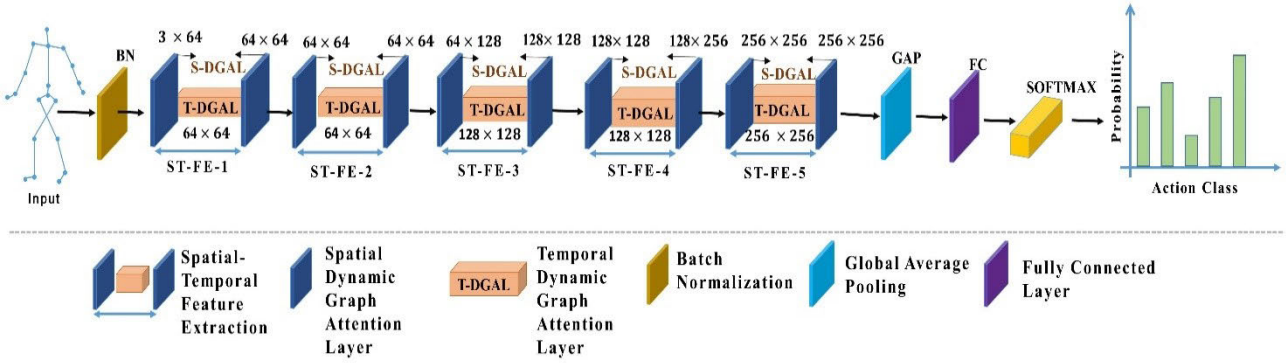

FIGURE 2. The Architecture of ST-DGATs.

spectral-domain [32], [33], ii) spatial perspective defines graph nodes and their neighbors are used to define convolutional filters directly [34]. Inspired by a successful GNN-based method Yan et al. [19] first proposed a spatial-temporal graph convolutional network (ST-GCN), in which captures skeleton feature by applying graph convolutional operator and learn temporal dynamics via 2D convolutional operator. Shi et al. [35] congregate adaptive graph approach and two-stream framework based on ST-GCN and suggested a two-stream adaptive GCN(2s-AGCN), which can connect the joint(first-order) and bones(second-order) features to improves the action recognition. In [36], the authors presented a directed acyclic graph (DAG) mechanism specially designed to extract relational patterns among joints and bones to enhance the action recognition task. In addition, [37] applied a part-based GCN to study the relation of human gestures using analogous joint coordinates and temporal displacements. In [38] model learns the spatial feature but loss the focus to learn the temporal features. Plizzari et al. [39] presented the ST-TR, a spatio-temporal transformer network that uses transformer technique to determine the self-attention score of each body joints. However, because the implicit connection between specific joints is not taken into account, this method will exaggerate the correlation among particular joints. The addition of interdependence between those joints not only raises the computing cost but also makes the model more difficult to understand when it comes to identifying actions. In a Spatio-temporal Graph, Penget al. [40] suggested graph triplet pooling, which can amalgamate or eliminate insignificant vertices.

C. GRAPH ATTENTIONAL MODEL

Attention mechanism attracted researchers widely and proposed numerous approaches [41], [42]. The key benefits of the attention mechanism are that they are competent with variable-sized inputs since they focus on informative parts of the input and allow them to make the appropriate decisions. During graph aggregation, the attention mechanism is used to dynamically compute the weight of every node's neighbor. Reference [43] designed an end-to-end memory attention network to execute temporal-then-spatial feature recalibration, but it lacks to find key joints in the spatial domain. In [44] the author proposed an end-to-end spatial-temporal attention architecture for learning attention weight and applying it over joints to focus on discriminative joints within each frame. The GAT proposed by Velickovic et al. [20] is considered the most popular GNN framework for learning with graphs. GAT calculates a pair-wise normalized attention score between two neighbors' nodes and pays attention to each node without the use of any expensive matrix operations. The GAT model has the drawback of sharing the attention score with every node and being unconstrained by the query node. However, Brody et al. [21] proposed GATv2 to address the constraint of GAT by changing the order of weighted vector operations in GAT. The GATv2 achieves a global approximator attention function, making it strictly superior to GAT.

III. PROPOSED APPROACH

Notation

A skeleton graph denoted as $G_g = (V_g, E_g)$ with N nodes in frame T, where $(V_g = [v_{ti}]_{t=1,i=1}^{T,N})$ denotes a vertex set of *ith* body joints on frame. *i*. E_g represents the relationship between nodes, denoted as $e_{e,j} = \langle (v_1, v_2), \ldots, (v_{k-1}, v_k)) \rangle \in E_g$, $i \neq j$. Let $A \in R^{NxN}$ an adjacency matrix that indicates whether pair of joints (i, j) are adjacent or not. The adjacency matrix represents the topological network by setting $A_{i,j} = 1$ when (v_i, v_j) is connected otherwise, $A_{i,j} = 0$. Let $D \in R^{NxN}$ where $D_{i,j} = \sum_i A_{i,j}$ represents the corresponding diagonal node degree matrix. The default graph is generally computed as:

$$G^{default} = \tilde{D}^{-0.5}\tilde{A}\tilde{D}^{-0.5} \quad (1)$$

where $\tilde{A} = A + I_N$ is the normalized adjancency matrix with self-loops, $\tilde{D}$ is the corresponding node degree matrix of $\tilde{A}$. The goal is to learn a function $f(G_i) \rightarrow L_i$, where G_i and L_i are represented as respectively. The notations used in this paper are illustrated in Table 1.

A. SPATIAL DYNAMIC GRAPH ATTENTION LAYER (S-DGAL)

1) SAMPLING FUNCTION

The sampling function is calculated on the neighbour set $F_S(V_i^t) = \{V_i^t \| d(V_i^t, V_j^t) \leq 1\}$ of a single frame w.r.t center

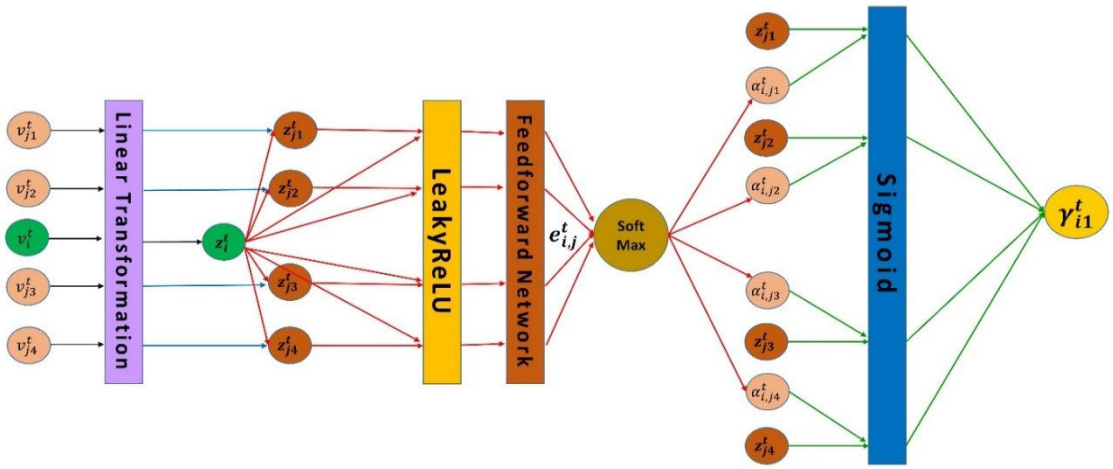

FIGURE 3. The Illustration of computing dynamic graph attention network.

TABLE 1. Commonly used notations.

Notation	Description
G	A Graph
V	The set of nodes in a graph.
E	The set of edges in a graph.
v_i^t	vertex set of ith body joints on frame i
w	Weight matrix
z_i^t	Embedding vector
$e_{i,j}^t$	An edge $e_i j \in E$
a^T	^{T} represents transposition and a is a single-layer feedforward neural network.
$\sigma(\cdot)$	The sigmoid activation function
$\alpha_{i,j}^t$	Attention Coefficient

node $v_t i$. Here $d(v_i^t, v_j^t)$ is the collection of neighbour nodes of v_i^t. Thus the sampling function is represented as

$$p(v_i^t v_j^t) = v_j^t \tag{2}$$

2) DYNAMIC ATTENTION COEFFICIENT

Inspired by [20], we applied a shared linear transformation, parametrized by a weight matrix W^t, to each body joint and convert input features into significantly high-level features at time t, which can be represented as

$$z_i^t = w v_i^t \tag{3}$$

Within each frame of given skeleton sequences, we use the self-attention technique to retrieve the essential features that encapsulate the relationship between human body parts. The attention coefficient (followed by LeakyReLU nonlinearity) can be represented as:

$$e_{i,j}^t = LeakyReLU[a^T(z_i^t, z_j^t)], \quad j \in F_s(v_i^t) \tag{4}$$

where ^{T} represents transposition and a is a single-layer feedforward neural network But GAT can compute only static attention which severely hurts and limits the expressivity of an attention mechanism. To fix this limitation, Brody et al. proposed GATv2 by switching the order of internal operations in the attention function of GAT. The GATv2 outperforms the GAT w.r.t accuracy. Equation (3) can be re-written as:

$$e_{i,j}^t = a^T(LeakyReLU[w(z_i^t, z_j^t)], \quad j \in F_s(v_i^t) \tag{5}$$

We apply masked attention mechanism on graph structure to determine the attention coefficient. This clearly demonstrates that the $e_(i, j)^t$ is exclusively calculated for first-order neighbors. we apply a softmax function, which normalizes the attention score across all neighbor nodes. This means how strong the correlations between each pair of body joints are. Thus, the attention coefficient function is represented as:

$$\alpha_{i,j}^t = soft_j(e_{i,j}^t) = \frac{exp(e_{i,j}^t)}{(\sum_{k \in F_s(v_i^t)} exp(e_{i,k}^t)} \tag{6}$$

Multi-head attention was presented as a way for the center node and neighboring nodes to concurrently attend multiple representation embedding, which can inhibit the overfitting of the network. Thus we apply the multi-head attention to obtain a new node embedding set $(\gamma_{i_1}^t, \ldots, \gamma_{i_H}^t)$ for the same node; the function is signed as:

$$\gamma_i^t = \|_{k=1}^k \sigma(\sum_{k \in F_s(v_i^t)} (\alpha_{i,j}^{t(k)}) W^{t(k)} z_i^{t(k)}) \tag{7}$$

where $\|$ represents concatenation,K is the number of independent attention heads. $\alpha_{i,j}^{t(k)}$ is the k th normalized attention weight among the root node i and neighbor node j. The $W^{t(k)}$ is the corresponding linearly projection weight matrix of k_{th} head.

B. TEMPORAL GRAPH ATTENTION LAYER (T-DGAL)

The notation used here is opposite to S-DGAL; subscripts denote time, whereas superscripts denote joint. With the T-DGAL module, temporal dynamics is captured for each

joint across consecutive frames. The correlations across frames are computed by changing the embedding of comparable joints throughout the temporal dimension. The final formulation of T-DGAL is symmetric to that given in Equations (6) and (7) for S-DGAL

$$\alpha^{v}_{t,u} = soft_{tu}(e^{v}_{t,u}) = \frac{exp(e^{v}_{t,u})}{(\sum_{k \in F_s(v^{v}_{t,u})} exp(e^{v}_{t,q})} \tag{8}$$

$$\gamma^{u}_{t} = \|^{k}_{k=1} \sigma(\sum_{k \in F_s(v^{v}_{t,u})} (\alpha^{v(k)}_{t,u}) W^{v(k)} z^{v(k)}_{t,u}) \tag{9}$$

Here, $\alpha^{v}_{t,u}$ is the correlation score. v_{ti},v_{ui} refer to the same joint v at two separate instants t,u and $\gamma^{v}_{t,u}$ is the resulting node embedding. Multi-head attention mechanism applied here is the same as S-DGAL.

IV. EXPERIMENTS AND RESULT

We undertake comprehensive ablation studies in this section to fairly examine the efficacy of leveraging dynamic graph attention compared to static graph attention. The proposed model outperforms the SOTA on three large-scale datasets. We also present the proposed model experiment settings as well as the datasets used to test the model

A. DATASETS

1) NTU-RGB+D 60 (NTU60) [45]

This is a KinectV2 camera capture dataset consisting of 56,880 skeleton sequences over 60 action classes captured from 40 daily actions, 9 actions depending on a medical condition and 11 joint actions. Each skeleton sequence has 25-joints for each subject. The author recommended two evaluation settings: 1) Cross-subject(X-Sub), for training and testing, data split into 40,320 samples and 16560 samples with 40 distinct subjects. 2) Cross-view(X-View), training set containing 37,920 samples captured by 1 camera and the rest 18,960 samples used for training.

2) NTURGB+D 120 (NTU120) [46]

This is a more complex and challenging version of the NTU60 dataset, enclosing an additional 60 classes, 57,600 video samples, and 4 million frames. Three cameras record all video clips in 32 different indoor scenario setups, performed by 106 diverse performers in a group of people ages between 10 and 57. The extended dataset recommended two evaluation protocols Cross-subject (X-Sub) and cross-setup (X-Setup).

3) KINETICS-400 [47]

This dataset contains raw videos of high-quality human actions. It incorporates 300,000 video clips with 400 action classes and at least 400 videos per action class. Each clip has a duration of around 10 seconds and is mined from diverse YouTube videos. OpenPose [48] toolbox is adopted to extract skeleton information from RGB videos and estimates 2D pixel coordinate (x,y) of the predicted joints and their corresponding confidence score c for all 18 body joints, and each joint is defined by a 3D vector (x,y,c). All videos are resized to 340 X 256 resolution with 30fps. To compare the proposed model with the literature, we reported top-1 and top-5 accuracies.

B. EXPERIMENTAL SETTINGS

Python, OpenPose [48], and the deep learning framework PyTorch [49] are used in our experiments to build the ST-DGAT model. A batch normalization layer initially normalizes the input skeletal sequences. The backbone of ST-DGAT is constructed with 9 blocks, where the feature dimension of the first three layers has 64 channel output, followed by three layers with 128 channel output and the last three layers have 256 channel output. The temporal kernel size is set to 9 for these layers. To alleviate the overfitting problem, we used DropAttention [50] for regularizing attention weight with probabilities 0.5 at each unit. The number of attention layers and the multi-head unit used for NTU60/120 and Kinetics-400 are (1,8) and (1,8), respectively. We started with a 0.0005 learning rate and reduced it by 0.1 per 10 epochs. The model applied a global average pooling layer on the resulting tensor. Finally, we train a softmax classifier with the feature vector using the conventional cross-entropy loss. The model is trained for epochs (80,40) with batch sizes of (32, 128) for NTU60/120 and Kinetics-400 respectively.

C. ABLATION STUDY

We conducted extensive experiments to determine the best K head on NTU60 and Kinetic dataset of our model. Furthermore, we compare the proposed model with a baseline to analyze the computational cost. In our ST-DGAT model, we want to determine the impact of varied head (K) counts. We set head value K=2 as lower limit and $K = 10$ as upper limit with an interval of 2. The output channel of ST-DGAT is 64,128,256 and 512. For $K = 6$, the outpu channel is set to 126,252,510 and for $K = 10$ the output channel is set to 120,250 and 510. Table 2 shows that ST-DGAT consistently outperforms baseline ST-GCN, which demonstrating the efficacy of our proposed module. In Table 3, we presented training and testing time per epoch on NTU60 and Kinetic dataset. As we can see our ST-DGAT model is 1.4h faster in training and 0.76h faster in testing than the baseline on NTU60 dataset. Our model is even faster on Kinetic dataset, which contains 2.4M video clips (300 frame/clip).

D. COMPARISON WITH STATE OF THE ART

On NTU60, NTU120, and Kinetics-400 datasets, our model ST-DGAT surpasses current SOTA methods on recommended benchmarks. In Table 4, we report the action recognition accuracy on NTU60 dataset with recommended benchmarks cross-subject and cross-view. Our model ST-DGAT achieves an accuracy of 91.1% (X-Sub) and 96.4%(X-View). The proposed model surpassed RNN-based method AGC-LSTM by 1.9% (X-Sub) and 1.4% (X-View). Our model still outperforms the CNN-based technique VACNN with an

TABLE 2. Comparison with baseline in on NTU60 and Kinetics dataset.

Method	NTU60		Kinetics	
	X-Sub	X-View	Top-1%	Top-5%
Baseline ST-GCN	80.7%	88.9%	31.8	53.6
ST-DGAT$_2h$	84.8	92.8	24.2	44.7
ST-DGAT$_4h$	87.9	94.8	29.6	50.4
ST-DGAT$_6h$	91.2	95.3	33.7	56.6
ST-DGAT$_8h$	**91.7**	**96.7**	38.2	61.0
ST-DGAT$_10h$	90.9	95.1	32.8	56.1

TABLE 3. Comparison of Training and Testing time with baseline on NTU60 and Kinetic datasets.

Method	No.of param. (NTU60)	No.of Skeleton Sequences (Kinetics)	TRN* (one epoch)	TST* (one epoch)
Baseline (ST-GCN)	3.2M	-	2.77h	0.950h
Baseline (ST-GCN)	-	2.4M	0.578h	0.186h
ST-DGAT	2.29M	-	1.32h	0.185h
ST-DGAT	2.29M	-	0.566h	130h

TABLE 4. On the NTU60 dataset, action recognition was compared to existing SOTA techniques for X-Sub and X-View benchmarks.

Method	X-Sub(%)	X-View(%)	Year
Part-aware LSTM [45]	62.9	70.3	2016
Two-Stream RNN [26]	71.3	79.5	2017
STA-LSTM [44]	73.4	81.2	2017
VA-LSTM [27]	79.2	87.7	2017
AGC-LSTM [28]	89.2	95.0	2019
MT-CNN [14]	83.2	89.3	2018
SAN [51]	87.2	92.7	2019
VACNN [25]	88.7	94.3	2019
ST-GCN [19]	81.5	88.3	2018
PB-GCN [37]	87.5	93.2	2018
GCMVT [52]	84.2	90.2	2019
2s-Shift-GCN [53]	89.7	95.0	2020
ST-TR [39]	89.9	96.1	2020
DSTA-Net [54]	**91.5**	96.4	2020
Tripool [40]	89.5	96.4	2021
AAM-GCN [38]	90.4	96.2	2021
ST-DGAT(ours)	91.1	96.4	-

accuracy of 2.4% (X-Sub) and 2.1% (X-View). Experiments in Table 4 show that our method surpasses the GCN and attention-based approaches, demonstrating that our model outperforms them in terms of accuracy.

Furthermore, we compare ST-DGAT to ST-GCN (the work's baseline) and determine that it outperforms by 9.6% and 8.1% on the X-Sub and X-View benchmarks, respectively. However, DSTA-Net[50] shows competitive results, which combine four streams and arouse a huge computational cost than our model.

Table 5 presents the performance of our model on NTU120 dataset, which is based on joint information only; our model achieves higher accuracy than SOTA approach. As can be

TABLE 5. On the NTU120 dataset, compare top-1 accuracy with current SOTA techniques.

Method	X-Sub(%)	X-Setup(%)	Year
ST-LSTM [29]	55.7	57.9	2016
GCA-LSTM [30]	61.2	63.3	2018
RotClip+MTCNN [55]	61.8	81.2	2018
2s-Shift-GCN [53]	86.6	87.7	2020
ST-TR [39]	85.1	87.1	2020
DSTA-Net [54][50]	86.6	89.0	2020
Tripool [40]	80.1	82.8	2021
ST-DGAT	**86.5**	**88.2**	-

TABLE 6. On the Kinetic-400 dataset, compare the top-1 and top-5 classification accuracies with current SOTA techniques.

Method	Top-1(%)	Top-5(%)	Year
TCN [23]	20.3	40.0	2017
ST-GCN [19]	30.7	52.8	2018
SAN [51]	35.1	55.7	2020
2s-Shift-GCN [53]	37.1	60.1	2020
ST-TR[35]	37.4	59.8	2020
AAM-GCN [38]	37.5	60.5	2010
Tripool [40]	34.1	56.2	2021
ST-DGAT	**38.2**	**61.0**	-

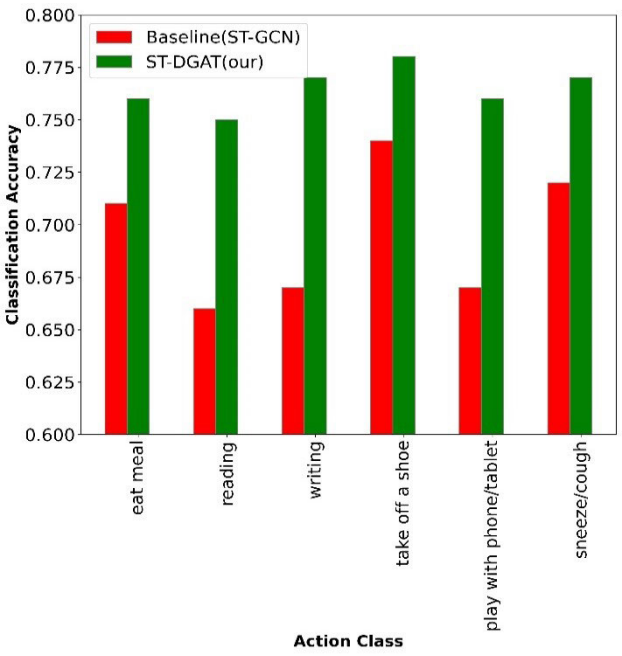

FIGURE 4. Comparison of the difficult classes with accuracies less than 80% on NTU60.

seen from Table 6, we achieved an accuracy of 86.5% on X-Sub and 88.2% on X-Setup. We compare our method on third-largest Kinetics-400 data, results presented in Table 6. Our model ST-DGAT surpassed baseline ST-GCN [20] by 7.5% in top-1 and 8.2% in top-5. Our method also surpassed recent SOTA methods

As shown in Figure 4, The baseline model, on the other hand, has trouble with many challenging classes, such as eating, reading, writing, taking off shoes, playing with

phone/tablet, and sneezing/coughing. Small differences in repetitive motion distinguish these classes, making action recognition more difficult.

V. CONCLUSION

In this work, we proposed a dynamic GAT for skeleton-based action recognition in spatial-temporal environment. The attention based GCN model shares the attention score with every node and is unconstrained by the query node. The ST-DGAT model computes dynamic graph attention by tweaking the order of weighted vector operations in GAT. We feed the extracted spatiotemporal joint features to classifier for action recognition. We conducted ablation studies to prove the effectiveness of our approach. Extensive experiments carried out in this work and results show that compared to ST-GCN(baseline) achieved 9.6% and 81% higher accuracies on both benchmarks of NTU60. Our model extortionate by 7.5% in top-1 and 8.2% in top-5 on Kinetics dataset. We prove that the model outperforms on NTU60 and NTU120 and achieved SOTA level on kinetic-400. In future work, our model will be investigating human-object interaction and scene information for better action
recognition.

REFERENCES

[1] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[2] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[3] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 1–43, 2011.

[4] F. Rezazadegan, S. Shirazi, B. Upcrofit, and M. Milford, "Action recognition: From static datasets to moving robots," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 3185–3191.

[5] T. Shu, X. Gao, M. S. Ryoo, and S.-C. Zhu, "Learning social affordance grammar from videos: Transferring human interactions to human–robot interactions," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 1669–1676.

[6] T. Saba, A. Rehman, R. Latif, S. M. Fati, M. Raza, and M. Sharif, "Suspicious activity recognition using proposed deep L4-branched-actionnet with entropy coded ant colony system optimization," *IEEE Access*, vol. 9, pp. 89181–89197, 2021.

[7] M. T. Ubaid, T. Saba, H. U. Draz, A. Rehman, M. U. Ghani, and H. Kolivand, "Intelligent traffic signal automation based on computer vision techniques using deep learning," *IT Prof.*, vol. 24, no. 1, pp. 27–33, Jan. 2022.

[8] L. Fan, W. Huang, C. Gan, S. Ermon, B. Gong, and J. Huang, "End-to-end learning of motion representation for video understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6016–6025.

[9] L. Wang, W. Li, W. Li, and L. Van Gool, "Appearance-and-relation networks for video classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1430–1439.

[10] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 803–818.

[11] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3288–3297.

[12] B. Li, X. Li, Z. Zhang, and F. Wu, "Spatio-temporal graph routing for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8561–8568.

[13] Y. Yan, J. Xu, B. Ni, W. Zhang, and X. Yang, "Skeleton-aided articulated motion generation," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 199–207.

[14] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 579–583.

[15] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1–7.

[16] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 588–595.

[17] M. A. Khan, Y.-D. Zhang, M. Alhaisoni, S. Kadry, S.-H. Wang, T. Saba, and T. Iqbal, "Correction to: A fused heterogeneous deep neural network and robust feature selection framework for human actions recognition," *Arabian J. Sci. Eng.*, vol. 48, pp. 1–16, Jan. 2022.

[18] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3595–3603.

[19] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–9.

[20] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.

[21] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" 2021, *arXiv:2105.14491*.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[23] T. S. Kim and A. Reiter, "Interpretable 3D human action analysis with temporal convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 20–28.

[24] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, Aug. 2017.

[25] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1963–1978, Aug. 2019.

[26] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 499–508.

[27] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2117–2126.

[28] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1227–1236.

[29] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 816–833.

[30] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention LSTM networks," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, Apr. 2018.

[31] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1110–1118.

[32] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[33] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," 2015, *arXiv:1506.05163*.

[34] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," 2013, *arXiv:1312.6203*.
[35] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12026–12035.
[36] Y. Xie, Y. Zhang, and F. Ren, "Temporal-enhanced graph convolution network for skeleton-based action recognition," *IET Comput. Vis.*, vol. 16, no. 3, pp. 266–279, Apr. 2022.
[37] K. Thakkar and P. J. Narayanan, "Part-based graph convolutional network for action recognition," 2018, *arXiv:1809.04983*.
[38] J. Xie, Q. Miao, R. Liu, W. Xin, L. Tang, S. Zhong, and X. Gao, "Attention adjacency matrix based graph convolutional networks for skeleton-based action recognition," *Neurocomputing*, vol. 440, pp. 230–239, Jun. 2021.
[39] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Comput. Vis. Image Understand.*, vols. 208–209, Jul. 2021, Art. no. 103219.
[40] W. Peng, X. Hong, and G. Zhao, "Tripool: Graph triplet pooling for 3D skeleton-based action recognition," *Pattern Recognit.*, vol. 115, Jul. 2021, Art. no. 107921.
[41] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," 2016, *arXiv:1611.01603*.
[42] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[43] C. Li, C. Xie, B. Zhang, J. Han, X. Zhen, and J. Chen, "Memory attention networks for skeleton-based action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4800–4814, Sep. 2022.
[44] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 1–8.
[45] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.
[46] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.
[47] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.
[48] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.
[49] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.
[50] L. Zehui, P. Liu, L. Huang, J. Chen, X. Qiu, and X. Huang, "DropAttention: A regularization method for fully-connected self-attention networks," 2019, *arXiv:1907.11065*.
[51] S. Cho, M. H. Maqbool, F. Liu, and H. Foroosh, "Self-attention network for skeleton-based human action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 635–644.
[52] Y.-H. Wen, L. Gao, H. Fu, F.-L. Zhang, and S. Xia, "Graph CNNS with motif and variable temporal block for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8989–8996.
[53] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 183–192.
[54] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Decoupled spatial–temporal attention network for skeleton-based action-gesture recognition," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 1–16.
[55] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning clip representations for skeleton-based 3D action recognition," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2842–2855, Jun. 2018.

MRUGENDRASINH RAHEVAR received the B.E.C.E. degree from Gujarat University, in 2006, and the M.E.C.S.E. degree from the Gujarat Technological University, in 2014. He is currently pursuing the Ph.D. degree in action recognition in computer vision. He is an Assistant Professor with the U & P U Patel Department of Computer Engineering, Charotar University of Science and Technology (CHARUSAT), Gujarat, India. His research interests include computer vision and deep learning.

AMIT GANATRA (Member, IEEE) served as a Provost for Parul University, Waghodia, Gujarat, India. His record of excellence has allowed him to not only establish his mark in academics through developing policies, designing curriculums, and furthering research. He is known for his contributions to the academic knowledge bank through his authorship of more than 130 research publications and his supervision of over 100 industry projects, more than 100 dissertations, and 12 Ph.D. research scholars. He holds the memberships across multiple academic bodies, where he is instrumental toward shaping the trajectory of technical education. He is a member of ACM and CSI Professional Society Chapters.

TANZILA SABA (Senior Member, IEEE) received the Ph.D. degree in document information security and management from the Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia, in 2012. Currently, she is an Associate Chair with the Information Systems Department, College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia, where she is the Leader of the Artificial Intelligence and Data Analytics Research Laboratory. Her primary research interests include medical imaging, pattern recognition, data mining, MRI analysis, and soft-computing. She is an Active Professional Member of ACM, AIS, and IAENG organizations. She is the PSU Women in Data Science (WiDS) Ambassador at Stanford University and the Global Women Tech Conference. She was a recipient of the Best Student Award from the Faculty of Computing, UTM, in 2012.

AMJAD REHMAN (Senior Member, IEEE) received the Ph.D. and Postdoctoral degrees (Hons.) from the Faculty of Computing, Universiti Teknologi Malaysia, with a specialization in forensic documents analysis and security, in 2010 and 2011, respectively. He is a Senior Researcher with the Artificial Intelligence and Data Analytics Laboratory, College of Computer and Information Sciences (CCIS), Prince Sultan University, Riyadh, Saudi Arabia. He is the author of more than 200 ISI journal articles and conferences. Currently, he is a PI in several funded projects and also completed projects funded from MOHE Malaysia and Saudi Arabia. His research interests include data mining, health informatics, and pattern recognition. He received the Rector Award for the 2010 Best Student from Universiti Teknologi Malaysia.

SAEED ALI BAHAJ (Member, IEEE) received the doctoral degree from Pune University, India, in 2006. He is an Associate Professor with Hadhramaut University, Hadhramaut, Yemen. He is also an Associate Professor with the MIS Department, CBA, PSAU, Alkharj, Saudi Arabia. His main research interests include artificial intelligence, information management, forecasting, information engineering, big data mining, and information security.