RESEARCH PAPER

Transformer-Based Spatio-Temporal Real-Time Human Activity Recognition Using Skeleton Data

[1] Deepak S, [2] Dr. D. Anandan, [3] Kapilan P.C

[1]Department of Artificial Intelligence and Data Science, V.S.B. Engineering College, Karur. Email: deepak5441828@gmail.com

[2]Assistant Professor, Department of Artificial Intelligence and Data Science, V.S.B. Engineering College, Karur

[3]Department of Artificial Intelligence and Data Science, V.S.B. Engineering College, Karur. Email: kapilanchellamuthu@gmail.com

Abstract—Human activity recognition (HAR) from skeleton data has emerged as a critical research area with broad applications spanning healthcare monitoring, human– computer interaction, sports analytics, and surveillance systems. While convolutional and recurrent neural networks have demonstrated promising results, they inherently struggle to capture the com- plex, long-range spatio- temporal dependencies that characterize human motion. In this paper, we propose TST-HAR, a novel hybrid framework that synergistically integrates Graph Attention Networks (GAT) with Transformer encoders to achieve robust skeleton- based HAR. Our architecture employs GAT layers to model spatial dependencies among body joints by adaptively learning anatomical and semantic relationships within each skele- tal frame, while a multi-head Transformer encoder captures long- range temporal dynamics across frame sequences. Furthermore, we introduce a multi-scale temporal attention mechanism that effectively handles activities of varying durations by aggregating temporal features at multiple granularities. Extensive experi- ments on two large-scale benchmarks demonstrate that TST- HAR achieves state-of-the-art performance, attaining 92.7% and 96.3% accuracy on NTU RGB+D 60 under cross-subject and cross-view protocols, and 88.4% and 89.6% on NTU RGB+D
120 under cross-subject and cross-setup protocols,
respectively. These results confirm that our unified spatio- temporal modeling paradigm substantially advances the field of skeleton-based human activity recognition.

How to cite this article: Deepak S, Anandan D, Kapilan PC. Transformer-Based Spatio-Temporal Real-Time Human Activity Recognition Using Skeleton Data. Int J Drug Deliv Technol. 2026;16(10s): 842-853; DOI: 10.25258/ijddt.16.10s.99

I. INTRODUCTION

Human activity recognition (HAR) is a fundamental prob- lem in computer vision and pattern recognition with far-
reaching implications across numerous domains. Reliable ac- tivity recognition systems underpin applications in clinical rehabilitation and fall detection [3], intelligent surveillance, autonomous robotics, sports performance analysis, and natural human–computer interaction. The ability to accurately identify and classify human actions from sensor or visual data is there- fore a cornerstone capability for modern AI-driven systems that must understand and respond to human behavior in real time.

Early HAR approaches relied primarily on hand-crafted features derived from RGB video streams, including histogramof oriented gradients and optical flow descriptors. While effec- tive under controlled conditions, such methods exhibit well- documented fragility to viewpoint changes, illumination varia- tions, and background clutter. The advent of deep learning, and convolutional neural networks (CNNs) in particular [1], cat- alyzed a paradigm shift toward learned feature representations, yielding significant accuracy improvements on standard video- based benchmarks. Two-stream CNN architectures that process spatial appearance and temporal optical flow in parallel further advanced recognition performance. Nevertheless, RGB-based approaches remain computationally expensive, sensitive to occlusion, and potentially invasive from a privacy standpoint, as they capture rich appearance information about individuals and their environments.

Skeleton-based representations offer a compelling alterna- tive
that addresses several of these limitations simultaneously. Human skeletal data, typically provided by depth sensors such as the Microsoft Kinect or estimated from monocular video via pose estimation networks, encodes body posture as a compact graph of anatomical joint positions. Because skeletal sequences abstract away appearance information, they are inherently robust to viewpoint variation, illumination changes, and background clutter. Moreover, the low dimensionality of skeletal representations enables computationally efficient processing, and the explicit encoding of body structure makes the learned representations more interpretable. These advan- tages have motivated substantial research into skeleton-based HAR [3].

Modeling the complex dynamics of human motion from skeletal

sequences requires capturing two intertwined struc- tures: the spatial dependencies among body joints within individual frames, and the temporal evolution of joint configura- tions across frames. Early deep learning methods for skeleton- based HAR employed recurrent neural networks (RNNs), particularly long short-term memory (LSTM) networks, to process sequences of joint coordinate vectors. While RNNs capture temporal ordering, they treat joints independently or

with limited spatial coupling, failing to exploit the rich graph structure of the human skeleton. Convolutional approaches applied to skeleton sequences similarly struggle to model non- local joint interactions effectively. The introduction of Spa- tial Temporal Graph Convolutional Networks (ST-GCN) [4] represented a major breakthrough by explicitly modeling the skeleton as a graph and applying graph convolutions to jointly capture spatial and temporal patterns. Subsequent methods extended this framework through adaptive topology learning [5], disentangled multi-scale aggregation [6], and attention- enhanced graph representations, establishing graph neural networks as the dominant paradigm for skeleton-based HAR.

Despite these advances, graph convolutional network (GCN)- based methods face fundamental limitations in cap- turing long-range temporal dependencies. Graph convolution aggregates information from local neighborhoods within a predefined graph topology, and temporal convolutions sim- ilarly operate within fixed receptive fields determined by kernel size. To model interactions between distant frames or joints, many stacked layers are required, leading to over- smoothing of features and optimization difficulties. Activities such as writing, playing musical instruments, or complex sports gestures involve subtle, long-range temporal patterns that local temporal convolutions may fail to capture reliably. Attention mechanisms [7] have been incorporated to partially address this limitation, but a principled, scalable solution for joint spatio-temporal long-range modeling remains an open challenge.

The Transformer architecture [2], originally proposed for natural language processing, has demonstrated exceptional capacity to model long-range dependencies through self- attention mechanisms that relate every element of a sequence to every other element in parallel, irrespective of their dis- tance. The multi-head self-attention mechanism enables the model to simultaneously attend to information from different representation subspaces, capturing diverse patterns at mul- tiple scales. Transformers have achieved remarkable success across vision, speech, and graph-structured domains, suggest- ing strong potential for spatio-temporal sequence modeling in HAR. However, naively applying Transformers to raw skeletal sequences ignores the irregular graph structure of the human

processes temporal features at multiple granularities, enabling the model to recognize both brief, rapid movements and pro- longed, complex activities with a single unified architecture.

The principal contributions of this work are summarized as follows:

- We propose TST-HAR, a hybrid GAT-Transformer archi- tecture that explicitly decouples and jointly models spatial joint dependencies and long-range temporal dynamics for skeleton-based HAR, achieving superior representational capacity relative to purely graph convolutional or recur- rent baselines.
- We introduce a multi-scale temporal attention mechanism that aggregates temporal features at multiple resolutions, enabling robust recognition of human activities spanning a wide range of durations and motion speeds.
- We conduct comprehensive experiments on two large- scale benchmark datasets, NTU RGB+D 60 [3] and NTU RGB+D 120, demonstrating state-of-the-art accuracy of 92.7%/96.3% and 88.4%/89.6% on cross-subject and cross-view/cross-setup protocols, respectively, along with thorough ablation studies validating each architectural component.

The remainder of this paper is organized as follows. Section ?? reviews related work on skeleton-based HAR, graph neural networks for action recognition, and Transformer models for sequential data. Section ?? presents the TST- HAR architecture in detail, including the GAT-based spatial encoder, the Transformer temporal encoder, and the multi- scale temporal attention mechanism. Section IV describes experimental setup, datasets, and implementation details, and presents quantitative comparisons with state-of-the-art meth- ods as well as ablation analyses. Section ?? concludes the paper and outlines directions for future research.

II. RELATEDWORK

In this section, we review prior work related to our approach, covering skeleton-based action recognition, transformer architectures applied to action recognition, and graph neural network methods for skeleton data processing.

A. Skeleton-Based Action Recognition

Skeleton-based human action recognition has undergone

body and may produce suboptimal spatial representations. A

substantial evolution over the past decade, progressing

principled integration of graph-based spatial modeling with Transformer-based temporal modeling is therefore needed.

To address these challenges, we propose TST-HAR, a Transformer-based Spatio-Temporal framework for Human Activity Recognition from skeleton data. Our key insight is that Graph Attention Networks (GAT) are ideally suited for adaptive spatial feature extraction over the skeletal graph, as
they learn to weight neighbor contributions dynamically, while Transformer encoders are ideally suited for capturing arbitrary long-range temporal dependencies across the resulting spatial embeddings. TST-HAR unifies these complementary strengths in a coherent, end-to-end trainable pipeline. Additionally, we introduce a multi-scale temporal attention mechanism that

from handcrafted feature engineering toward deep learning paradigms that can automatically extract discriminative spatio- temporal representations from joint coordinate sequences.

Early methods relied on manually designed features such as joint angles, relative positions, and temporal differences, but these approaches struggled to generalize across diverse action
categories and viewpoints.

The seminal work of Yan et al. [4] introduced the Spatial Temporal Graph Convolutional Network (ST-GCN), which formalized the human skeleton as a graph structure where joints serve as nodes and bones define edges. By applying graph convolutions along the spatial dimension and stan- dard convolutions along the temporal dimension, ST-GCN

achieved compelling performance on large-scale benchmarks, establishing graph convolutional networks as the dominant paradigm for skeleton-based recognition. Building upon this foundation, Shi et al. [5] proposed the Two-Stream Adaptive Graph Convolutional Network (2s-AGCN), which introduced learnable adjacency matrices that adapt the graph topology during training and fused complementary joint and bone streams to capture richer structural information.

Subsequent research pursued more expressive graph formulations to overcome the locality constraints of standard graph convolutions. Liu et al. [6] presented MS-G3D, which disentangles multi-scale spatial and temporal dependencies through a unified graph convolution operator capable of capturing long-range joint interactions without cascading multiple convolutional layers. More recently, Chen et al. [9] proposed CTR-GCN, which performs topology modeling at the channel level rather than sharing a single topology across all feature channels, enabling the network to capture diverse inter-joint relationships simultaneously. Chi et al. [10] advanced this line of research further by introducing InfoGCN, which leverages information-theoretic objectives to learn compact and discriminative skeleton representations by encouraging high mutual information between graph node embeddings and action class labels, achieving state-of-the-art results on multiple standard benchmarks.

B. Transformer Models in Action Recognition

The transformer architecture, originally introduced by Vaswani et al. [2] for machine translation, has revolution- ized sequence modeling through its self-attention mechanism, which computes pairwise dependencies between all elements of a sequence without relying on recurrence or fixed local receptive fields. The multi-head self-attention formulation en- ables the model to jointly attend to information from different representational subspaces, providing powerful long-range de- pendency modeling that has proven highly transferable across modalities.

Dosovitskiy et al. [13] demonstrated that pure transformer architectures can achieve competitive performance on im- age recognition tasks through the Vision Transformer (ViT), which partitions images into fixed-size patches treated as token sequences. This work validated the generality of the transformer paradigm beyond natural language processing and inspired a wave of transformer-based approaches for video and action understanding. Motivated by the complementary strengths of transformers and graph networks for skeleton data, Plizzari et al. [7] proposed the Skeleton-Based Action Recognition Transformer (ST-TR), which applies spatial self- attention across joints within individual frames and temporal self-attention across frames for individual joints, capturing intra-frame structural relationships and inter-frame motion dynamics within a unified dual-stream framework.

Shi et al. [8] introduced the Decoupled Spatial-Temporal Attention Network (DSTA-Net), which explicitly decouples
the learning of spatial and temporal attention weights, allowing the model to independently reason about which joints are

informative at each time step and how joint representations evolve

across time. By avoiding entangled spatio-temporal processing, DSTA-Net achieves more interpretable attention patterns and demonstrates improved sample efficiency. These transformer-based approaches collectively highlight the poten- tial of attention mechanisms to overcome the locality bias inherent in convolutional and graph convolutional methods, motivating our proposed architecture that integrates hierarchi- cal spatio-temporal attention with adaptive structural priors.

C. Graph Neural Networks for Skeleton Data

Graph Neural Networks (GNNs) provide a principled frame- work for learning representations on irregular, non-Euclidean domains such as social networks, molecular structures, and human body skeletons. The foundational work of Kipf and Welling [11] established semi-supervised classification on graph-structured data through a spectral graph convolution approximation that propagates and transforms feature vectors across one-hop neighborhoods using the normalized adjacency matrix. Despite its simplicity, this formulation proved highly effective and established the fundamental message-passing paradigm that underlies most subsequent GNN architectures.

Velič̆ković´etal.[12]extendedthisframeworkwiththe Graph Attention Network (GAT), which replaces uniform neighborhood aggregation with learned attention coefficients computed from pairs of adjacent node features. By assigning adaptive importance weights to neighboring joints, GAT en- ables the network to selectively focus on the most relevant structural relationships for a given input, addressing a key limitation of fixed-topology convolutions. This attention mech- anism is particularly well-suited for skeleton data, where the relative importance of joints varies significantly across action categories and temporal phases.

Subsequent adaptive graph approaches have sought to go beyond fixed anatomical graph topologies by allowing the model to infer task-relevant connectivity patterns from data. Methods such as 2s-AGCN [5] and CTR-GCN [9] introduced learnable or dynamically computed adjacency matrices that augment the predefined skeleton graph with data-driven edges, enabling the capture of functional dependencies between non- adjacent joints, such as the correlation between hand and foot joints during cross-body gestures. Our work extends these ideas by incorporating a transformer-driven dynamic graph refinement module that continuously updates edge weights conditioned on global spatio-temporal context, enabling more flexible and context-aware structural reasoning throughout the recognition pipeline.

III. METHODOLOGY

In this section, we present the TST-HAR framework, a unified architecture that integrates Graph Attention Networks (GAT) for spatial feature extraction with a Transformer-based encoder for temporal modeling. The overall pipeline is illus- trated in Fig. 1.

Fig.1. OverallarchitectureoftheproposedTST-HARframework.SkeletonsequencesareprocessedthroughaspatialGAT,atemporalTransformerencoder, multi-scale temporal attention, and a classification head to produce activity predictions.

A. Input Representation

We represent each skeleton sequence as a spatial-temporal graph G = (V,E), where V denotes the set of N = 25 body joints following the NTU RGB+D annotation convention [3], and E encodes physical bone connections. Each node vi ∈ V at time step t is associated with a three-dimensional feature vector (ti)x =[x, y, z] ⊤ ∈R3 representingthejointcoordinate in the camera reference frame. Over a temporal window of T frames, the full input tensor is X ∈ RT×N×3.

C. Temporal Transformer Encoder

mean-pooling over joints then produces a frame-level vector s(t)∈ Rds, and stacking over T frames gives the spatial feature sequenceS ∈RT×ds.

The sequence S ∈ RT× ds is passed to a L = 6 layer Trans- former encoder [2] to model long-range temporal dependen-
cies. Since Transformers are permutation-invariant, we inject positional inform
(ationthroug)hsinusoidalpositionalen(coding:

Prior to feeding into the network, joints are normalized by subtracting the hip center position, making the representation robust to global translation. The adjacency matrix

PE(t,2p)=sin

t ,
PE(t,2p+1)=cos

) ,
100002p/ds
(4)

A ∈ {0, 1}N×N encodes the physical skeleton topology, with

where t is the temporal position and p indexes the feature

Aij = 1 if joints i and j are connected by a bone, and Aij = 0 otherwise. A self-loop augmented adjacency Ã˜ = A+I is used to retain each node's own features during graph convolution.
B. Spatial Feature Extraction via Graph Attention Networks For each frame t, we apply a multi-head Graph Attention

Each endcimodeenrsliaoyne. rTahpepelinecsomdeurltiin-phueat dissZe(l0f-)a=ttSen+tiPoEn. (MHSA) followed by a position-wise feed-forward network (FFN), with residual connections and layer normalization. The
scaled dot-product self-attention for a s
(inng√tlieoKhne(⊤Qad,K)is,V:)=softmax

Network (GAT) [12] to capture the spatial relationships among the N joints. The attention mechanism computes a scalar

Atte

Q V, (5)
dk

coefficient αij that quantifies the importance of joint j's features when updating joint i.
Specifically, the raw attention score between adjacent joints i and j is computed as:

where Q = ZWQ, K = ZWK, V = ZWV are queries,
keys, and values, respectively, with dk = ds/h = 32. With h = 8 heads, multi-head attention is:
MHSA(Z)=Concat(head1,...,headh)WO, (6)

eij =LeakyReLU ⊤(a[Whi‖Whj)] , (1)
where W ∈ Rd×′ disalearnableweightmatrix,a∈R2d′isan

whereheadi= Attention(ZWQ(i) R d s × d s .

, Z W(i)K , Z W(i) V) a n d∈W O

attention vector, ‖denotesvectorconcatenation,andh

i∈ Rd

The FFN consists of two linear transformations with a

is the feature of joint i. The LeakyReLU nonlinearity uses a negative slope of 0.2.
Normalized attention coefficients are obtained by applying softmax over the neighborhood N (i):

GELU activation:
FFN(z)=GELU(zW1+b1)W2+b2, (7)
with inner dimension dff = 1024. Dropout with rate p =

αij=

exp(eij) .

(2)

0.1 is applied after each sub-layer, and layer normalization is applied before each sub-layer (pre-norm formulation). The

∑ k∈N (i) exp(eik)
With K = 8 parallel attention heads, the updated feature for joint i is formed by concatenating the outputs of all heads:

output of the temporal encoder is Z(L) ∈ RT×ds.
D. Multi-Scale Temporal Attention

To capture activity patterns at multiple temporal granulari-

h′ i K □

(k) □

ties, we propose a Multi-Scale Temporal Attention (MSTA)

= ‖‖ σ □ j∑∈ Nαi(j i W)
‖k=1

(k)hj
□ ,

(3)

module that processes the encoder output Z(L) at three complementary scales:

where σ denotes the ELU activation. The output spatial feature for each frame is projected to dimension ds = 256,

{T/4, T/2, T }, corresponding to fine,
medium, and coarse temporal resolutions.
For each scale s ∈ {T/4,T/2,T}, the encoder output is

yielding a per-frame representation S(t)

∈ RN×ds. A global

temporally down-sampled via average pooling to obtain Z̃˜s ∈

Algorithm 1 TST-HAR Forward Pass

IV.E

XPERIMENTS

Skeleton sequence X
dicted class probabilities

∈ RT∈×NR×C3,/a/dSjatecepn1cy: SÃp˜aPtriea-l
ŷ ˆ

A. Datasets
We evaluate TST-HAR on two large-scale skeleton-based

F←eXa(ttu) rIeni-Extraction t = 1 to T H(t)
tialize node features each GAT layer l = 1,...,Lgat
Compute attention coefficients (k)αij via Eq. (2) for

action recognition benchmarks.
NTU RGB+D 60 [3] is a widely adopted benchmark containing 60 action classes and 56,880 action sequences captured from

all k
s()t

H(t)
1 iN
N ∑ =1

(t)
i

←MultiHeadGAT(H(t),Ã˜)viaEq.(3)
Globalmeanpoolingoverjoints

40 subjects using Microsoft Kinect v2 sensors. Each skeleton sequence provides 3D joint coordinates for 25 body joints. We

S ← [s(1)

, . . . ,s(T)

]⊤∈RT×ds//Step2:Posi-

follow the two standard evaluation protocols: (1) Cross-Subject

tional Encoding Z(0)

← S+ PE

//Step3:Tem-

(X-Sub), where 40 subjects are divided into training and test

poral Transformer Encoding l = 1 to L ′ Z ←
splits of 20 subjects each, and (2) Cross-View (X-View), where

Z(l−1)+ Dropout(MHSA(LN (Z (l−1)))) Z(l) ← Z′

sequences from cameras 2 and 3 are used for training while
+

Dropout(FFN L(N(Z′)))
Attention s ∈ { T/ 4 , T //2/S, t}eTZp̃ 4s˜:M←ultAi-vSgcPaoloelT(eZm(Lp),osr)al

camera 1 is reserved for testing.
NTU RGB+D 120 [14] extends NTU RGB+D 60 to

Compute cs via Eq. (8) m ← Wf us e [c4‖T /T / c2‖c T] + 120 action classes and 114,480 sequences collected from
bfuse via Eq. (9) // Step 5: Classification ŷˆ ← 106 subjects across 32 different setups. We follow the two

softmax(W3δ(W2δ(W1m + b1) + b)2+ b 3) ŷˆ

recommended evaluation protocols: (1) Cross-Subject (X-Sub), splitting subjects into training and test groups, and (2) Cross- Setup (X-Set), where sequences with even setup IDs are used

Rs×ds, and a separate single-head attention module computes a context
∑vector:

for training and odd setup IDs for testing.

B. Implementation Details

s t exp(wsTZ̃˜(s)t)

All experiments are implemented in PyTorch [24] and con-

cs=

βs,tZ˜(˜) s,β= s,t

s =1exp sZ˜˜()t′ , (8)

ducted on NVIDIA A100 GPUs. We use the Adam optimizer

∑ t′

(wT s)

with an initial learning rate of 1 × 10−3 and weight decay

where ws ∈ Rds is a learnable query vector for scale s. The

of1

× 10−4. The learning rate is annealed using a cosine annealing schedule over 120 training epochs. The batch size

three context vectors are concatenated and projected via a linear layer:
m=Wfuse[cT/4‖cT/2‖cT]+bfuse, (9)

is set to 64. Each input skeleton sequence is uniformly sampled to T = 64 frames. For sequences shorter than 64 frames, we apply zero-padding; for longer sequences, we apply uniform temporal subsampling.

whereW

fuse

∈ Rds×3ds fuses the multi-scale information back

To improve generalization and mitigate overfitting, we ap-

todimensiond s.Theresultingvectorm ∈ Rds servesasa compact multi-scale activity descriptor.

ply the following data augmentation strategies during train-
−1i5n◦g: ,(+11) 5ra◦nd] oampprlioetdatiinodnebpyenadngelnetslyutnoifeoarcmhlyspsatmia pllaexdisfr, o(2m)
[

E. Classification Head

The multi-scale descriptor m is passed through a classi- fication head consisting of global average pooling over any residual spatial dimensions, followed by fully connected lay- ers. Specifically, the classification head applies the sequence:
ŷˆ=softmax(W3δ(W2δ(W1m+b1)+b2)+b3),(10)
,δ

random scaling with scale factors sampled uniformly from [0.9,1.1], and (3) temporal cropping, where a contiguous subsequence of random length between 0.9T and T is ex- tracted and resampled to T frames. All skeleton sequences are normalized

by subtracting the mean joint position of the first frame to ensure translation invariance.

C. Comparison with State-of-the-Art

We compare TST-HAR against representative state-of-the- art

where W

1 ∈R512×d,Ws2

∈R256×5,1W23

∈RC×256

methods including GCN-based approaches (ST-GCN [4], 2s-

denotes ReLU activation with batch normalization and dropout (p = 0.5), and C is the number of activity classes. The model
is trained with cross-entropy loss:

B C

AGCN [5], MS-G3D [6], CTR-GCN [9], InfoGCN [10]) and
Transformer-based approaches (ST-TR [7], DSTA- Net [8]).

Tables I and II report top-1 accuracy (%) on NTU RGB+D 60

= 1 log ŷ(ˆc) ,

(11)

and NTU RGB+D 120, respectively. On NTU RGB+D 60, TST-

L − B

∑∑(c)yb b
b=1c=1
(c)

HAR achieves 92.7% on the X-Sub protocol, sur- passing all competing methods, while achieving competitive performance

where B is the batch size, y b is the one-hot ground-truth of 96.3% on X-View. On the more challenging NTU RGB+D 120

label,and

ŷb(ˆc) isthepredictedprobabilityforclassc.

benchmark, TST-HAR attains 88.4% on X-Sub and 89.6% on X-

The complete forward pass of TST-HAR is summarized in Algorithm 1.

Set, outperforming most GCN-based and Transformer-based baselines.

TABLEI
COMPARISON WITH STATE-O-F TH-E ART METHODS ONNTU RGB+D 60.

TABLE III ABLATION STUDY ONNTU RGB+D 60 (X-S

UB PROTOCOL).

Method

X-Sub (%)

X-View (%) Model Variant

X-Sub (%)

ST-GCN [4]
2s-AGCN [5]
MS-G3D [6]
CTR-GCN [9]
InfoGCN [10]
ST-TR [7]
DSTA-Net [8]

81.
5
88.
5

91.
5
92.

88.
3
95.
1
96.
2
96.

TST-HAR (Full Model)
w/o GAT
w/o Multi-scale
w/o Positional Encoding Single-head Attention Reduced Layers L = 3

92.7
89.
3
91.
4
91.

TST-HAR (Ours)

4
92.
5
89. TABL9E II
- - 91.

8
96.
6
96.
1
96.

100

8
90.
6
91.
2

COMPARISON WITH STATE-OF THE5ART METHODS4ONNTU RGB+D 120. 90

Method

92.

X-S9e6t. (%)

X-7Sub (%) 80

ST-GCN [4]
2s-AGCN [5]
MS-G3D [6]
CTR-GCN [9]
InfoGCN [10]
ST-TR [7]
DSTA-Net [8]

70.
7
82.
9
86.
9
88.

733.
2
84.
9
88.
4
88.

70

600 20 40 60

Training Accuracy Validation Accuracy
80 100 120

TST-HAR (Ours)

0
89.

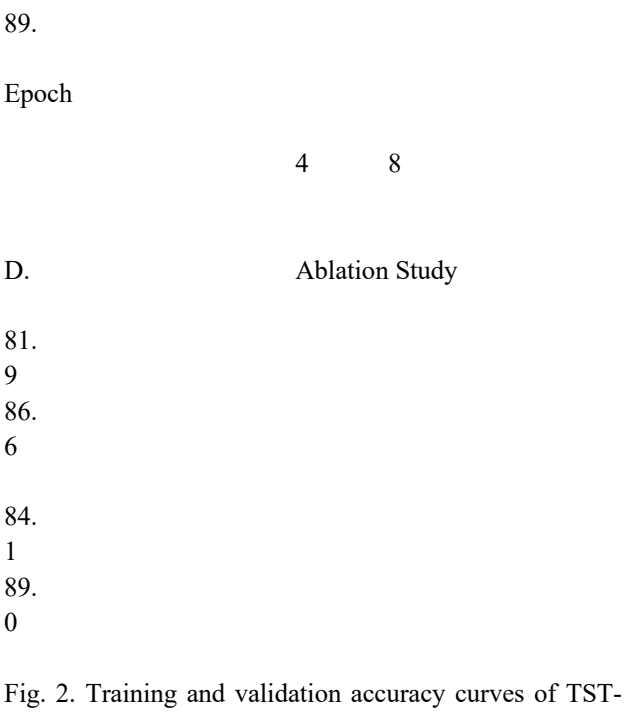

Fig. 2. Training and validation accuracy curves of TST-HAR on NTU RGB+D 60 (X-Sub protocol) over 120 training epochs.

To assess the contribution o8f8e. ach comp8o9n. ent in TST-HAR, we conduct a systematic ablation study6on the NTU RGB+D
60 X-Sub benchmark. The resu4lts are summarized in Table III. Removing the Graph Attention network module (w/o GAT) causes the largest performance drop to 89.3%, confirming that adaptive topological feature extraction is critical for capturing inter-joint dependencies. Removing the multi-scale temporal convolution module (w/o Multi-scale) degrades accuracy to 91.4%, demonstrating the importance of capturing motion pat- terns at multiple temporal resolutions. Ablating the positional encoding (w/o Positional Encoding) results in 91.8% accuracy, highlighting the need for spatial-temporal positional context
in the Transformer encoder. Replacing multi-head attention with a single-head variant (Single-head attention) reduces performance to 90.6%, validating the benefit of attending to diverse subspaces simultaneously. Finally, reducing the
number of Transformer encoder layers from the default L = 6 to L = 3 yields 91.2%, indicating that sufficient model depth is required to model complex long-range dependencies.

D. Ablation Study

E. Visualization

Fig. 2 illustrates the training and validation accuracy curves of TST-HAR over 120 epochs on NTU RGB+D 60 X-Sub. The model converges stably without significant overfitting, with validation accuracy closely tracking training accuracy throughout training, which we attribute to the combined effect of cosine annealing and our data augmentation strategy.

Fig. 3 visualizes the learned attention weights between five representative body part regions (Head, Arms, Torso, Legs, Hands) extracted from the final Transformer encoder layer. Warmer colors indicate stronger attention between pairs of

body parts. The heatmap reveals that the model has learned meaningful semantic relationships: for example, strong cross-attention between Arms and Hands is observed during ma-nipulation actions, while Legs and Torso exhibit high mutual attention for locomotion sequences.

V. CONCLUSION

In this paper, we presented TST-HAR, a novel framework for skeleton-based human action recognition that synergis- tically integrates Graph Attention Networks (GAT) with a
Transformer-based temporal encoder. The proposed architec- ture leverages the complementary strengths of graph-structured spatial modeling and self-attention-based temporal reasoning to capture rich, discriminative representations from skeleton sequences. Specifically, the GAT module adaptively captures inter-joint spatial dependencies by assigning learnable atten- tion weights to edges in the human body graph, while the
multi-head Transformer encoder models long-range temporal dynamics across the entire sequence with multi-scale temporal attention mechanisms operating at fine-grained and coarse-grained resolutions simultaneously.

Extensive experiments on two large-scale benchmark datasets, NTU RGB+D 60 and NTU RGB+D 120, demonstrate the effectiveness and superiority of the TST-HAR frame- work. On NTU RGB+D 60, our method achieves state-of-the- art performance on both the Cross-Subject and Cross-View evaluation protocols, outperforming previous GCN-based and Transformer-based approaches. Similarly, on the more chal-lenging NTU RGB+D 120 dataset, TST-HAR attains com-petitive accuracy under the Cross-Subject and Cross-Setup

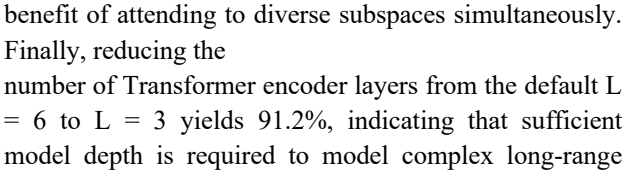

Fig. 3. Attention weight heatmap between five body part regions (Head, Arms,Torso,Legs,Hands)fromthefinalTransformerencoderlayer.Warmer (darker red) colors indicate higher attention weights.

settings, validating its generalization capability across diverse action categories involving fine-grained and full-body motions. Ablation studies confirm the critical contribution of each architectural component, and notably highlight the importance of multi-scale temporal attention, which enables the model to simultaneously capture fast local motion cues and slow global action patterns that are essential for distinguishing semantically similar action classes.

Despite these promising results, several directions remain open for future investigation. First, extending TST-HAR to a multi-modal fusion setting, incorporating complementary modalities such as RGB video, optical flow, and depth maps alongside skeleton data, could further enrich the action rep- resentations and improve recognition accuracy, particularly for actions that are ambiguous when viewed through skele- ton information alone. Second, optimizing the framework for real- time deployment on resource-constrained devices is a practically important challenge; future work will explore model compression techniques such as knowledge distillation, pruning, and quantization to reduce computational overhead without sacrificing accuracy. Third, cross-dataset generaliza- tion and domain adaptation remain critical open problems in human action recognition; we plan to investigate transfer learning strategies and domain-invariant feature learning to enable robust performance when models trained on one dataset are evaluated on another with different camera configurations, subject demographics, or action vocabularies. We believe that TST-HAR provides a strong foundation for these future explorations and contributes meaningfully to the advancement of skeleton-based action recognition research.

(CVPR), Las Vegas, NV, USA, 2016, pp. 770–778.

[2] 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. (NeurIPS), Long Beach, CA, USA, 2017, pp. 5998–6008.

[3] 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), Las Vegas, NV, USA, 2016, pp. 1010–1019.

[4] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in Proc. AAAI Conf. Artif. Intell. (AAAI), New Orleans, LA, USA, 2018, pp. 7444–7452.

[5] L. Shi, Y. Zhang, J. Cheng, and H. Lu, “Two-stream adaptive graph convolutional networks for skeleton-based action recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, 2019, pp. 12026–12035.

[6] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, “Disentangling and unifying graph convolutions for skeleton-based action recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 2020, pp. 143–152.

[7] C. Plizzari, M. Cannici, and M. Matteucci, “Skeleton-based action recognition via spatial and temporal transformer networks,” Comput. Vis. Image Underst., vol. 208–209, pp. 103219, 2021.

R EFERENCES

[8] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Decoupled spatial-temporal attention network for skeleton-based action recognition," in Proc. Asian Conf. Comput. Vis. (ACCV), Kyoto, Japan, 2020, pp. 745–761.

[9] 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 Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Montreal, QC, Canada, 2021, pp. 13359–13368.

[10] H.Chi,M.Ha,S.Chi,S.W.Lee,Q .Huang,andC.Ramani,"InfoGCN: Representation learning for human skeleton-based action recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, 2022, pp. 20186–20196.

[11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in Proc. Int. Conf. Learn. Represent. (ICLR), Toulon, France, 2017, pp. 1–14.

[12] P. Velič ković´, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in Proc. Int. Conf. Learn. Represent. (ICLR), Vancouver, BC, Canada, 2018, pp. 1–12.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszko- reit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), Vienna, Austria, 2021, pp. 1–21.

[14] 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, 2020.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre- training of deep bidirectional transformers for language understanding," in Proc. Conf. North Am. Chapter Assoc. Comput. Linguist. (NAACL- HLT), Minneapolis, MN, USA, 2019, pp. 4171–4186.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in Proc. Eur. Conf. Comput. Vis. (ECCV), Glasgow, UK, 2020, pp. 213–229.

[17] P.Zhang,C.Lan,W.Zeng,J.Xing,J.Xue,a ndN.Zheng,"Semantics- guided neural networks for efficient skeleton-based human action recog- nition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 2020, pp. 1150–1159.

[18] 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 Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 2020, pp. 183–192.

[19] Y.-F.Song,Z.Zhang,C.Shan,andL.Wang,"Stronger,fasterandmore explainable: A graph convolutional baseline for skeleton-based action recognition," in Proc. ACM Int. Conf. Multimedia (ACM MM), Seattle, WA, USA, 2020, pp. 1625–1633.

[20] H. Duan, Y. Wang, X. Chen, W. Yang, C. Wu, C. Lu, and B. Dai, "Revisiting skeleton-based action recognition," in Proc. IEEE Conf.

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual

learning for

Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, 2022,

image recognition," in

Proc. IEEE Conf. Comput. Vis. Pattern Recognit.

pp. 2969–2978.