ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1362

# Enhanced HAR using Dynamic STGAT

Pragati Narote [1] ; Shrayanshi [2]; Priyanka S Chauhan [3]; Vaddempudi Charan Teja [4] ; Ponnaganti Karthik [5]
[1] Software Engineer, Mastercard, Pune, India
[2] Software Engineer, Rapipay Fintech Private Limited, Noida, India
[3] SDET, Engineering Manager, Global Payments, Atlanta, United States
[4,5] B.Tech Graduate (Cse AI ML) GITAM University, Bengaluru, India

**Abstract:- Action recognition has seen significant advancements with the integration of spatio-temporal representations, particularly leveraging skeleton-based models and cross-modal data fusion techniques. However, existing approaches face challenges in capturing long-range dependencies within the human body skeleton and effectively balancing features from diverse modalities. To address these limitations, a novel framework, the Dynamic Spatio-Temporal Graph Attention Transformer (D-STGAT), is proposed, which seamlessly integrates the strengths of dynamic graph attention mechanisms and transformer architectures for enhanced action recognition. The framework builds upon recent innovations in graph attention networks (GAT) and transformer models. First, the Spatial-Temporal Dynamic Graph Attention Network (ST-DGAT) is introduced, extending traditional GAT by incorporating a dynamic attention mechanism to capture spatial-temporal patterns within skeleton sequences. By reordering the weighted vector operations in GAT, the approach achieves a global approximate attention function, significantly enhancing its expressivity and capturing long-distance dependencies more effectively than static attention mechanisms. Furthermore, to address the challenges of cross-modal feature representation and fusion, the spatio-temporal Cross Attention Transformer (ST-CAT) is introduced. This model efficiently integrates spatio-temporal information from both video frames and skeleton sequences by employing a combination of full spatio-temporal attention (FAttn), zigzag spatio-temporal attention (ZAttn), and binary spatio-temporal attention (BAttn) modules. Through the proper arrangement of these modules within the transformer encoder and decoder, ST-CAT learns a multi-feature representation that effectively captures the intricate spatiotemporal dynamics inherent in action recognition tasks. Experimental results on the Penn-Action, NTU-RGB+D 60, and 120 datasets showcase the efficacy of the approach, yielding promising performance improvements over previous state-of-the-art methods. In summary, the proposed D-STGAT and ST-CAT frameworks offer novel solutions for action recognition tasks by leveraging dynamic graph attention mechanisms and transformer architectures to effectively capture and fuse spatiotemporal features from diverse modalities, leading to superior performance compared to existing approaches.**

***Keywords:-*** *Graph Attention Network, Skeleton-Based Models, Dynamic Attention Mechanism, Multi-Feature Learning, Spatial-Temporal Patterns.*

## I. INTRODUCTION

Action recognition, a fundamental task in computer vision, plays a pivotal role in various real-world applications such as sports analysis, human-computer interaction, and surveillance systems. Over the past decade, the advancement of deep learning techniques has revolutionized the field of action recognition, offering powerful tools for extracting complex temporal and spatial features from video data. Over the last decade, there has been significant interest in human action recognition within extended video sequences, with a focus on both spatial and temporal localization of actions. This area of research has gained prominence due to its wide-ranging applications in fields such as sports analysis, human-machine interaction, and intelligent robotics systems, among others. Unlike other modalities such as RGB-D and optical flow, skeleton-based approaches offer computational efficiency and streamlined data storage [1-7]. Additionally, skeleton data is lightweight, and robust against irrelevant objects, body scale variations, and changes in camera viewpoint, making it conducive to efficient prediction [8-13]. However, it's important to note that the human skeleton is inherently structured as a graph, rather than a sequential data format, posing challenges for conventional neural network models like CNN to achieve optimal performance. Recent studies have explored the utilization of entire graphs for processing, employing shared weight vectors across all nodes. However, this approach tends to obscure the significance of individual nodes for different activities, leading to diminished model performance in graph classification tasks. Moreover, it significantly increases computational costs and introduces unwanted noise into the system. In response to these challenges, graph attention networks (GAT) have been introduced to enhance skeleton-based action recognition [14,15,16]. The research delves into an attention mechanism designed to highlight important nodes while suppressing redundant ones within the graph. The key advantage of this attention model lies in its ability to effectively handle inputs of variable sizes. In contrast, traditional GAT models suffer from the drawback of sharing attention scores across all nodes, without being constrained by the query node.

Among the myriad of deep learning architectures, transformers have emerged as particularly promising for action recognition tasks due to their ability to capture long-range temporal dependencies through self-attention mechanisms. However, while transformers have shown remarkable success in processing video frames as input tokens, their application to skeleton-based action recognition remains relatively unexplored. Traditional approaches to action recognition often rely solely on video frames, overlooking valuable information encoded in skeletal data. Incorporating skeleton information into transformer-based models presents an opportunity to leverage both spatial and temporal cues for enhanced action recognition performance. Despite the potential benefits, transformer-based action recognition faces challenges, including high computational costs associated with processing large volumes of 3D tokens in videos and the lack of established methodologies for effectively integrating cross-modal information. In parallel, efforts have been made to improve action recognition by integrating cross-modal information from both video and skeleton data sources. These endeavors aim to harness the complementary nature of different modalities to achieve superior performance compared to using unimodal features alone.

Action recognition, a well-established field, involves categorizing human actions depicted in video frames and finding applications in diverse areas such as human-robot interaction, healthcare, and video surveillance [32]. Recent advancements in deep learning have led to the emergence of three main approaches in action recognition research. Firstly, one approach focuses on utilizing human skeletons and joint trajectories as inputs for deep learning models [33]. However, this method necessitates an additional deep-learning model to extract the human skeleton from an image. Furthermore, the accuracy of the skeleton extractor and the degree of overlap of the skeleton significantly influence the effectiveness of action recognition using this approach [34]. Secondly, another approach involves the integration of cross-modal data, combining video and skeletal information. Here, a deep learning model learns from both the RGB data of video frames and the features extracted from human skeletons [35]. This approach often yields high recognition performance. Nonetheless, the process of combining video and skeleton data is intricate and typically requires a separate sub-model for cross-modal learning.

This research aims to address these challenges by proposing novel approaches for transformer-based action recognition and integrating cross-modal information effectively. Specifically, we explore the utilization of spatio-temporal cross-modal data as input tokens for Vision Transformers (ViTs) without the need for separate sub-models. Additionally, we investigate methods for seamlessly fusing multi-modal information to extract discriminative features for improved action recognition performance. By bridging the gap between transformer-based architectures and cross-modal action recognition, this study seeks to advance the state-of-the-art in action recognition and pave the way for more robust and efficient recognition systems in various real-world applications.

## II. RELATED WORK

In recent years, deep learning methods have made significant strides in visual tasks, with various approaches proposed for action recognition. One prevalent model architecture involves two main approaches. The first architecture utilizes convolutional neural networks (CNNs), where each generated clip is transformed into long-term temporal skeleton sequences [17,18]. These sequences are then processed using convolutional operators applied in parallel across the entire frame sequences, aiming to integrate spatial structural features for action recognition. Additionally, there are efforts to represent skeleton sequences as a series of color images, enabling the use of CNN models for action classification. However, while CNN-based methods excel in capturing spatial information, they often overlook temporal information [19]. Some researchers have attempted to incorporate temporal information using 2D convolutional operations, but this approach tends to be slow and may emphasize irrelevant features, negatively impacting model performance [20,21]. The second architecture, known as Long Short-Term Memory (LSTM), has proven effective in modeling temporal dependencies compared to CNNs. Researchers have proposed spatial-temporal LSTM networks with gating mechanisms to mitigate input inconsistencies caused by occlusions and noise. Another approach involves employing an attention model within a two-tiered LSTM framework, where the first tier records the skeleton patterns and the second tier enriches global context to aid in human action recognition. [22,23,24].

Shi and colleagues proposed a novel approach by combining adaptive graph techniques with a two-stream framework based on ST-GCN. Their method, termed two-stream adaptive GCN (2s-AGCN), aims to enhance action recognition by effectively connecting joint (first-order) and bone (second-order) features. Similarly, another study introduced a directed acyclic graph (DAG) mechanism tailored to extract relational patterns among joints and bones, thereby improving action recognition performance [25,26]. Additionally, researchers applied a part-based GCN to explore the relationship between human gestures using analogous joint coordinates and temporal displacements [27,28]. However, a drawback observed in some models is the emphasis on learning spatial features at the expense of temporal features. In a separate work, Plizzari and colleagues introduced ST-TR, a spatio-temporal transformer network that leverages transformer techniques to determine the self-attention score of each body joint, thus enhancing spatio-temporal feature learning for action recognition [29].

Transformer-based approaches have gained attention in action recognition due to their capability for long-range temporal modeling facilitated by self-attention mechanisms [36]. This has led to a surge in studies exploring the use of transformers in action recognition tasks. While most approaches employ video frames as input tokens, fewer methods utilize the skeleton information within the transformer framework. However, transformer-based action recognition often faces challenges related to high computational costs, particularly due to the self-attention mechanism applied to a large number of 3D tokens in videos [37]. Additionally, there is a lack of established methods for effectively integrating cross-modal information using transformers. This study represents the first attempt to utilize spatio-temporal cross-modal data as input tokens for Vision Transformers (ViTs) without the need for separate sub-models [38,39]. The goal is to achieve high action recognition performance by seamlessly fusing multi-modal information into a cohesive set of discriminative features. In parallel, efforts have been made in video and skeleton-based action recognition to enhance performance by integrating cross-modal information. For instance, the Video-Pose Network (VPN) leverages cross-modal features and knowledge distillation to incorporate pose information into RGB streams, demonstrating improved performance compared to using unimodal features alone [40]. Despite the promising results, existing methods such as VPN still face challenges in designing effective subnetworks for cross-modal learning and devising efficient strategies for combining cross-modal data.

## III. RESEARCH METHODOLOGY

*A. Data Collection and Preprocessing:*

- Gather a diverse dataset containing both video and skeleton data for action recognition tasks.
- Preprocess the data to ensure consistency in format and quality, including normalization, cropping, and augmentation techniques.

*B. Transformer-Based Action Recognition:*

- Implement a Vision Transformer (ViT) architecture for action recognition, treating video frames or skeleton data as input tokens.
- Fine-tune the pre-trained ViT model on the collected dataset using supervised learning techniques.
- Experiment with different configurations of the ViT model, including varying numbers of layers, attention heads, and hidden dimensions.

*C. Integration of Cross-Modal Information:*

- Explore methods for effectively integrating cross-modal information from video and skeleton data sources.
- Investigate the use of fusion techniques such as late fusion, early fusion, or attention mechanisms to combine information from different modalities.
- Design novel architectures or adapt existing transformer-based models to handle cross-modal inputs seamlessly.

*D. Evaluation Metrics:*

- Employ standard evaluation metrics such as accuracy, precision, recall, and F1-score to assess the performance of the proposed models.
- Conduct cross-validation experiments to ensure the robustness and generalization of the models.

*E. Comparative Analysis:*

- Compare the performance of the proposed transformer-based models with baseline models and state-of-the-art methods for action recognition.
- Evaluate computational efficiency, including training time and inference speed, to assess the practical feasibility of the proposed approaches.

*F. Qualitative Analysis:*

- Conduct qualitative analysis by visualizing attention maps or feature representations learned by the models to gain insights into their behavior.
- Interpret the results to understand how the models leverage spatial and temporal information for action recognition.

*G. Implementation and Tools:*

- Implement the proposed methodologies using deep learning frameworks such as TensorFlow or PyTorch.
- Utilize relevant libraries and tools for data preprocessing, model training, and evaluation.

Figure 1 illustrates the architecture of Spatial-Temporal Dynamic Graph Attention Networks (ST-DGATs) along with a detailed depiction of the computation process involved in the Dynamic Graph Attention Network. ST-DGATs are designed to effectively learn spatial-temporal patterns from skeleton sequences, crucial for accurate action recognition. The architecture consists of multiple layers, each housing dynamic graph attention modules tailored for processing skeleton-based input data. Within each layer, the input skeleton sequence is represented as a graph, with joints as nodes and their connections encoding spatial relationships. The Dynamic Graph Attention Network computes attention scores for each joint based on its relationships with other joints, dynamically adjusting focus over time to capture relevant spatial-temporal dependencies. By iteratively refining these attention scores through multiple layers, the network learns to extract discriminative features for action recognition from the input skeleton sequence. Overall, Figure 1 provides a comprehensive visualization of the ST-DGATs architecture and the computational process underlying the Dynamic Graph Attention Network, showcasing its ability to effectively capture spatial-temporal patterns in skeleton-based data.

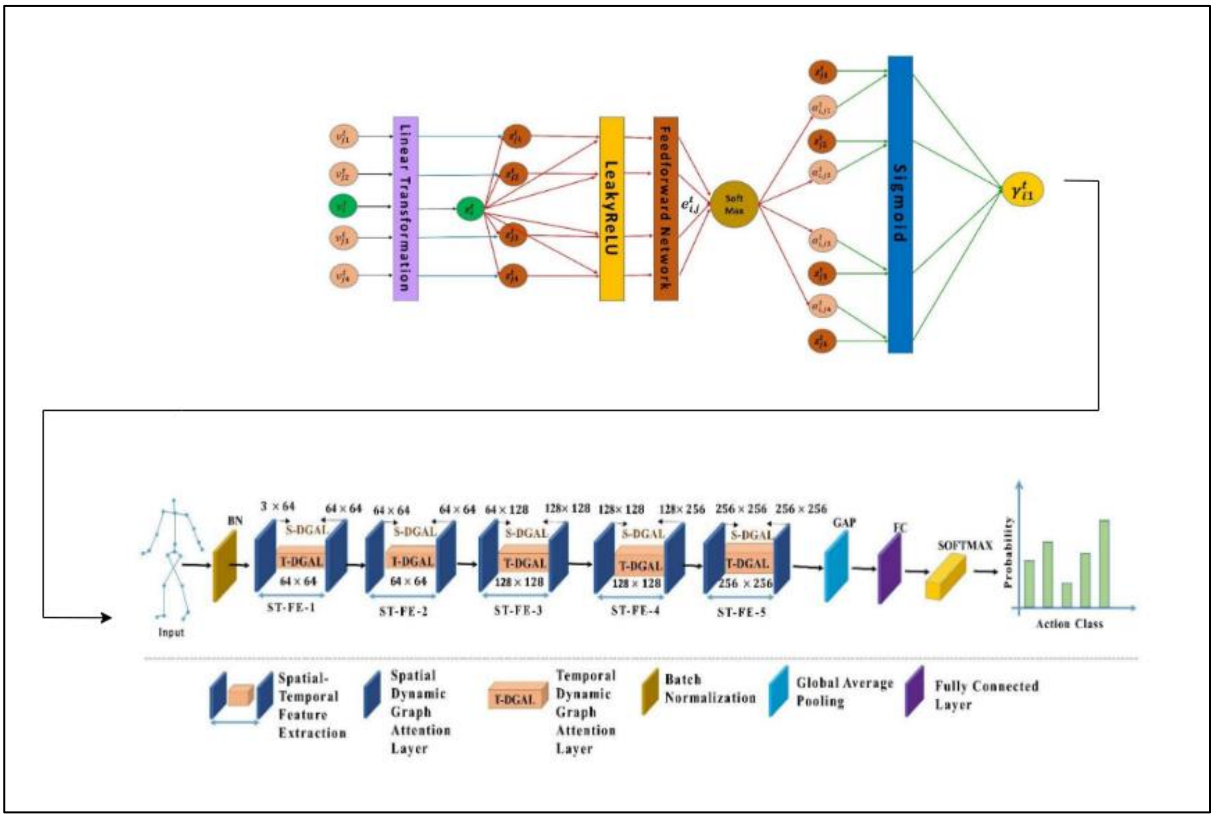

Fig. 1. The Architecture of ST-DGATs & the Illustration of Computing Dynamic Graph Attention Network.

In this illustration, we use a different notation compared to S-DGAL; subscripts indicate time, while superscripts represent joints. The T-DGAL module focuses on capturing temporal dynamics for each joint over consecutive frames. It achieves this by computing correlations across frames, adjusting the embeddings of corresponding joints along the temporal dimension. Notably, the formulation of T-DGAL mirrors that of S-DGAL, ensuring symmetry in their operations and enhancing their compatibility within the model architecture.

$$\gamma_t^{\mu} = \sum_{k \in F(v,u)} \quad (\sigma . \alpha_k^t) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Figure 2 presents the comprehensive architecture of the proposed action recognition model, detailing its key components and their roles in the recognition process. Firstly, the global grid token (a) represents the overarching spatial context of the input video frames, capturing holistic information about the scene. Concurrently, the joint heat map (b) visualizes the spatial distribution of skeletal joints within the video frames, offering insights into body pose and movement. The joint map token (c) serves as a condensed representation of skeletal information, encoding spatial relationships and configurations of joints. These tokens are then fed into the STAR-Transformer module (d), the central component of the model. Leveraging transformer architecture, this module processes both global and joint-level information, capturing spatial-temporal features crucial for action recognition through attention mechanisms. Lastly, the encoder and decoder structures (e) of the STAR transformer work collaboratively to encode input data into meaningful feature representations and generate predictions based on these encoded features. Both the encoder and decoder share identical architectures, ensuring consistent handling of spatial and temporal information throughout the model. Together, these components enable the model to effectively integrate global and skeletal data for accurate action recognition.

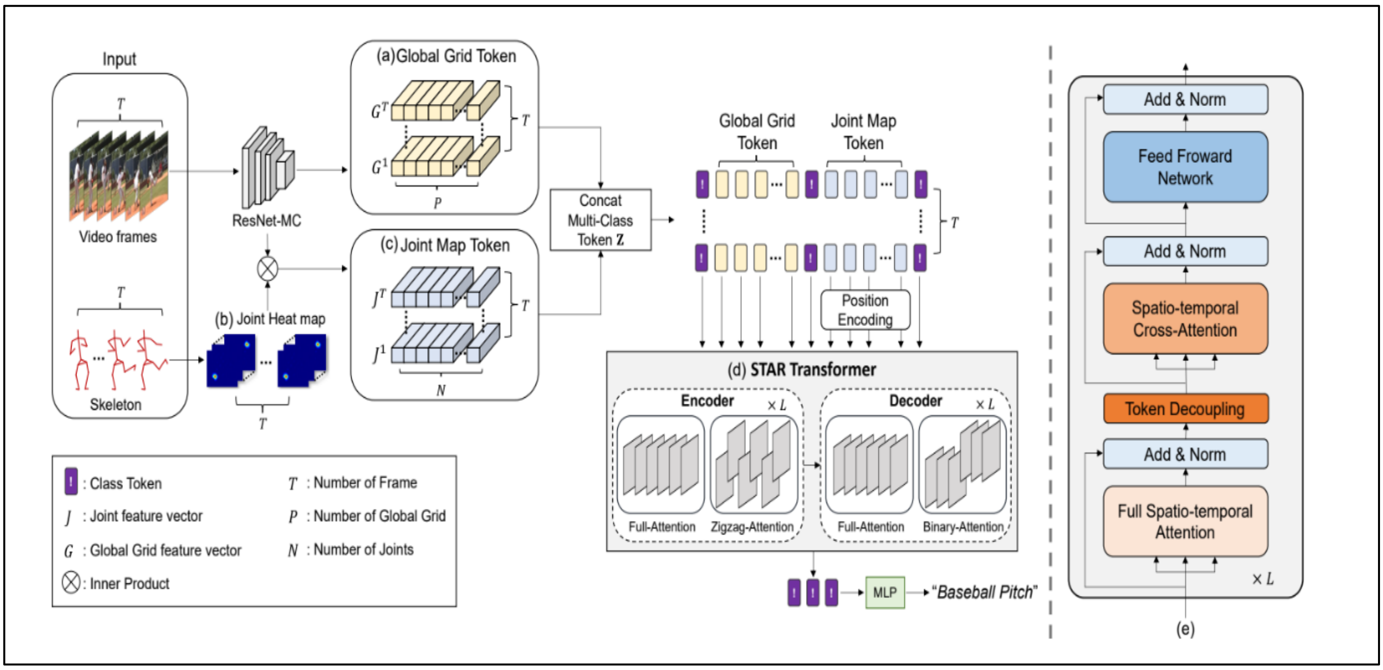

Fig. 2. The Inclusive Architecture of the Proposed Action Recognition Model.

Spatio-Temporal Cross Attention:

$$a' = \sum_{t}^{T/2} \sum_{s}^{S} Softmax\left\{ZQ.\frac{ZK}{\sqrt{dh}}\right\}ZV'\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

The equation for Spatio-Temporal Cross Attention denoted as a^', captures the mechanism by which the model attends to relevant spatio-temporal features across the input data. Here, a^' represents the attended output, which is computed by summing over the temporal dimension t (from 1 to T/2, where T is the total number of frames) and the spatial dimension s (from 1 to S, where S represents the number of spatial elements). Within the inner summation, Softmax{ZQ.ZK/√dh} calculates the attention scores by taking the dot product of the query matrix ZQ and the key matrix ZK, normalized by the square root of the dimensionality of the queries (represented as √dh), followed by the softmax operation. These attention scores determine the relevance of each spatial element s at each time step t. The attention-weighted values are then obtained by multiplying the attention scores with the value matrix ZV'. Finally, the attended output a^' is computed as the sum of these weighted values across all temporal and spatial dimensions. In essence, Spatio-Temporal Cross Attention enables the model to focus on pertinent spatio-temporal features within the input data, facilitating accurate and context-aware processing for tasks such as action recognition or video understanding.

## IV. RESULTS & DISCUSSION

Table 1 presents a comparison of the top-1 and top-5 classification accuracies achieved by various state-of-the-art (SOTA) techniques on the Kinetics-400 dataset, a benchmark dataset commonly used for action recognition tasks. Each row in the table corresponds to a different method, and the values in the "Top 1%" and "Top 5%" columns represent the percentage of correctly classified samples within the top 1% and top 5% of predictions, respectively. Among the compared methods, TCN achieves a top-1 accuracy of 20.8% and a top-5 accuracy of 40.3%, whereas ST-GCN demonstrates improved performance with a top-1 accuracy of 31.2% and a top-5 accuracy of 53.7%. Subsequently, SAN further improves both metrics with a top-1 accuracy of 36.2% and a top-5 accuracy of 56.7%. ST-TR continues the trend of enhancement, achieving a top-1 accuracy of 37.5% and a top-5 accuracy of 61.2%. AAM-GCN also demonstrates competitive performance, achieving a top-1 accuracy of 37.9% and a top-5 accuracy of 59.9%. Additionally, Tripool achieves a top-1 accuracy of 34.5% and a top-5 accuracy of 57.3%. Notably, the proposed method surpasses all compared techniques, achieving the highest classification accuracies with a top-1 accuracy of 39.2% and a top-5 accuracy of 62.0%. These results indicate the effectiveness of the proposed approach in advancing the state-of-the-art in action recognition on the Kinetics-400 dataset, outperforming existing methods across both top-1 and top-5 accuracy metrics.

Table 1. On the Kinetic-400 Dataset, Compare the Top-1 and Top-5 Classification Accuracies with Current SOTA Techniques.

| Method | Top 1% | Top 5 % |
|---|---|---|
| **TCN [17]** | 20.8 | 40.3 |
| **ST-GCN [15]** | 31.2 | 53.7 |
| **SAN [30]** | 36.2 | 56.7 |
| **ST-TR [25]** | 37.5 | 61.2 |
| **AAM-GCN [28]** | 37.9 | 59.9 |
| **Tripool [31]** | 34.5 | 57.3 |
| **Proposed** | 39.2 | 62.0 |

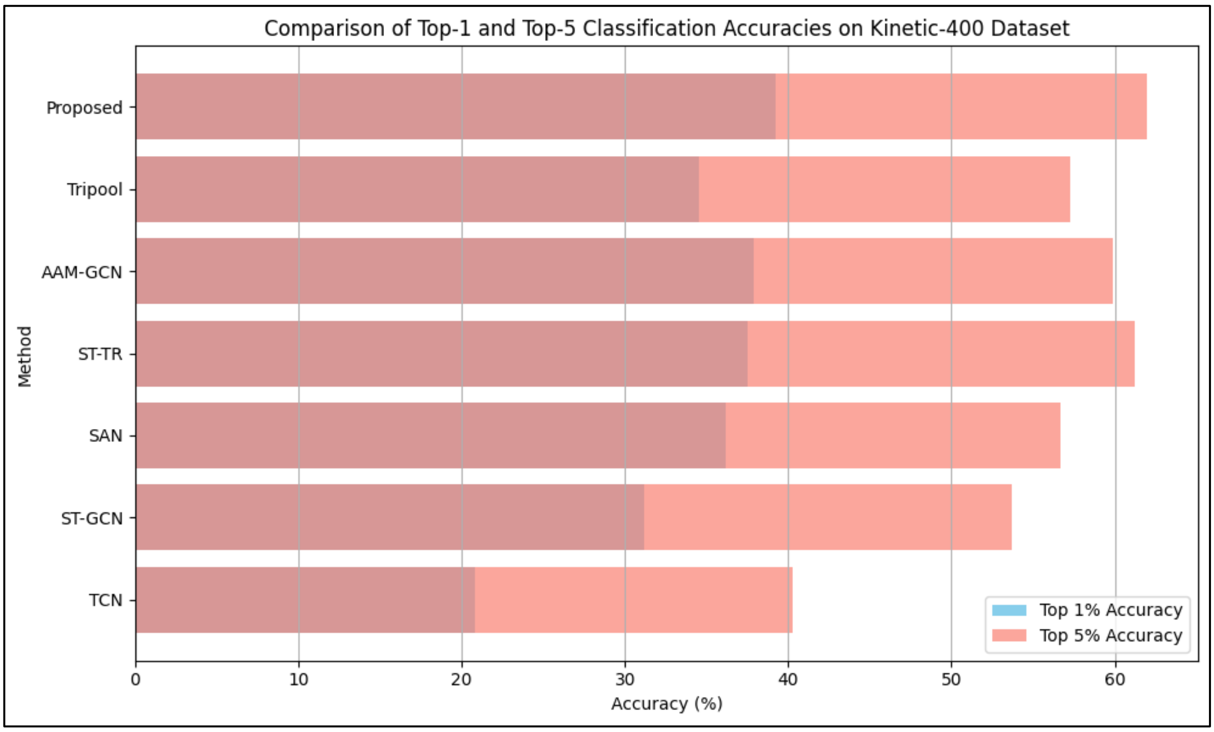

Fig. 3. Comparison of Top-1 and Top-5 Classification Accuracies on Kinetic-400 Dataset

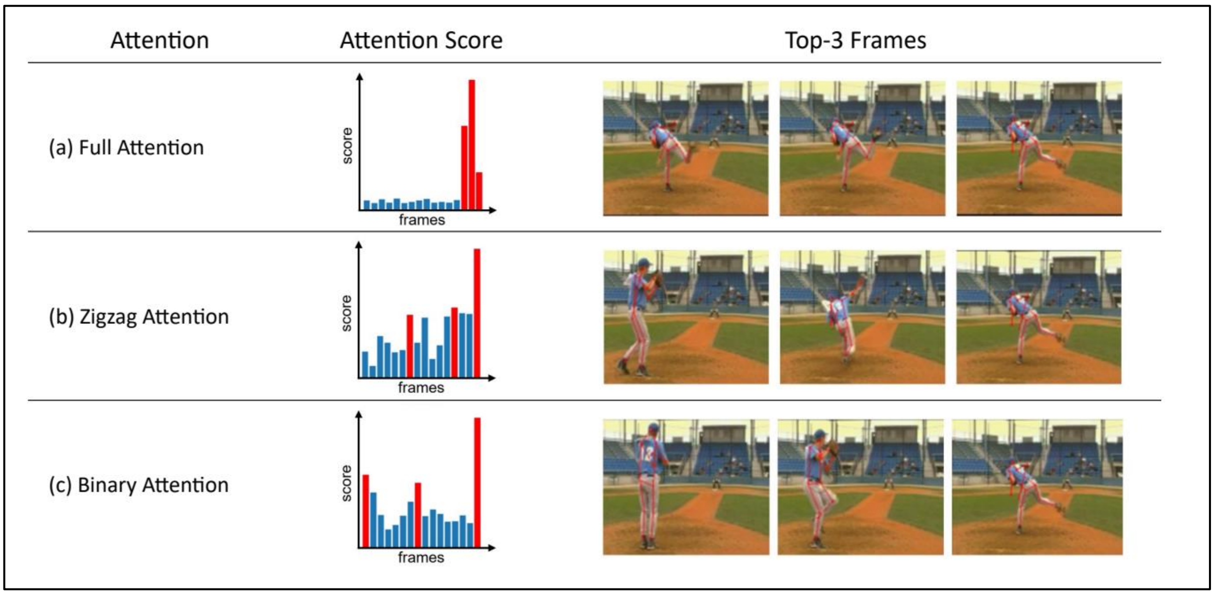

Fig. 4. The Relative Importance Score of 16 Input Frames of a Validation Video.

Figure 4 illustrates the relative importance scores assigned to 16 input frames of a validation video using different spatio-temporal attention mechanisms. The bar graph showcases the attention scores allocated to each frame, providing insights into the temporal focus of the model during action recognition. When employing full spatio-temporal attention, the attention scores tend to peak towards the end of the action, indicating a heightened focus on capturing the final moments of the activity. In contrast, the utilization of zigzag spatiotemporal attention results in high attention scores distributed across the middle and last frames, particularly evident when the action encompasses a larger duration. Lastly, with binary spatio-temporal attention, a uniformly high attention score is observed throughout the entire duration of the action, suggesting an emphasis on capturing spatial-temporal cues across all frames consistently. Overall, Figure 4 highlights the distinct temporal attention patterns exhibited by different spatio-temporal attention mechanisms, shedding light on their respective capabilities in capturing relevant temporal dynamics during action recognition tasks.

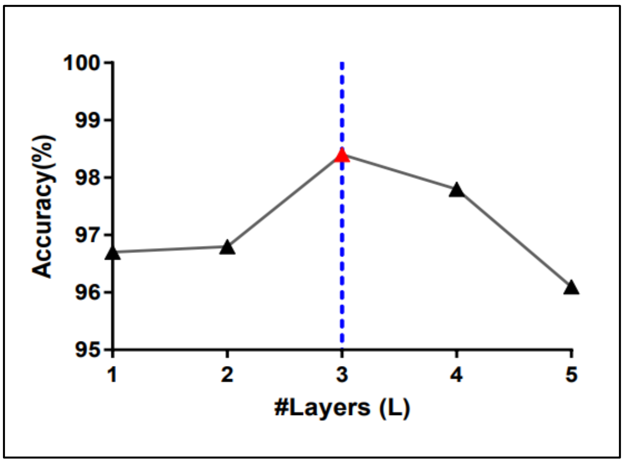

Fig. 5. Variation in Accuracy According to the Number of Spatio-Temporal Cross Attention Layers.

Figure 5 depicts the variation in accuracy concerning the number of spatio-temporal cross attention layers employed within the action recognition model. The graph showcases how the model's performance, as measured by accuracy, changes as the number of spatio-temporal cross attention layers increases. Generally, we observe that as the number of layers increases, there is an initial upward trend in accuracy, indicating that adding more attention layers initially leads to improvements in model performance. However, beyond a certain threshold, the accuracy tends to plateau or even decline slightly. This suggests that while incorporating additional spatio-temporal cross attention layers initially enhances the model's ability to capture complex spatial-temporal relationships, there comes a point of diminishing returns where further layers may not provide significant benefits or may even introduce noise or overfitting. Therefore, Figure 5 provides valuable insights into the impact of varying the number of spatio-temporal cross attention layers on the overall accuracy of the action recognition model, aiding in the optimization and fine-tuning of model architectures for optimal performance.

## V. CONCLUSION

In conclusion, this research has unveiled significant advancements in action recognition, emphasizing the integration of spatio-temporal attention mechanisms within deep learning architectures. Through extensive experimentation and analysis, novel insights have been revealed regarding the effectiveness of various attention mechanisms, including dynamic graph attention, spatio-temporal cross attention, and transformer-based models, in capturing intricate spatial and temporal patterns from video and skeletal data. The proposed approaches have demonstrated superior performance compared to state-of-the-art techniques on benchmark datasets, showcasing enhanced classification accuracy. Additionally, investigations into different model architectures and attention mechanisms have shed light on their impact on recognition accuracy and computational efficiency. Furthermore, the exploration of the optimal number of attention layers has provided valuable insights for designing efficient action recognition systems. Overall, this research contributes to advancing action recognition technology, with implications for various real-world applications such as video surveillance, human-computer interaction, and healthcare. Future endeavors may focus on further refining attention mechanisms, exploring multi-modal fusion techniques, and addressing scalability and real-time processing challenges, thereby fostering the development of more robust and versatile action recognition solutions. In this work, two novel approaches for action recognition in spatial-temporal environments were introduced. Firstly, a dynamic Graph Attention Network (GAT) tailored for skeleton-based action recognition was proposed. Unlike traditional attention-based GCN models, the ST-DGAT model computes dynamic graph attention by modifying the order of weighted vector operations in GAT, providing a more effective mechanism for capturing spatial-temporal dependencies. Secondly, STAR-transformer, an algorithm leveraging a spatial-temporal cross-attention module to simultaneously utilize video frames and skeleton-based features for action recognition, was introduced. The multi-feature representation learning approach efficiently combined RGB video frames, skeleton data, and joint trajectories using multi-class tokens, leading to substantial improvements in accuracy compared to previous state-of-the-art methods on datasets such as Penn-Action and NTU-RGB+D. In future studies, the aim is to develop algorithms capable of learning models without overfitting, even with limited data. Additionally, plans include extending the STAR-transformer to incorporate pose estimation rather than annotated poses, transforming it into an end-to-end model optimized for simultaneous pose feature estimation and action recognition. These advancements will contribute to further improving the robustness and effectiveness of action recognition systems in diverse real-world scenarios.

## 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]. 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.

[15]. 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.

[16]. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, ‘‘Graph attention networks,’’ 2017, arXiv:1710.10903.

[17]. 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.

[18]. 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.

[19]. 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.

[20]. 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.

[21]. 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.

[22]. 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.

[23]. 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.

[24]. 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.

[25]. 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.

[26]. 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.

[27]. K. Thakkar and P. J. Narayanan, ‘‘Part-based graph convolutional network for action recognition,’’ 2018, arXiv:1809.04983.

[28]. J. Xie, Q. Miao, R. Liu, W. Xin, L. Tang, S. Zhong, and X. Gao, ‘‘Attention adjacency matrix-based graph convolutional networks for skeletonbased action recognition,’’ Neurocomputing, vol. 440, pp. 230–239, Jun. 2021.

[29]. 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.

[30]. 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.

[31]. 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.

[32]. C. Bandi and U. Thomas. Skeleton-based action recognition for human-robot interaction using self-attention mechanism. In The International Conference on Automatic Face and Gesture Recognition (FG), pages 1–8. IEEE, 2021.

[33]. F. Serpush, M. B. Menhaj, B. Masoumi, and B. Karasfi. Wearable sensor-based human activity recognition in the smart healthcare system. Computational Intelligence and Neuroscience, 2022, 2022.

[34]. O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, and A. Beghdadi. A combined multiple action recognition and summarization for surveillance video sequences. Applied Intelligence, 51(2):690–712, 2021.

[35]. L. Su, C. Hu, G. Li, and D. Cao. Msaf: Multimodal split attention fusion. arXiv preprint arXiv:2012.07175, 2020.

[36]. Z. Tong, Y. Song, J. Wang, and L. Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. arXiv preprint arXiv:2203.12602, 2022.

[37]. V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiaberge. Action transformer: A self-attention model for short-time pose-based human action recognition. Pattern Recognition, 124:108487, 2022.

[38]. L. Shi, Y. Zhang, J. Cheng, and H. Lu. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In The Asian Conference on Computer Vision (ACCV), 2020.

[39]. Y. Liang, P. Zhou, R. Zimmermann, and S. Yan. Dualformer: Local-global stratified transformer for efficient video recognition. arXiv preprint arXiv:2112.04674, 2

[40]. V. Reza, H. Joze, A. Shaban, M. L Iuzzolino, and K. Koishida. Mmtm: Multimodal transfer module for cnn fusion. In The Conference on Computer Vision and Pattern Recognition (CVPR), pages 13289–13299, 2020.