scientific reports

Check for updates

OPEN

Two-stream spatio-temporal GCN-transformer networks for skeleton-based action recognition

Dong Chen[1,2,3]✉, Mingdong Chen[2,3], Peisong Wu[2,3,4], Mengtao Wu[2,3,4], Tao Zhang[2,3,4] & Chuanqi Li[3]✉

For the purpose of achieving accurate skeleton-based action recognition, the majority of prior approaches have adopted a serial strategy that combines Graph Convolutional Networks (GCNs) with attention-based methods. However, this approach frequently treats the human skeleton as an isolated and complete structure, neglecting the significance of highly correlated yet indirectly connected skeletal parts, finally hindering recognition accuracy. This study proposes a novel architecture addressing this limitation by implementing a parallel configuration of GCNs and the Transformer model (SA-TDGFormer). This parallel structure integrates the advantages of both the GCN model and the Transformer model, facilitating the extraction of both local and global spatio-temporal features, leading to more accurate motion information encoding and improved recognition performance. The proposed model distinguishes itself through its dual-stream structure: a spatiotemporal GCN stream and a spatiotemporal Transformer stream. The former focuses on capturing the topological structure and motion representations of human skeletons. In contrast, the latter seeks to capture motion representations that consist of global inter-joint relationships. Recognizing the unique feature representations generated by these streams and their limited mutual understanding, the model also incorporates a late fusion strategy to merge the results from the two streams. This fusion allows the spatiotemporal GCN and Transformer streams to complement each other, enriching action features and maximizing information exchange between the two representation types. Empirical validation on three established benchmark datasets, NTU RGB + D 60, NTU RGB + D 120, and Kinetics-Skeleton, substantiates the model's effectiveness. The experimental results indicate that, compared to existing classification frameworks, the method proposed in this paper improves the accuracy of human action recognition by 1–5% (NTU RGB + D 60 dataset). This improvement demonstrates the superior performance of the model in action recognition.

Keywords Action recognition, Graph convolutional networks, Transformer

Human action recognition has become a critical task in the domain of video understanding. Existing research comprises diverse representations of human action features, including RGB frames[1–4], optical stream[5–8] and human skeletons[9–12]. Among these representation modalities, the utilization of human skeletal features derived from depth skeleton data enables the recognition model to disregard background elements during the training process. This enables the analysis of the human skeleton in both temporal and spatial dimensions. This method effectively extracts spatial features from the topological graph for learning purposes, exhibiting wide-ranging potential applications in behavior recognition, action prediction, video understanding, intelligent monitoring, pedestrian tracking, human–computer interaction, and various other fields[13–16].

Numerous established methodologies leverage graph convolutions in conjunction with Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs) to describe temporal relationships. These methods commonly represent the human skeleton as a sequential vector of joint coordinates or a pseudo-image, thus inputting this representation into RNNs or CNNs for predictive purposes. However, this approach to skeletal data representation, whether as vector sequences or pseudo-images, does not fully capitalize on the graph structure present in skeletal data, thereby limiting its generalizability to diverse skeletal forms. Specifically,

[1]Guangxi Normal University, College of Computer Science and Engineering, Guilin 541000, China. [2]Nanning Normal University, College of Physics and Electronic Engineering, Nanning 530000, China. [3]Guangxi Key Laboratory of Functional Information Materials and Intelligent Information Processing, Nanning 530000, China. [4]Peisong Wu, Mengtao Wu and Tao Zhang contributed equally to this work. ✉email: hgccd@qq.com; lcq@mailbox.gxnu.edu.cn

skeletons are naturally structured as graphs in non-Euclidean space, with joints as vertices and the natural anatomical connections of the human body constituting edges.

In 2018, Yan et al.[17] introduced the ST-GCN model. ST-GCN effectively addressed the limitations in traditional manual modeling methods and traversal rules by constructing a human skeleton model designed to extract temporal features from consecutive frames and spatial features from intra-frame skeletal joints for action recognition tasks. ST-GCN achieved significantly superior performance on two prominent action recognition datasets, namely Kinetics and NTU RGB + D. This seminal work spurred the development of numerous GCN-based methods. Despite the demonstrated effectiveness of GCNs in action recognition, certain limitations persist. For instance, the attention mechanism in GCNs requires enhanced flexibility, as it struggles to establish connections between nodes exhibiting a higher degree of movement-related relevance and fails to fully utilize features such as human body skeletal length and direction. To reduce these limitations, Shi et al.[18] proposed the Dual-Stream Adaptive Graph Convolutional Network (2s-AGCN). This architecture addressed two key shortcomings of ST-GCN through the integration of dual-stream networks and adaptive structures. This data-driven approach enhanced the flexibility of graph construction models, enabling greater adaptability to diverse data samples. However, the constraint imposed by convolutional kernel size restricts information exchange to adjacent regions in the spatiotemporal domain. Besides, Plizzari et al.[19] proposed the innovative Spatio-Temporal Transformer Network (ST-TR), which leverages the self-attention mechanism in transformers to model inter-node dependencies. In contrast to disregarding the spatial relationships of natural connections, ST-TR models each skeleton as an independent node in the graph, thereby enhancing the accuracy of action recognition by capturing partially independent yet strongly correlated skeletal features. Both GCN models and Transformer models have inherent limitations that hinder their ability to fully capture motion features across both local and global perspectives, thereby restricting the effectiveness of action recognition. GCNs, while adept at modeling local graph structures and node relationships, may struggle to capture broader, context-aware features that are crucial for understanding complex human movements. On the other hand, Transformer models, which excel at capturing long-range dependencies and global context, may overlook fine-grained, localized motion details that are essential for accurate action recognition. As a result, neither model alone can provide a comprehensive representation of human motion, limiting the overall performance of action recognition systems.

Therefore, this paper proposes a network architecture based on a GCN parallel transformer (SA-TDGFormer), aiming to extract comprehensive spatiotemporal features from action data by simultaneously leveraging the advantages of both GCN and Transformer models. This architecture is characterized by two parallel streams: a spatiotemporal GCN stream and a spatiotemporal Transformer stream. The spatiotemporal GCN stream delves into the intricate joint relationships within the human skeletal system's topological structure, integrating an innovative Adaptive GCN (AGCN) module that dynamically refines graph representations for unparalleled accuracy in capturing inter-joint connectivities. Complementing this, the Temporal Convolutional Network (TDCN) component introduces a pioneering convolutional architecture that significantly diverges from traditional methods, offering a novel approach to temporal feature extraction. In parallel, the spatiotemporal Transformer stream stands out by meticulously capturing relationships among all nodes, both within and across frames, through its sophisticated Spatial and Temporal Transformer components. This stream excels in modeling complex, non-linear dependencies that transcend simple adjacency. Recognizing the distinct yet complementary insights offered by these two streams, one grounded in the inherent topology of the human skeleton, the other exploring relationships beyond immediate connections, this work introduces a refined late fusion strategy. This strategy harmonizes the rich action representations from both streams, fostering a more comprehensive and nuanced comparative analysis. By integrating these diverse perspectives, our approach not only leverages the natural skeletal structure but also embraces the broader, non-adjacent relationships captured by the Transformer, thereby enhancing the overall understanding and performance in action recognition tasks.

In summary, the main contributions of our work can be summarized as follows:

- We propose an adaptive Graph Convolutional Network model, which introduces an Adaptive Graph Convolutional Network (AGCN) to skillfully capture the complex inter-joint connections within the topological structure of the human skeletal system. Furthermore, we present a sophisticated Temporal Convolutional Network (TDCN) that meticulously uncovers the temporal dependencies between consecutive frames, providing a fine-grained depiction of motion evolution.
- We have developed a skeletal feature learning model exclusively using Transformers. The spatiotemporal Transformer stream employs advanced Spatial and Temporal Transformer modules to meticulously capture complex node interactions within individual frames and across time, transcending simple adjacency. This innovative approach reveals profound, non-linear dependencies that are often overlooked by traditional methods.
- We have introduced a novel parallel network structure, designated as the contrast GCN-Transformer network, capable of more accurately identifying information between any joints while simultaneously preserving the topological structure of the human skeletal graph. In detail, the skeleton data is inputted into the GCN and Transformer backbones for modeling advanced features. These features will be effectively combined through a late fusion strategy to achieve higher action recognition accuracy. Specifically, our proposed model attains comparable accuracy to models employing the four-stream fusion method, albeit utilizing only joint and skeleton fusion.

Related works

Graphical convolutional network (GCN)

The work on the theoretical underpinnings of graph neural networks was presented by Scarselli et al.[20], in 2009. In contrast to methods that transform images into sequential node representations for list-based processing,

GNNs operate directly on graphs in graph theory, mapping nodes to vector representations for the following analysis. The advent of GNNs addressed the limitations of conventional neural networks in processing non-Euclidean graph structures, prompting significant interest in the ubiquitous nature of graph data in the research community. Bruna et al.[21], in 2013, introduced the concept of graph convolutional neural networks, generalizing the principles of convolutional neural networks to the non-Euclidean domain. Their work outlined two primary GCN classification approaches: frequency domain and spatial domain methods. Then, ChebyNet[22] integrated spectral graph theory, parameterizing the convolutional kernel and drawing an analogy between convolution on images and convolution on general graphs. This approach achieved parameter sharing, effectively reducing computational costs. Kipf and Welling, in 2017[23], introduced a series of optimizations to the first-generation GCN and ChebyNet models, simplifying computations through the approximation of spectral convolutional kernels. This work facilitates the learning of hidden layer representations that encode both local graph structures and node features, while achieving state-of-the-art performance for graph-based semi-supervised classification tasks. In 2016, Niepert et al.[24] proposed normalized non-Euclidean structures to achieve node-level convolutions with fixed local structures. Leveraging Learnable Graph Convolutional Layers (LGCN), a predetermined number of neighboring nodes are autonomously selected in the receptive field of each central node. This process effectively transforms graph data into a grid-like structure in a 1-D format, enabling the application of standard convolution operations to general graphs. The applications of graph data modeling are remarkably diverse, and our work builds upon this foundation.

Human action recognition (HAR)

In the field of computer vision, Human Action Recognition (HAR) has developed rapidly and found wide-ranging applications. Deep learning, particularly methods based on skeleton sequences, has emerged as a key technology for HAR. These methods are primarily categorized into four types: CNN, RNN, GCN, and Transformer-based techniques.

The CNN-based methods are primarily employed for the analysis of two-dimensional images, exhibiting proficiency in the extraction of advanced semantic features. Therefore, numerous CNN-based methods encode backbone features into two-dimensional pseudo-images. In the study conducted by A. Tran et al.[25], videos are decomposed into spatial and temporal components. The spatial component comprises surface information derived from individual frames, including objects and scenes, while the temporal component comprises optical stream between frames, conveying motion information. To address this multidimensional information, the researchers proposed a novel network architecture consisting of two deep networks, each dedicated to processing the temporal and spatial dimensions, respectively. Donahue et al.[26] introduced a network structure that integrates CNN and Long Short-Term Memory (LSTM), demonstrating applicability in behavior recognition, image description, and video description. In 2017, the I3D model[27] was presented, constituting an extension of the 2DCNN Inception-V1 architecture. This model leverages pre-trained parameters from ImageNet. Experimental evaluations indicated that this model attained state-of-the-art performance on several benchmark datasets at that time. Dhiman et al.[28] introduced a novel viewpoint-invariant framework for human action recognition that leverages skeletal joint coordinates and RGB dynamic images (DIs) to explore the spatial shape structure of the skeleton. By applying the concept of transfer learning, this framework projects a hybrid representation of human features into a higher-dimensional space, enhancing the robustness of the model. Nigam et al.[29] have proposed a deep neural network architecture, named FactorNet, for efficiently recognizing actions in videos with long temporal durations. Then, Duan et al.[30] proposed the C3D network as a replacement for 2DCNN, utilizing heatmaps to explain the spatial correlation of joints and employing stacks to represent temporal sequences. Chaturvedi et al.[16] proposed an architecture that combines convolutional neural networks with the proposed Spatial and Channel-wise Attention-based ConvLSTM encoder(SCan-ConvLSTM). In this architecture, the blended attention mechanism adjusts the weights of the outputs at different locations and across different channels, enhancing the focus on critical regions. While this methodology achieved notable results, it neglected the motion correlation in skeletal data, a prevalent challenge encountered by CNN-based methodologies.

Based on RNNs, the utilization of output from the preceding time step as input in the current time step, establishing internal recursive connections, has been significantly efficacious in processing sequential data. In the domain of action recognition, RNN-based methodologies represent skeletal data as a sequence of vectors, wherein each frame contains the positional information of all joints. This approach has yielded promising outcomes in action recognition. Sun et al.[31] introduced L2STM, an extension of LSTM that learns independent hidden state transitions for individual spatial locations. In addition, Zhao et al.[32] determined that the integration of recursive neural networks with convolutional neural networks has yielded exceptional results at a system level. Wang et al.[33] introduced a novel dual-stream RNN architecture for modeling the temporal dynamics and spatial configuration of skeleton-based action recognition. Despite the aforementioned methodologies, a plethora of valuable approaches based on Recurrent Neural Networks (RNNs) have been continuously explored. Nonetheless, in terms of specific modeling aspects, RNN-based methodologies still demonstrate certain limitations.

Based on GCN, the ST-GCN, an action recognition method, has successfully overcome previous limitations in effectively extracting spatiotemporal features of actions. In 2019, Li et al.[34] introduced an enhancement to ST-GCN known as the Action Structure Graph Convolutional Network (AS-GCN). AS-GCN leverages A-links and S-links to extract features, thereby capturing more detailed action patterns and thus improving recognition performance. Si et al.[35] proposed the Attention-Enhanced Graph Convolutional LSTM Network (AGC-LSTM) for the purpose of recognizing human behavior based on skeletal data. AGC-LSTM not only captures discriminative features related to spatial configurations and temporal dynamics, but also evaluates the symbiotic relationship between the spatial and temporal domains. DGNN[36] employs an alternating spatial aggregation scheme to update joint and skeletal features. Cheng et al.[37] introduced the innovative Shift-GCN, which comprises a Shift-GCN operation and lightweight point-wise convolution. This method achieves superior model performance

despite utilizing fewer parameters and computational resources. Dong et al.[38] proposed the Multi-Scale Spatio-Temporal Graph Neural Network (MSTGNN), capable of simultaneously identifying discriminative features at multiple scales in both spatial and temporal domains. Wei et al.[39] introduced a novel approach to skeleton-based action recognition utilizing the Dynamic Hypergraph Convolutional Network (DHGCN). This model represents the skeleton structure utilizing hypergraphs and assigns weights to each joint based on its motion dynamics. Dhiman[40] proposed a view-invariant deep human action recognition framework, which is a novel integration of two important action cues: motion and shape temporal dynamics (STD). By leveraging fine-tuned InceptionV3 for motion capture and view-invariant features from HPM and SSIM for shape dynamics, this algorithm demonstrates significant performance improvements over the most advanced methods. Duan et al.[41] proposed a novel framework for skeleton-based action recognition known as Dynamic Group Spatiotemporal GCN (DG-STGCN). This framework consists of two modules: DG-GCN for spatial modeling and DG-TCN for temporal modeling. Specifically, DG-GCN utilizes a learned affinity matrix to capture dynamic graph structures, as opposed to relying on predetermined structures. DG-TCN employs grouped temporal convolutions with varying receptive fields and integrates a dynamic joint-skeleton fusion module for adaptive multi-level temporal modeling. Li et al.[42] have introduced a novel Scale-aware Graph Convolutional Network with Part-level Refinement (SaPR-GCN), which comprises a Part-level Refined Spatial Graph Convolution (PR-SGC) and a Scale-aware Temporal Graph Convolution (Sa-TGC). This enhances the precise recognition of actions with limited spatio-temporal information. Wang et al.[43] propose a two-stream graph convolution method, named Spatial-Structural GCN (SpSt-GCN). The Spatial GCN component performs information aggregation based on the topological structure of the human body, while the structural GCN component performs differentiation based on the similarity of edge node sequences, greatly enhancing flexibility. Yang et al.[44] introduced ViA, a novel self-supervised view-invariant autoencoder for skeleton action representation learning. It utilizes motion retargeting between different performers as a pretext task to disentangle the latent action-specific "Motion" features from the visual representation of 2D or 3D skeleton sequences. Chi et al.[45] proposed the InfoGCN++ model, specifically developed for online skeleton-based action recognition. By learning from both current and anticipated future movements, InfoGCN++ creates a more comprehensive representation of the entire sequence, demonstrating exceptional performance in online action recognition.

The Transformer model[46], characterized by its central attention mechanism, has become a prominent research area in recent years due to its remarkable capabilities and extensive potential. Its application in RGB-based object detection has yielded unprecedented performance levels. The critical contribution of the Transformer lies in its self-attention mechanism, enabling dynamic focus on global contextual information. Alexey et al.[47] successfully implemented a pure Transformer architecture on image patch sequences, achieving exceptional performance in image classification tasks. Shi et al.[48] introduced DSTA-Net, a novel attention network model captures spatial–temporal dependencies between skeleton nodes using three key technologies: spatial–temporal attention separation, position encoding decoupling, and spatial global normalization. By incorporating skeleton data decoupling, it emphasizes features from different scales, providing a comprehensive understanding of human movements. Bertasius et al.[49] extended the pure Transformer model to the domain of videos, proposing the TimeSformer model, which decomposes each video into a series of frame-level patches. Then, they introduced an attention mechanism that segregates spatial and temporal attention in each block of the model. ViViT[50] decomposes the Transformer encoder components along the spatial and temporal dimensions, proposing three pure Transformer models for video classification based on the ViT model. In addition, the authors demonstrate effective model regularization methods during training and leverage pre-trained image models for training on relatively small datasets. Since then, a consistent emergence of outstanding models has been observed, with VTN[51], Mformer[52], MViT[53], and STAR-transformer[54] being representative examples. While Transformer-based computer vision methods can achieve significant performance, they necessitate significant computational memory resources. Moreover, these methods primarily rely on raw RGB images or pseudo-images generated from skeleton points and heatmaps as input data, with limited research directly processing skeleton point data. Our work focuses on the utilization of complete skeleton point data.

Methods

In this section, we present a detailed explanation of the proposed SA-TDGFormer, outlining its architecture and component functionalities.

The SA-TDGFormer comprises two streams: the spatial–temporal graph convolution stream and the spatial–temporal Transformer stream, as depicted in Fig. 1. The spatial–temporal graph convolution stream, specialized in extracting motion features grounded in the topological framework of the human skeleton graph, encompasses two fundamental elements: the innovative Adaptive Graph Convolutional Network (AGCN) and the sophisticated Temporal Convolutional Network (TDCN). The AGCN component is adept at dynamically comprehending the topological intricacies of the graph, seamlessly adapting to disparities across layers and skeletal instances. Concurrently, the TDCN component meticulously unravels the temporal interdependencies woven between consecutive frames, providing a fine-grained portrayal of motion evolution. Complementing this, the spatial–temporal Transformer stream integrates spatial and temporal Transformer modules, meticulously engineered to capture the intricate interplay between arbitrary joint pairs, spanning both individual frames and the temporal continuum. This stream excels in unearthing the hidden correlations that often elude traditional methods. Specifically, both streams generate motion representations with unique characteristics, each operating with limited awareness of the other's output. In response to this, we adopted a late fusion strategy, integrating the outputs from the GCN and the Transformer stream, in order to fully exploit their respective strengths in capturing both temporal and spatial features. In addition, we introduce label smoothing(LS) during model training to optimize the conventional cross-entropy loss function. By incorporating noise and attenuating the weight of true sample labels in the loss function calculation, we effectively reduce overfitting.

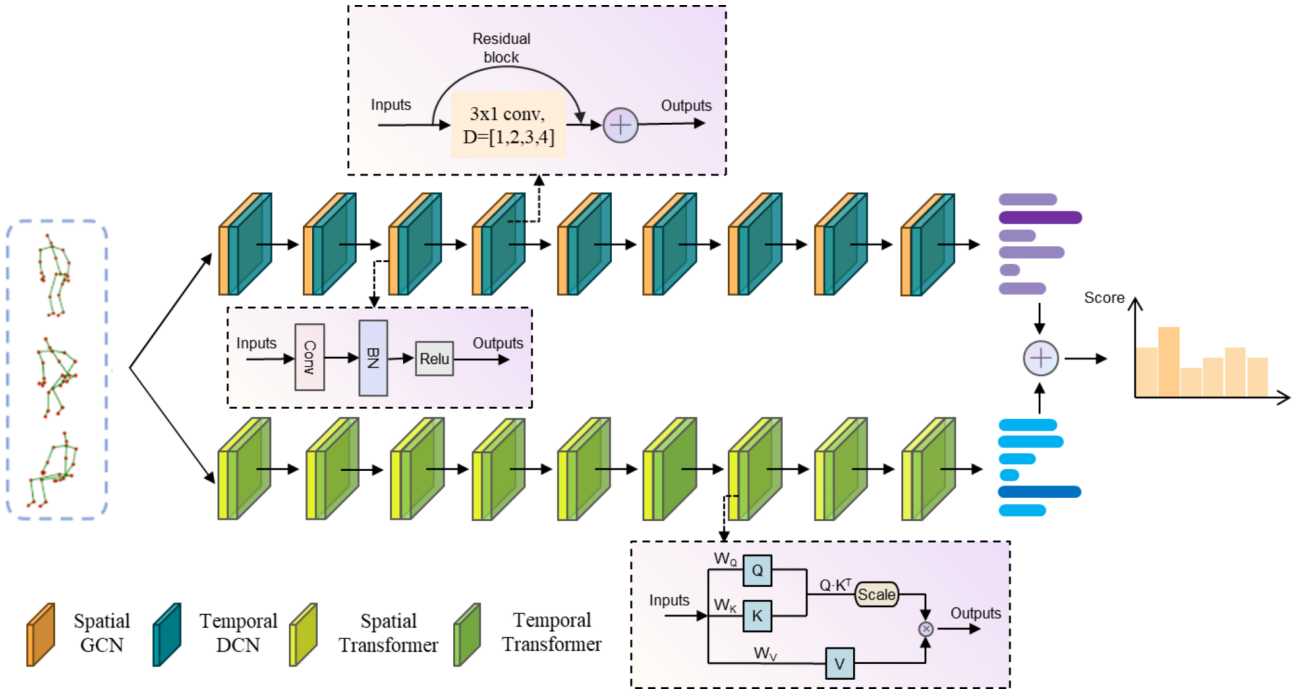

Fig. 1. The overall architecture of the SA-TDGFormer.

The stream of GCN

When given a frame-length skeletal sequence spacetime diagram $G = (V, E)$,V forms a collection of skeletal points. $|V| = n$ symbolizes the essential nodes including all moments. The set of edges E is divided into two subsets.In the depth of each frame lies the skeletal edge E_s, interconnecting every joint with accuracy, reflecting the spatial essence in a single frame.Another subset represents the connecting edges E_F between corresponding joints across frames, reflecting the temporal properties between multiple frames.The feature of vertex $G = (V, E)$ is the key node coordinate vector $F(v_{ti})$, where v_{ti} represents the coordinates of joint i in frame t.

The form of spatial graph convolutional neural networks can be obtained:

$$H^{l+1} = \sigma\left(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

The input of the $H^{(l)} \in R^{N\times D}$ layer network is , The input signal begins as $H^{(0)} = X$. Every node possesses a D-dimensional feature vector, $W^{(l)} \in R^{D\times D}$ represents the weights of the -th layer, and σ stands for the activation function.

Adaptive graph convolutional layers optimize the graph's topological structure simultaneously with other network parameters in an end-to-end learning paradigm. This adaptation results a unique graph for each layer and sample, significantly enhancing model flexibility. Additionally, designing the layers as residual branches ensures stability of the original model. Importantly, the graph's topological structure is determined by the adjacency matrix and mask. To achieve adaptive graph structuring, adaptive graph convolutional layers can be formulated as:

$$H^{l+1} = \sigma\left(\widetilde{D}^{-\frac{1}{2}}(\widetilde{A} + L + E)\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

Besides the traditional adjacency matrix $\widetilde{A}$, the connection strength between two points is represented by a matrix L, and is fully trained. The matrix E, related to the input data, learns a unique matrix for each sample, reflecting the connection strength between any two points.

In contrast to the spatial topology, the temporal topology of joints exhibits a linear structure. Therefore, temporal relationships are conventionally captured utilizing regular convolution operations instead of graph convolutions. In contrast to these traditional operations, we introduce a novel temporal convolution operation employing dilated convolutions, termed TDCN, which can be represented as:

$$H^{l+1} = TDCN\left(AGCN(H^{(l)})\right)$$

Here, $TDCN(*)$ denotes a convolution kernel with a nucleus size of K and a multi-scale expansion rate, as illustrated in Fig. 2. In comparison to regular convolution, the dilation convolution expands the receptive field of the kernel without increasing the computation cost, enabling each convolutional output to consist of a wider

range of information at each step while maintaining a consistent output feature map size. Enhancing the overall network architecture and incorporating residual connections further enhances recognition accuracy.

The stream of transformer

The Transformer model, originally designed for Natural Language Processing tasks, employs a self-attention mechanism, a non-local operator that enhances the embedding of each word by considering its surrounding context. In the Transformer, the embedding of a new word is calculated by comparing pairs of words and then blending their embeddings based on their relevance to each other. The self-attention mechanism, by gathering contextual clues, extracts a deeper meaning from each word, dynamically establishing relationships in and between phrases. The traditional Transformer matrix formulation is as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The spatial Transformer proposed in this paper abandons the connections between key nodes, initializing the connections between all body key nodes with equal strength. For skeletal data at a given frame t, for each node V_i^t of the skeleton, a trainable linear transformation is applied to the node features, extracting query vector q, index vector k, and value vector v for all nodes. The dot product of the query and index vectors is then utilized to obtain the representation of the correlation strength a_{ij}^t between nodes i and j. This process can be expressed as follows:

$$a_{ij}^t = q_i^t\left(k_j^t\right)^T$$

The matrix multiplication of strength a_{ij}^t with the value vector v is defined as follows:

$$Z_t^i = \sum_j softmax\left(\frac{a_{ij}^t}{\sqrt{d_k}}\right)V_i^t$$

In the same frame, each keypoint contains the characteristics of other keypoints, thereby establishing a mutual correlation among all keypoints in a single frame. This allows for the allocation of appropriate keypoint attention for the input data, offering a high level of correlation strength for keypoints exhibiting a strong, albeit unnatural, connection in a specific action.

The individual frame data is then transformed into the Time Transformer stream, where the relative strength of a single joint across different frames is analyzed. Each individual joint is considered independent, and the correlation between frames is calculated by comparing the changes in embeddings of the same body joint along the temporal dimension. Considering skeletal point v across different frames on the timeline, trainable linear transformations are applied between frames to extract the query vector q, index vector k, and value vector v of a single node for all frames. The dot product of the query and index vectors yields the representation of the correlation strength a_{ts}^v between frames t and s. The specific formulation is as follows:

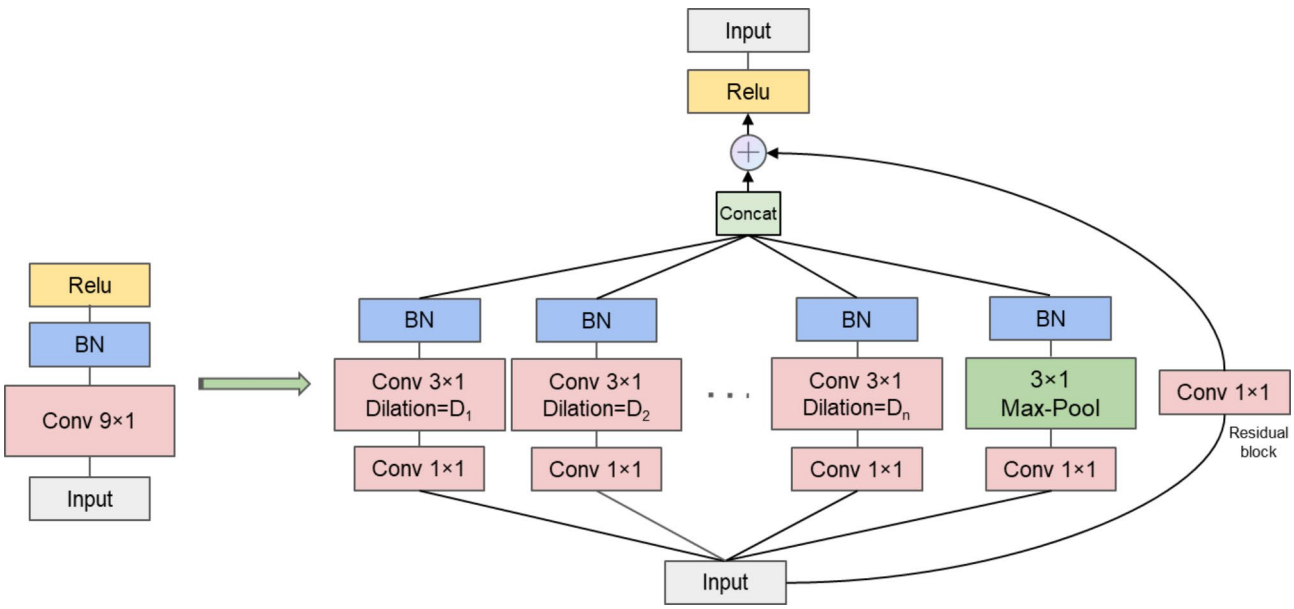

Fig. 2. Conventional convolution and enhanced convolution form.

$$a_{ts}^{v} = q_{t}^{v}(k_{s}^{v})^{T}$$

Multiplying the matrix composed of strengths and value vector v, the specific formula is as follows:

$$Z_{v}^{t} = \sum_{j} softmax\left(\frac{a_{ts}^{v}}{\sqrt{d_k}}\right) V_{t}^{v}$$

Arrange the Spatial Transformer stream and the Temporal Transformer stream in sequence, with the output of the Spatial Transformer stream serving as the input for the Temporal Transformer stream. The spatiotemporal Transformer stream can be represented as:

$$H^{l+1} = TT\left(ST(H^{(l)})\right)$$

In the equation, $TT(*)$ and $ST(*)$ represent the time transformer component and space transformer component, respectively. H $(*)$ denotes the action representation generated in the $(*)$ th layer transformer.

To maintain structural symmetry with the spatiotemporal GCN stream, the spatiotemporal Transformer layer is replicated nine times, extracting features layer by layer before finally employing Softmax for categorical predictions.

Late fusion strategy

Lastly, a late fusion strategy is adopted to better integrate the classification scores from the GCN and Transformer streams. Generally, late fusion is both efficient and concise when the dual-stream model can complement each other's features. Specifically, the classification scores from the GCN branch and the Transformer branch are fused through a decision-making layer, with the formula presented as follows:

$$O = \alpha \cdot G(V) + \beta \cdot T(V)$$

With V representing the human pose data, we process the feature maps and skeleton data separately using the GCN and Transformer streams, obtaining classification scores for each modality. Subsequently, the classification scores from the two modalities are effectively fused using two integration rates, α and β. The final result O that integrates two streams will be processed by a softmax layer for final action recognition.

Experiments

To verify the effectiveness of the proposed method, a comprehensive series of experiments were conducted using three widely recognized and frequently used datasets in the field: NTU RGB + D 60[55], NTU RGB + D 120[56], and Kinetics-Skeleton[57]. This section will offer a detailed description of these three datasets, followed by a detailed description of the experimental setup. Then, a comparative analysis will be presented, contrasting the proposed method with several state-of-the-art approaches. Finally, an study will be conducted to assess the contribution of each individual component in the proposed method.

Datasets and evaluation protocol

NTU RGB + D 60 and NTU RGB + D 120: The NTU RGB + D 60 dataset, a large-scale benchmark dataset for three-dimensional human action recognition, was collected by Shahroudy et al. using Microsoft Kinect v2. The skeleton information embedded in this dataset comprises the three-dimensional coordinates of 25 human joints, including a total of 60 action classes. The NTU RGB + D 60 dataset adheres to two evaluation standards. The first standard, denoted as Cross-View evaluation (X-View), segregates 37,920 training samples and 18,960 testing samples based on the specific camera viewpoints from which the actions were recorded. The second standard, known as Cross-Subject evaluation (X-Sub), is composed of 40,320 training samples and 26,560 testing samples collected from 40 different individuals. These individuals are divided into two groups: one designated for training and the other for testing. NTU RGB + D 120, an extension of the NTU RGB + D 60 dataset, incorporates an additional 57,367 skeletal sequences, representing 60 novel actions. For evaluation purposes, the extended dataset adopts two standards. The first standard is cross-subject evaluation (X-Sub), similar to the standard employed in NTU RGB + D 60. The second standard is cross-setup evaluation (X-Set), which replaces cross-view evaluation by partitioning training and testing samples based on unique camera setup identifiers.

Kinetics-Skeleton: The Kinetics-Skeleton dataset was derived by extracting skeleton annotations from videos constituting the Kinetics 400 dataset, using the OpenPose toolbox[58]. This dataset comprises 240,436 training samples and 19,796 testing samples, representing a comprehensive collection of 400 action classes. Each skeleton is characterized by 18 joints, with each joint defined by its two-dimensional coordinates and associated confidence levels. For each frame, a maximum of two individuals are selected based on the highest confidence score.

Implementation setup

All experiments were conducted utilizing the PyTorch deep learning framework. Optimization was achieved through the implementation of stochastic gradient descent (SGD) with Nesterov momentum (0.9). The batch size for the spatiotemporal GCN stream was set to 32, and cross-entropy was employed as the loss function for gradient backpropagation. A weight decay of 0.0001 was applied. For the NTU RGB + D 60 and NTU RGB + D 120 datasets, a maximum of two individuals were present in each sample. In instances where a sample contained fewer than two individuals, padding was applied to the second object with a value of 0. The maximum number

of frames per sample was set to 300. For samples with fewer than 300 frames, repetitive sampling was conducted until the 300-frame limit was reached. The learning rate was initialized at 0.1 and divided by 10 at the 30th, 40th, and 60th epochs. The training process concluded at the 100th epoch. The spatiotemporal Transformer stream was trained for a total of 100 epochs for both the NTU RGB + D 60 and NTU RGB + D 120 datasets. A batch size of 64 was utilized, and SGD was implemented as the optimizer. The initial learning rate was set to 0.1, and it was reduced by a factor of 10 at the 60th and 80th epochs, respectively. For the Kinetics-Skeleton dataset, both the spatiotemporal GCN stream and spatiotemporal Transformer stream were configured identically. The batch size was set to 32, with each batch containing 150 frames and 2 bodies per frame. The learning rate was initialized at 0.1 and was then divided by 10 at the 45th and 55th epochs. The training process concluded at the 80th epoch.

Comparative analysis of experimental results

To verify the effectiveness of the proposed network, its predictive accuracy was compared against several state-of-the-art methods under two evaluation protocols, utilizing the NTU RGB + D 60, NTU RGB + D 120, and Kinetics-Skeleton datasets. Table 1 presents the experimental results obtained on the NTU RGB + D 60 dataset for both individual and overall performance, where "*" denotes results obtained through experimentation employing a four-stream fusion approach (joint, skeleton, joint motion, and skeleton motion).

Table 1 offers a comprehensive comparison between the proposed work and previously established relevant methods on the NTU RGB + D 60 dataset. The proposed approach demonstrates a performance improvement of 8.5% on X-View compared to the baseline ST-GCN method and a 11.2% improvement on X-Sub. The model achieves outstanding performance, exhibiting comparable effectiveness utilizing only dual-stream data compared to methods that partially leverage four-stream data.

Table 2 presents the performance of the proposed method on the NTU RGB + D 120 dataset. As in Table 1, where "*" denotes results derived from experimentation utilizing four-stream fusion.

In Table 2, comparative experiments were conducted on the NTU RGB + D 120 dataset utilizing X-Sub and X-Set benchmarks. The competitive results presented in Table 2 confirm that the proposed method outperforms all other methods on the X-Set evaluation of the NTU RGB + D 120 dataset. Specifically, it achieves performance levels close to the current state-of-the-art on X-Sub.

Table 3 demonstrates the performance of our proposed SA-TDGFormer on the Kinetics-Skeleton dataset.

As illustrated in Table 3, our proposed SA-TDGFormer achieves a superior accuracy of 39.0% on the Kinetics-Skeleton dataset, outperforming all existing state-of-the-art methods presented in the table. Despite this advancement, the complexity and large scale of the Kinetics-Skeleton dataset pose considerable challenges for traditional models, rendering it a demanding benchmark for researchers in this domain.

Figure 3 presents a comparative analysis of our method's accuracy against the traditional ST-GCN network across individual categories in the NTU RGB + D 60 dataset.

The results shown in Fig. 3 indicate that our method consistently outperforms the ST-GCN method in terms of accuracy across all categories. However, it is also evident that there are notable shortcomings, with the accuracy of some categories being significantly lower than that of others. This trend is further evaluated in Fig. 4.

The confusion matrix obtained by SA-TDGFormer on the X-View evaluation benchmark of the NTU RGB + D dataset is shown in Fig. 4, where each row represents the actual category of the data and each column represents

Model	X-View (%)	X-Sub (%)
ST-GCN[17]	88.3	81.5
AS-GCN[34]	94.2	86.8
2S-AGCN[18]	95.1	88.5
Shift-GCN[37]	96.0	89.7
DGNN[36]	96.1	89.9
SkeletonGCL[59]	96.1	91.6
MS-G3D[60]	96.2	91.5
EfficientGCN-B4[61]	96.1	92.1
SkeletonGCL*[59]	96.4	92.2
Shift-GCN*[37]	96.5	90.7
CTR-GCN*[62]	**96.8**	92.4
2s-Shuffle-GCN[63]	96.0	90.6
2s-GATCN[64]	95.9	89.6
SARGCN[65]	94.8	88.9
PA-GCN*[66]	96.7	92.1
CD-JBF-GCN[67]	95.4	89.0
Gnet	96.4	92.4
Tnet	53.7	52.3
SA-TDGFormer	**96.8**	**92.7**

Table 1. Classification accuracy comparisons with state-of-the-art methods on NTU RGB + D 60. Best results are in bold.

Model	X-Set (%)	X-Sub (%)
ST-GCN[17]	73.2	70.7
AS-GCN[34]	79.8	78.3
Shift-GCN[37]	86.6	85.3
Shift-GCN*[37]	87.6	85.9
MS-G3D[60]	88.4	86.9
MST-GCN[68]	88.8	**87.5**
NLB-ACSE[69]	88.1	86.2
SARGCN[65]	85.1	83.8
AT-Shift-GCN[70]	88.2	86.8
Gnet	88.3	86.4
Tnet	49.2	48.9
SA-TDGFormer	**88.9**	86.8

Table 2. Classification accuracy comparisons with state-of-the-art methods on NTU RGB + D 120. Best results are in bold.

Model	Accuracy (%)
ST-GCN[17]	30.7
AS-GCN[34]	34.8
AGCN[71]	36.1
DGNN[36]	36.9
2s-GATCN[64]	36.7
MS-G3D[60]	38.0
Gnet	38.2
Tnet	30.7
SA-TDGFormer	**39.0**

Table 3. Classification accuracy comparisons with state-of-the-art methods on on Kinetics-Skeleton. Best results are in bold.

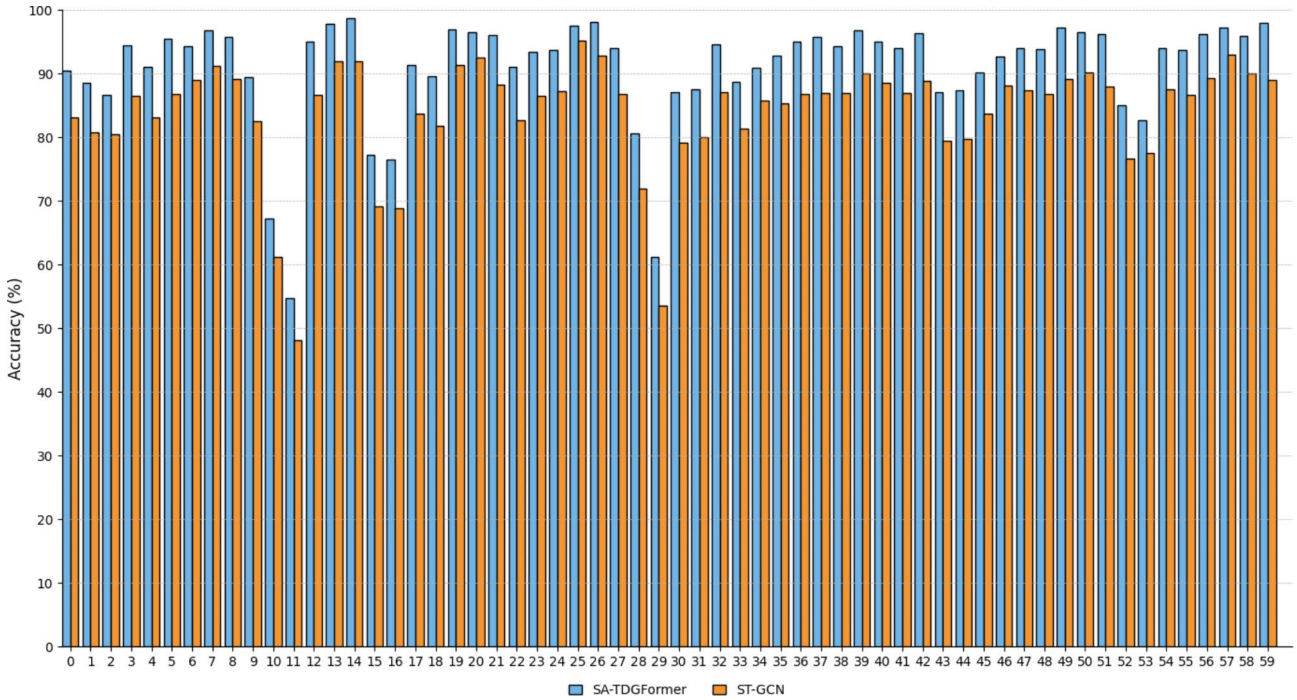

Fig. 3. Comparison of accuracy between ST-GCN and this work (NTU RGB + D 60, X-View).

the predicted category. It can be observed that excellent accuracy has been achieved in most action classes, and the recognition of "reading," "writing," "play with phone/tablet," and "type on a keyboard" action classes needs to be improved. This observation suggests a limitation of our model in accurately identifying actions that lack unique skeletal point representations. Actions such as writing and reading, characterized by subtle hand movements involving objects, present a challenge for solely skeleton-based recognition. The similarity in skeletal poses for these actions necessitates the integration of additional data modalities, such as image data, to enhance the differentiation of similar actions. This highlights a promising direction for future research.

Ablation study

In this section, we conducted a series of ablation experiments to rigorously assess the individual contributions of each module in the overall algorithmic model. Table 4 summarizes the findings of these ablation experiments.

Table 4 presents the Accuracy, FLOPs, and parameters of our models. The SA-TDGFormer model demonstrated exceptional performance, achieving an accuracy of 95.04% with 21.59G FLOPs and 3.97M parameters. When bone data was incorporated, the SA-TDGFormer model further improved its accuracy to 96.83%, albeit with an increase in computational complexity (32.78G FLOPs) and parameter count (7.36 M). The results also highlighted the significant contribution of the GCN stream, which exhibited relatively high accuracy and substantially influenced the overall model performance.

In Fig. 5, we applied t-SNE to visualize how our models gradually learn the intrinsic structure of the data. To better illustrate the distribution of action representations for each category, we randomly selected 10 classes from the X-View of the NTU-60 dataset. Each color represents a different action class, and each point represents a skeleton sequence. From the visual results, we can observe that as the network layers increase, both Gnet and Tnet become more adept at distinguishing action representations of different classes, while the representations of samples within the same class tend to converge. Furthermore, the Gnet stream demonstrates superior feature extraction capabilities, whereas the Tnet stream exhibits varying degrees of mixing among most classes by the end of training.

To further appraise the GCN algorithm model, we performed ablative experiments to isolate the effect of each component. Table 5 presents the results of these experiments.

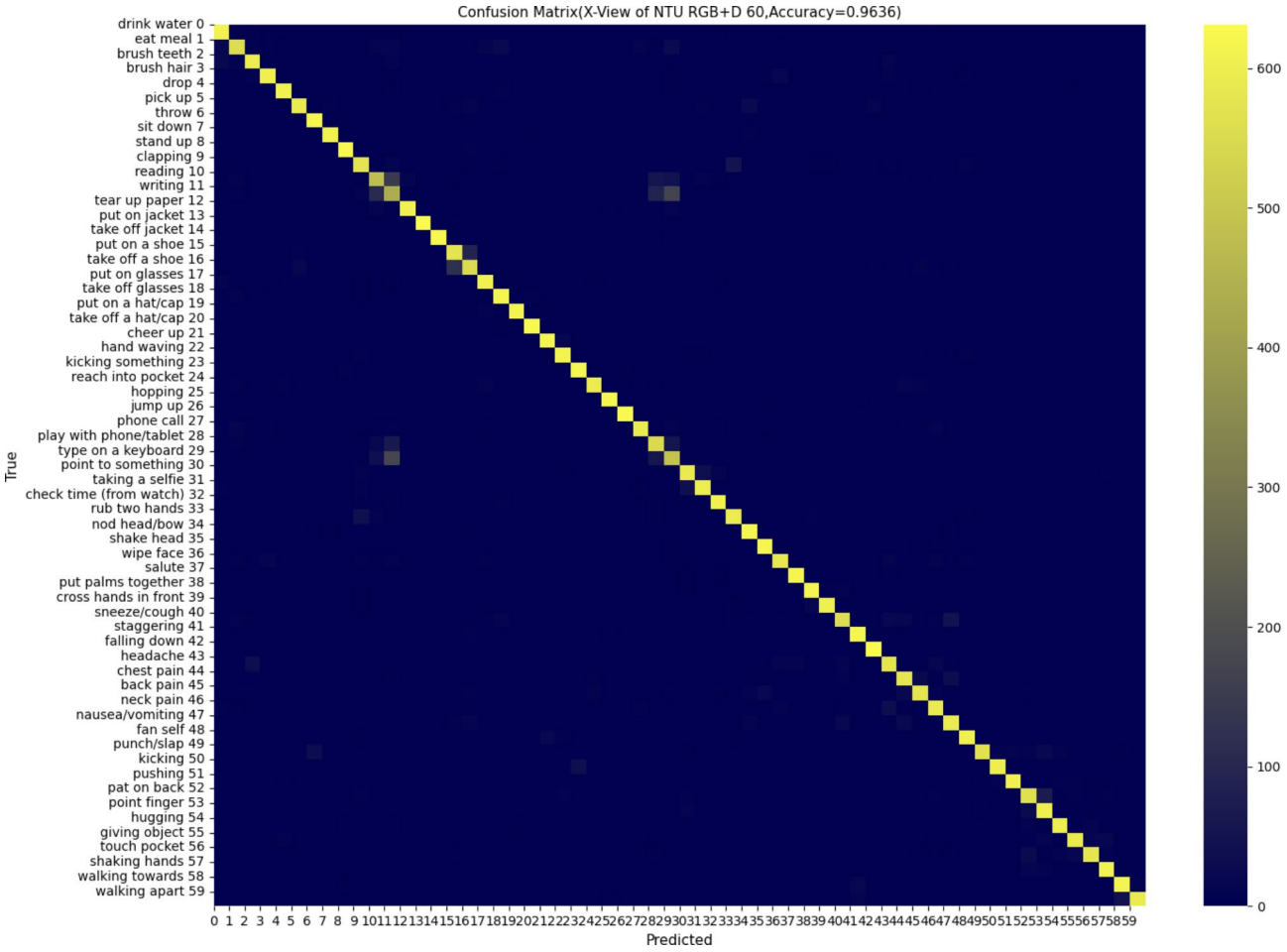

Fig. 4. Confusion matrix on the NTU RGB + D dataset with the X-View evaluation benchmark.

Model	Bones	Accuracy (%)	FLOPs ($\times 10^9$)	# Param. ($\times 10^6$)
Gnet		94.35	21.59	3.97
Tnet		52.49	14.86	2.85
SA-TDGFormer		95.04	21.59	3.97
Gnet	√	96.36	32.78	7.36
Tnet	√	53.67	25.24	4.33
SA-TDGFormer	√	**96.83**	32.78	7.36

Table 4. Ablation study of dual-stream approach on X-View of NTU RGB + D 60. Best results are in bold.

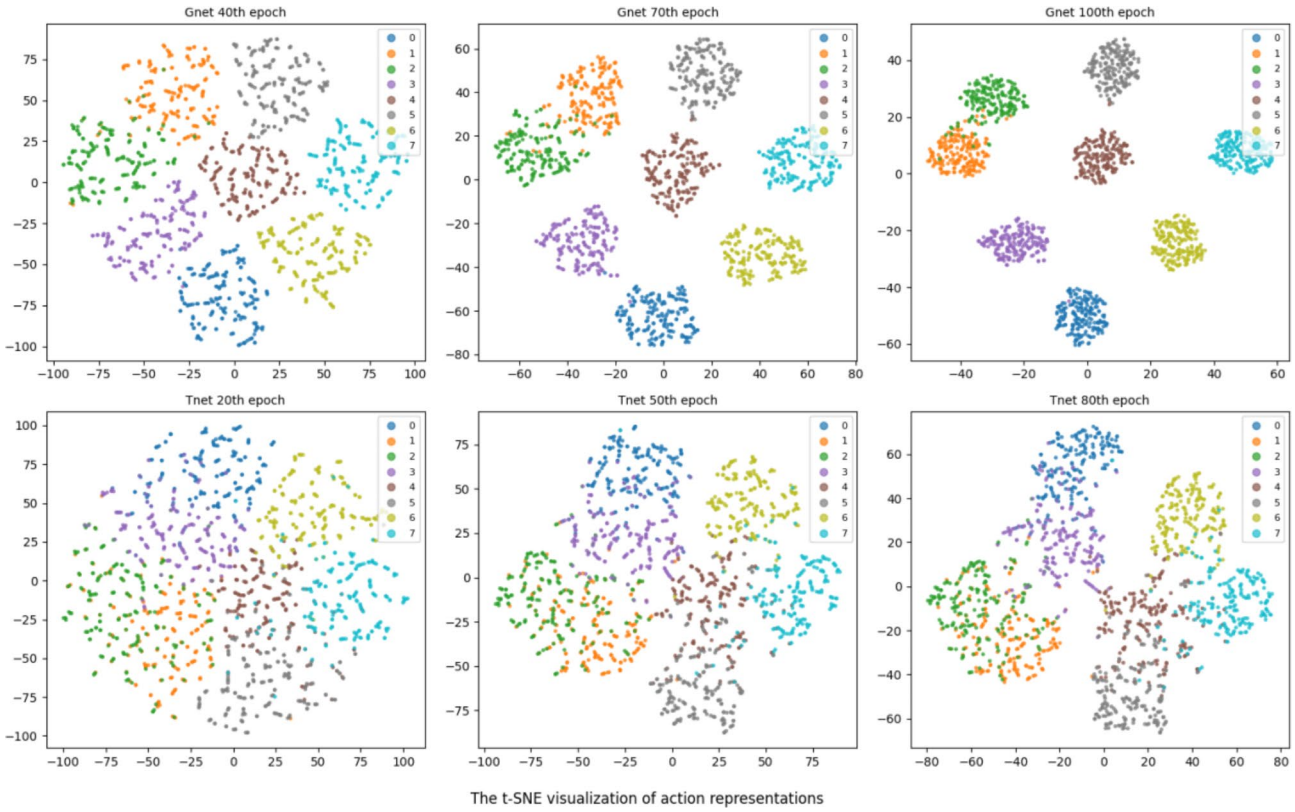

Fig. 5. The t-SNE visualization of action representations on X-View of NTU RGB + D 60.

Model	AGCN	TDCN	LS	Accuracy (%)
Baseline				92.95
Gnet(A)	√			95.13
Gnet(T)		√		94.87
Gnet(A + T)	√	√		96.12
Gnet(A + T + L)	√	√	√	**96.36**

Table 5. Ablation study of AGCN, TDCN and LS components on X-View of NTU RGB + D 60. Best results are in bold.

Table 5 demonstrates the necessity of the three main components of our proposed Gnet: AGCN, TDCN, and LS, for the model. Compared to the baseline model, our designed AGCN and TDCN components significantly enhance performance. Furthermore, utilizing the LS to optimize the loss during training yields noticeable improvements over traditional loss functions.

To investigate and demonstrate the issue of over-smoothing in GCNs, an experiment was conducted comparing the recognition performance of the Gnet model with varying numbers of GCN layers—specifically, 4, 6, 9, and 12 layers—in the X-View(Using only joint data.) setting of the NTU RGB + D 60 dataset. The results,

Layers	Accuracy (%)	# Param. ($\times 10^6$)
4	92.63	2.54
6	93.31	3.09
9	**94.40**	3.97
12	92.98	5.46

Table 6. Ablation study of GCN layer numbers' effect on performance in Gnet stream on X-View (joint) of NTU RGB + D 60. Best results are in bold.

presented in Table 6, reveal that as the number of GCN layers increases to 12, there is a significant decline in recognition accuracy. This decrement is accompanied by an increased probability of encountering the over-smoothing problem, which correlates with the notion that deeper GCNs are more prone to aggregating features from too many neighbors, resulting in the homogenization of node features and subsequently diminished classification performance[72].

Conclusion

This paper introduces a novel network architecture for action recognition, according to the parallel fusion of GCN and Transformer (SA-TDGFormer). The architecture integrates GCN and Transformer streams in a parallel manner to effectively extract both local and global spatio-temporal features from action data. The spatiotemporal GCN stream extracts action representations that reflect the complex topological structure of the human skeleton. By incorporating Adaptive Graph Convolutional (AGCN) and novel Temporal Convolutional (TDCN) components, it skillfully captures the intricate inter-joint connections within this topology. The spatiotemporal Transformer stream, on the other hand, utilizes advanced Spatial and Temporal Transformer modules to transcend simple adjacency and meticulously capture profound nonlinear node interactions both within individual frames and across time. This innovative approach uncovers global relationships between joints that are often overlooked by traditional methods, offering a more comprehensive understanding of action dynamics. Through the late fusion of dual-stream results, we significantly boost the recognition power of the dual-stream model. The experimental findings indicate that the proposed method surpasses existing classification frameworks, achieving a 1–5% improvement in accuracy for human action recognition on the NTU RGB + D 60 dataset. This improvement highlights the exceptional performance of the model in the field of action recognition. In addition, the study rigorously employs ablation studies to verify the effectiveness of individual modules in the proposed architecture. Experimental results, conducted on three publicly available datasets, demonstrate the outstanding performance of the proposed method in action recognition tasks.

Data availability

The datasets generated during and/or analyzed during our study are available from the corresponding author on reasonable request.

Received: 12 July 2024; Accepted: 21 January 2025
Published online: 10 February 2025

References

1. Ke, Q., Bennamoun, M., An, S., Boussaid, F. & Sohel, F. Human interaction prediction using deep temporal features. In *European Conference on Computer Vision*, 403–414 (Springer, 2016).
2. Feichtenhofer, C. X3d: Expanding architectures for efficient video recognition. In *CVPR,* 203–213 (2020).
3. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y. & Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 6450–6459 (2018).
4. Diba, A., Sharma, V., Van Gool, L. Deep temporal linear encoding networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2329–2338 (2017).
5. Piergiovanni, A., Ryoo, M.S. Representation stream for action recognition. In: CVPR.pp. 9945–9953 (2019).
6. Sevilla-Lara, L., Liao, Y., Güney, F., Jampani, V., Geiger, A. & Black, M. J. On the integration of optical stream and action recognition. In *GCPR*, 281–297 (2018).
7. Wang, P., Li, W., Gao, Z., Tang, C. & Ogunbona, P. O. Depth pooling based large-scale 3-d action recognition with convolutional neural networks. *IEEE Trans. Multimed.* **20**(5), 1051–1061 (2018).
8. Li, Z., Zheng, Z., Lin, F., Leung, H. & Li, Q. Action recognition from depth sequence using depth motion maps-based local ternary patterns and CNN. *Multimed. Tools Appl.* **78**(14), 19587–19601 (2019).
9. Caetano, C., Brémond, F. & Schwartz, W. R. Skeleton image representation for 3d action recognition based on tree structure and reference joints. In *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIB-GRAPI)*, 16–23 (IEEE, 2019).
10. Liu, J., Shahroudy, A., Xu, D. & Wang, G. Spatio-temporal LSTM with trust gates for 3d human action recognition. In *ECCV*, 816–833 (2016).
11. Pham, H.-H., Khoudour, L., Crouzil, A., Zegers, P. & Velastin, S. A. Exploiting deep residual networks for human action recognition from skeletal data. *Comput. Vis. Underst.* **170**, 51–66 (2018).
12. Zhang, J., Lin, L. & Liu, J. Shap-Mix: Shapley value guided mixing for long-tailed skeleton based action recognition. In *IJCAI* (2024).
13. Yu, Q., Dai, Y., Hirota, K., Shao, S. & Dai, W. Shuffle graph convolutional network for skeleton-based action recognition. *J. Adv. Comput. Intell. Intell. Inform.* **27**(5), 790–800 (2023).
14. Singh, K., Dhiman, C., Vishwakarma, D. K., Makhija, H. & Walia, G. S. A sparse coded composite descriptor for human activity recognition. *Expert Syst.* **39**(1), e12805 (2022).
15. Zhou, Y., Duan, H., Rao, A., Su, B. & Wang, J. Self-supervised action representation learning from partial spatio-temporal skeleton sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37(3), 3825–3833 (2023).

16. Chaturvedi, K., Dhiman, C. & Vishwakarma, D. K. Fight detection with spatial and channel wise attention-based ConvLSTM model. *Expert Syst.* **41**(1), e13474 (2024).
17. Yan, S. J., Xiong, Y. J. & Lin, D. H. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the Thirty-Second Association for the Advance of Artificial Intelligence, New Orleans*, 7444–7452 (2018).
18. Shi, L., Zhang, Y., Cheng, J. & Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach*, 12018–12027 (2019).
19. Plizzari, C., Cannici, M. & Matteucci, M. Skeleton-based action recognition via spatial and temporal transformer networks. *Comput. Vis. Image Underst.* **208**, 103219 (2020).
20. Scarselli, F., Gori, M., Tsoi, A., Hagenbuchner, M. & Monfardini, G. The graph neural network model. *IEEE Trans Neural Netw* **20**(1), 61–80 (2009).
21. Bruna, J., Zaremba, W., Szlam, A. & LeCun, Y. Spectral networks and locally connected networks on graphs. arXiv preprint arXiv:1312.6203 (2013).
22. Defferrard, M., Bresson, X. & Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. CoRR (2016).
23. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. CoRR (2016).
24. Niepert, M., Ahmed, M. & Kutzkov, K. Learning convolutional neural networks for graphs, 1–10. arXiv:1605.05273 (2017).
25. Tran, A. & Cheong, L.-F. Two-stream stream-guided convolutional attention networks for action recognition. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice*, 3110–3119 (2017).
26. Donahue, J. et al. Long-term recurrent convolutional networks for visual recognition and description. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston*, 2625–2634 (2015).
27. Carreira, J. & Zisserman, A. Quo Vadis. Action Recognition? A New Model and the Kinetics Dataset. IEEE (2017).
28. Dhiman, C. & Vishwakarma, D. K. View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. *IEEE Trans Image Process* **29**, 3835–3844 (2020).
29. Nigam, N., Dutta, T. & Gupta, H. P. Factornet: Holistic actor, object, and scene factorization for action recognition in videos. *IEEE Trans Circuits Syst. Video Technol.* **32**(3), 976–991 (2021).
30. Duan, H., Zhao, Y., Chen, K., Lin, D. & Dai, B. Revisiting skeleton-based action recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans*, 2959–2968 (2022).
31. Sun, L., Jia, K., Chen, K., Yeung, D. Y., Shi, B. E. & Savarese, S. Lattice long short-term memory for human action recognition. In *2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy*, 2166–2175 (2017).
32. Zhao, R., Ali, H. & van der Smagt, P. Two-stream RNN/CNN for action recognition in 3D videos. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver*, 4260–4267 (2017).
33. Wang, H. 7& Wang, L. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks (2017).
34. Li, M., Chen, S., Chen, X. et al. Actional-structural graph convolutional networks for skeleton-based action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach* (2019).
35. Si, C., Chen, W., Wang, W. et al. an attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach* (2019).
36. Shi, L., Zhang, Y., Cheng, J. & Lu, H. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7912–7921 (2019).
37. Cheng, K., Zhang, Y., He, X. et al. Skeleton-based action recognition with shift graph convolutional network. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle* (2020).
38. Feng, D., Wu, Z., Zhang, J. & Ren, T. Multi-scale spatial temporal graph neural network for skeleton-based action recognition. *IEEE Access* **9**, 58256–58265 (2021).
39. Wei, J. et al. Dynamic hypergraph convolutional networks for skeleton-based action recognition. ArXiv abs/2112.10570 (2021).
40. Dhiman, C., Vishwakarma, D. K. & Agarwal, P. Part-wise spatio-temporal attention driven CNN-based 3D human action recognition. *ACM Trans. Multimid. Comput. Commun. Appl.* **17**(3), 1–24 (2021).
41. Duan, H. et al. DG-STGCN: Dynamic spatial-temporal modeling for skeleton-based action recognition. arXiv:2210.05895 (2022).
42. Li, C., Mao, Y., Huang, Q., Zhu, X. & Wu, J. Scale-aware graph convolutional network with part-level refinement for skeleton-based human action recognition. *IEEE Trans. Circuits Syst. Video Technol.* **34**, 4311–4324 (2023).
43. Wang, J., Falih, I. & Bergeret, E. Skeleton-based action recognition with spatial-structural graph convolution. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–9. IEEE (2024).
44. Yang, D., Wang, Y., Dantcheva, A., Garattoni, L., Francesca, G. & Brémond, F. View-invariant skeleton action representation learning via motion retargeting. Int. J. Comput. Vis. 1–16 (2024).
45. Chi, S., Chi, H. G., Huang, Q. & Ramani, K. InfoGCN++: Learning representation by predicting the future for online skeleton-based action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* (2024).
46. Vaswani, A. et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, 6000–6010 (Curran Associates Inc., Red Hook, 2017).
47. Dosovitskiy, A., Beyer, L., Kolesnikov, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv: Computer Vision and Pattern Recognition (2020).
48. Shi, L. et al. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *Asian Conference on Computer Vision* (2020).
49. Bertasius, G. et al. Is space-time attention all you need for video understanding? arXiv:2102.05095 (2021).
50. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M. & Schmid, C. ViViT: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal*, 6816–6826 (2021).
51. Neimark, D., Bar, O., Zohar, M. & Asselmann, D. Video transformer network. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal*, 3156–3165 (2021).
52. Patrick, M. et al. Keeping your eye on the ball: Trajectory attention in video transformers. In *Neural Information Processing Systems Foundation*, 12493–12506 (2021).
53. Fan, H. et al. Multiscale vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal*, 6804–6815 (2021).
54. Ahn, D., Kim, S., Hong, H. & Ko, B. C. Star-transformer: a spatio-temporal cross attention transformer for human action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3330–3339 (2023).
55. Shahroudy, A., Liu, J. Ng, T.-T. & Wang, G. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas* (2016).
56. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y. & Kot, A. C. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42(10), 2684–2701 (2020).
57. Carreira, J. & Zisserman, A. Quo Vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 4724–4733 (2017).
58. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E. & Sheikh, Y. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. In IEEE *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43(1), 172–186 (2021).
59. Huang, X., Zhou, H., Wang, J., Feng, H., Han, J., Ding, E. et al. Graph contrastive learning for skeleton-based action recognition. arXiv preprint arXiv:2301.10900 (2023).

60. Liu, Z., Zhang, H., Chen, Z., Wang, Z. & Ouyang, W. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 143–152 (2020).
61. Song, Y. F., Zhang, Z., Shan, C. & Wang, L. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(2), 1474–1488 (2022).
62. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y. & Hu, W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13359–13368 (2021).
63. Yu, Q., Dai, Y., Hirota, K., Shao, S. & Dai, W. Shuffle graph convolutional network for skeleton-based action recognition. *J. Adv. Comput. Intell. Intell. Inf.* **27**(5), 790–800 (2023).
64. Zhou, S. B., Chen, R. R., Jiang, X. Q. & Pan, F. 2s-GATCN: Two-stream graph attentional convolutional networks for skeleton-based action recognition. *Electronics* **12**(7), 1711 (2023).
65. Zhu, Q. & Deng, H. Spatial adaptive graph convolutional network for skeleton-based action recognition. *Appl. Intell.* **53**(14), 17796–17808 (2023).
66. Guo, Q., Yang, X., Zhang, F. & Xu, T. Perturbation-augmented graph convolutional networks: A graph contrastive learning architecture for effective node classification tasks. *Eng. Appl. Artif. Intell.* **129**, 107616 (2024).
67. Tu, Z., Zhang, J., Li, H., Chen, Y. & Yuan, J. Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition. *IEEE Trans. Multimed.* **25**, 1819–1831 (2022).
68. Chen, Z., Li, S., Yang, B., Li, Q. & Liu, H. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35(2), 1113–1122 (2021).
69. Chen, H., Li, M., Jing, L. & Cheng, Z. Lightweight long and short-range spatial-temporal graph convolutional network for skeleton-based action recognition. *IEEE Access* **9**, 161374–161382 (2021).
70. Lu, C., Chen, H., Li, M. & Jing, L. Attention-guided and topology-enhanced shift graph convolutional network for skeleton-based action recognition. *Electronics* **13**(18), 3737 (2024).
71. Park, C., Park, J. & Park, S. AGCN: Attention-based graph convolutional networks for drug-drug interaction extraction. *Expert Syst. Appl.* **159**, 113538 (2020).
72. Li, Q., Han, Z. & Wu, X. M. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32(1) (2018).

Acknowledgements

This work was supported by the Key Laboratory of AI and Information Processing, Education Department of Guangxi Zhuang Autonomous Region (Hechi University) (Project No. 2024GXZDSY015), and the Guangxi Science and Technology Program (Project No. 2023AB29003 and No. 2023AB01005).

Author contributions

Dong Chen and Chuanqi Li conceived the experiments, Mingdong Chen conducted the experiments, Peisong Wu, Mengtao Wu and Tao Zhang analysed the results. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.C. or C.L.