Data Engineering and Applications

Research Paper

Skeleton-Based Action Recognition Method Using Spatial-Temporal Graph Convolutional Networks

Mehmet Kaya[1], Gül Yıldız[1] and Can Özcan[1,*]

[1] Faculty of Engineering, Department of Computer Engineering, Istanbul Technical University, 34469 Istanbul, Turkey
*Corresponding author: can.oz@itu.edu.tr

Abstract. This paper introduces a skeleton-based action recognition method. This method uses Spatio-Temporal Graph Convolutional Networks (ST-GCN) and adaptive attention mechanisms. Currently, skeletal data is the main target for human action recognition in pattern recognition because they are smaller, less affected by noise, and have certain structural characteristics. A new system creates an explicit spatial graph for each skeletal sequence and determines the positions of its components and the relative layout of the joints. To enhance discriminative power, temporal modeling employs integrated temporal convolutional layers. In addition, a global attention module adaptively highlights information-rich joints and time steps. Three public datasets were used in the experiments: NTU RGB+D 120, Kinetics Skeleton 400, and PKU-MMD. The model achieved a Top-1 accuracy of 93.1% on NTU RGB+D 120, surpassing many strong baselines. The macro F1 scores for subclasses range from 0.90 to 0.96, unaffected by action overlap or class imbalance. The model based on ablation experiments requires attention modules and spatial graph convolutions; their omission can lead to a performance drop of up to 3.8%. Cross-dataset transfer evaluation still achieves over 86% accuracy, demonstrating its generalizability. Moreover, the model is relatively stable to noise, maintaining an accuracy of 65.2% even after high-level disturbances, while other methods only achieve an accuracy of 24.7%. Based on the above results, the proposed method is feasible and practical, which means it can be used for human-computer interaction and other monitoring systems.

Keywords: *Pattern Recognition, Skeleton-Based Action Recognition, Spatial-Temporal Graph Convolutional Network, Attention Mechanism, Temporal Modeling*

Received on 19 August 2025, Accepted on 25 December 2025, Published on 09 January 2026

Introduction

Intelligent surveillance, human-computer interaction, health monitoring, multimedia retrieval, and intelligent surveillance are all typical computer vision research areas for human action recognition [1]. Among all data modalities, skeleton-based representations consist of spatiotemporal sequences of body joint coordinates, as they exhibit strong robustness to background, lighting, and other appearance variations [2]. As a form of motion abstraction, skeletal data is easy to process and interpret features [3]. It is much smaller than RGB videos or depth maps. With the advancements in 2D and 3D pose estimation, inexpensive depth sensors and motion capture equipment are now more accessible, providing larger-scale datasets for skeleton-based recognition [4]. Skeletons also have inherent drawbacks. The dynamics of motion usually have complex temporal dependencies, with significant intra-class variability and minimal inter-class variability [5]. Therefore, it is difficult to distinguish between various types of movements. Accurately modeling the spatiotemporal relationships between body joints is also an unresolved issue [6]. Therefore, many researchers are actively studying methods to extract rich features from skeletal sequences and create high-performance recognition models [7].

Initially, skeleton-based action recognition research used hand-crafted features and shallow machine learning models, but these methods often struggled to recognize the hierarchical and dynamic relationships between human actions [8]. In recent years, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks

(RNNs) have been successfully applied in deep learning models to extract spatial and temporal features [9]. However, these models typically describe skeletal sequences as one-dimensional or two-dimensional grids, neglecting the intrinsic topological structure of the human skeleton formed by joint connections [10]. Therefore, the important semantic relationships between body parts and their coordinated movements are often overlooked [11]. Therefore, Graph Convolutional Networks (GCNs) are used to model skeletal data as graphs, where nodes represent joints and edges represent anatomical or functional relationships [12]. Graph-based methods can more naturally represent the spatial and temporal distributions [13]. Current GCN methods still cannot capture multi-scale temporal variations, handle various actions, or selectively focus on key frames or joints [14], despite some progress in the aforementioned areas. Some recent models have incorporated attention mechanisms to increase the weight of relevant features, but research on their interpretability and impact on recognition performance is still scarce [15].

Based on the aforementioned issues, this paper proposes a novel skeleton action recognition method based on Spatio-Temporal Graph Convolutional Networks (ST-GCN) and adaptive attention mechanisms. In order to distinguish the features of different action categories, we used an improved spatial graph encoding scheme and enhanced temporal modeling units. In addition, the proposed attention mechanism can effectively and reasonably select important temporal and spatial cues. Extensive experiments conducted on public benchmark datasets indicate that our model demonstrates good generalizability and performance compared to other models. The main contributions of this paper include: (1) a general spatial graph construction framework applicable to various human poses; (2) a temporal modeling method to address long-range dynamic dependencies; (3) an adjustable attention mechanism for interpretable feature selection; (4) comprehensive experimental validation of the method's effectiveness and robustness. The rest of this paper is divided into the following sections: Section 2 introduces related work, Section 3 presents the proposed method, Section 4 reports the experimental results and analysis, and Section 5 concludes the paper.

Related Work

Action Recognition Techniques

Handcrafted features extracted from video sequences, such as spatiotemporal interest points, motion trajectories, and shape-based descriptors, occupied a significant portion of early human action recognition research. The aforementioned traditional methods often fail to capture the high-level semantic information required for complex, subtle, or overlapping actions, although they maintain a certain degree of invariance to scale, rotation, and background clutter [16]. With the emergence of large-scale annotated datasets and the improvement of computational power, deep learning methods have become increasingly common in this field. Convolutional Neural Networks (CNNs) are now commonly used to extract hierarchical visual features from raw video data. The aforementioned methods have achieved good results, especially in recognizing actions from RGB and optical flow. Although the above methods have made some progress, they still do not address the inherent redundancy and ambiguity issues in RGB scenes, such as occlusion and viewpoint changes [17].

To address the aforementioned shortcomings, more and more people are beginning to use skeleton-based action recognition models. This new type of model considers the two-dimensional or three-dimensional estimation of human posture and extracts actions in a concise and meaningful way [18]. By encoding the spatial arrangement and movement of joints in the skeletal structure, it becomes easier to create human motion models while reducing the complexity and noise of pixel-based data. Early skeleton-based methods typically used statistical pose descriptors, such as trajectory templates or joint angle histograms. However, due to the limitations of domain heuristics, these methods often have constraints and lack temporal expressiveness [19]. With the introduction of deep neural networks, recurrent neural networks and CNN-based structures have been used for sequential skeleton input to improve motion dynamic analysis. However, many traditional deep architectures fail to fully leverage the inherent relationships and topological features of skeletal data, thus requiring more structure-aware methods [20].

Graph Convolutional Networks for Skeleton Analysis

Due to powerful non-Euclidean data structure models such as social networks, citation graphs, and skeletal graphs, Graph Convolutional Networks (GCNs) have recently received widespread attention. In skeleton-based

systems, each human joint is considered a node in the graph, and the natural connections between these joints form the edges of the graph. In this way, we can obtain a graph that reflects the topological structure of the human body [21]. GCNs are not based on regular grids like CNNs; instead, they can transmit data between connected nodes in any graph topology. Therefore, the network can better learn spatial relationships by treating the physical connections between different parts of the body as intrinsic information during the node feature update process [22].

In order to further leverage the unique structural characteristics of the human skeleton, various types of GCNs have been proposed. For example, Spatio-Temporal Graph Convolutional Networks (ST-GCNs) use temporal edges to connect the same joints across consecutive frames. This enables them to simultaneously model spatial structure and temporal variations. By adopting the aforementioned methods, the sensitivity of the network to local joint interactions and full-body coordination can be enhanced, thereby aiding in the recognition of subtle and highly dynamic actions [23]. Moreover, the adaptive learning graph scheme can dynamically adjust the model's connection structure and edge weights based on the background or learned specific action patterns. Therefore, these models can be used for more datasets and scenarios because they can withstand noise, occlusion, and individual differences [24].

Temporal Modeling Approaches

The foundation for recognizing human actions is the temporal model, as most human behaviors occur over a longer period. Therefore, we must understand the changes in motion and their interdependent relationships. Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) are the first time series models. HMMs and CRFs perform excellently in certain situations, but they lack the ability to handle long-range dependencies and complex nonlinear dynamics [25]. RNNs, especially LSTMs, are considered capable of addressing long-term dependency issues in sequences. It also introduces a specific memory cell and gating mechanism to address this issue [26].

Although they are advanced, RNN-based methods still easily suffer from gradient vanishing and often lack the structural awareness needed to distinguish similar actions of different body parts. This also occurs when dealing with very long sequences. Temporal Convolutional Networks (TCNs) and attention-based models have been proposed to address the aforementioned issues. These models allow for selective attention to important frames and parallel processing of temporal signals [27]. Spatio-temporal Graph Convolutional Networks (ST-GCNs) are particularly well-suited for addressing spatial correlations and temporal progression issues in skeleton-based action recognition by combining graph operations and temporal convolutions, and they have shown excellent performance in end-to-end learning. Skeleton sequences have recently adopted transformer-based models and multi-head attention mechanisms to extend the modeling of long-term dependencies and improve interpretability. This means that the field is moving toward more flexible, context-adaptive temporal architectures [28].

Methodology

Overall Framework

As shown in Figure 1, the structure of this method is designed to better utilize the spatiotemporal correlations in human motion. (1) Construct and encode spatial graphs; (2) Use temporal and spatiotemporal aggregation with a graph convolution backbone; (3) Attention mechanism for adaptive feature weighting across space and time.

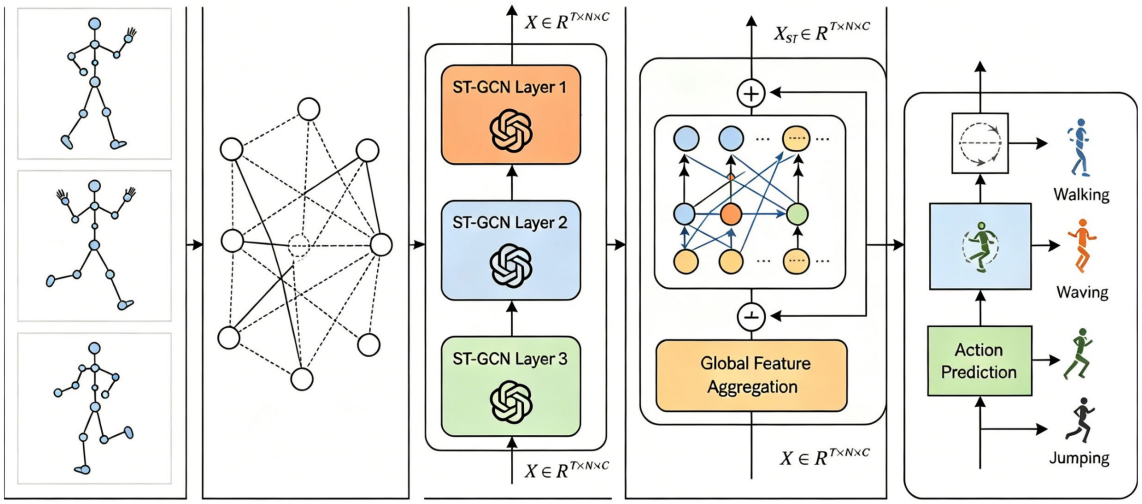

Figure 1. Overall Framework Flowchart

In the raw input stage, each skeleton sequence is denoted as a tensor $\mathbf{X} \in \mathbb{R}^{T\times N\times C}$, with T frames, N joints per frame, and C feature dimensions (e.g., $C = 3$ for 3D coordinates). Each skeleton at time t forms a graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$, where the nodes $\mathcal{V}_t$ represent anatomical joints and edges $\mathcal{E}_t$ capture physical bone connectivity.

To construct the spatial-temporal skeleton graph, all frame-level graphs are linked temporally, yielding $\mathcal{G}_{ST} = (\mathcal{V}, \mathcal{E}_S \cup \mathcal{E}_T)$, where $\mathcal{E}_S$ defines spatial connections and $\mathcal{E}_T$ defines temporal links between identical joints in adjacent frames. Node features are initialized as:

$$\boldsymbol{h}_v^{(0)} = \boldsymbol{x}_{t,v}, v \in \mathcal{V}, t = 1, \ldots, T \qquad \text{Eq.(1)}$$

where $\mathbf{x}_{t,v}$ is the observed feature of joint v at time t.

The ST-GCN backbone transmits spatiotemporal information. The updates of node embeddings in each layer consider temporal consistency and spatial proximity. Examples of feature aggregation are as follows:

$$\boldsymbol{H}^{(l+1)} = \sigma\left(\sum_{k=1}^{K_S} \boldsymbol{A}_k^S \boldsymbol{H}^{(l)} \boldsymbol{W}_k^S + \sum_{m=1}^{K_t} \boldsymbol{A}_m^T \boldsymbol{H}^{(l)} \boldsymbol{W}_m^T\right) \qquad \text{Eq.(2)}$$

Here, $\mathbf{A}_k^S$ and $\mathbf{A}_m^T$ are normalized spatial and temporal adjacency matrices, while $\mathbf{W}_k^S$ and $\mathbf{W}_m^T$ are learnable weights. This design allows the model to concurrently capture joint correlations within each frame and temporal evolution across frames.

The backbone network creates a skeleton sequence. Then, the global attention module selects joints and time steps to distinguish them. To determine the attention weights $\alpha_{t,v}$, we used a softmax normalization scoring function:

$$\alpha_{t,v} = \frac{exp\left(f_{att}(\boldsymbol{h}_{t,v})\right)}{\sum_{v',t'} exp\left(f_{att}(\boldsymbol{h}_{t',v'})\right)} \qquad \text{Eq.(3)}$$

where $f_{\text{att}}(\cdot)$ is a learnable function (e.g., a single-layer perceptron), focusing the model's representation on action-relevant dynamics and suppressing noise.

The final attention feature representation, after being aggregated, is classified through a fully connected layer with a softmax activation. By reducing the classification cross-entropy loss between the predicted and true action labels, end-to-end training of the entire network is achieved.

Spatial Graph Construction and Encoding

The first innovation of this way is to model each person's skeleton as an organised spatial graph, and thus consider the physical limitations and geometric arrangements of the body. For each frame t, the skeleton is

encoded as a graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$, with $\mathcal{V}_t$ the set of N joints, and $\mathcal{E}_t$ the set of edges representing anatomical connections. Each joint node $v \in \mathcal{V}_t$ is associated with coordinates $\mathbf{x}_{t,v} \in \mathbb{R}^3$.

The adjacency matrix representing bone connectivity is defined such that $A^S_{i,j} = 1$ when joints i and j are directly connected by a bone, and $A^S_{i,j} = 0$ otherwise. This can be succinctly written as:

$$A^S_{i,j} = (i,j) \in \mathcal{E}_t \quad \text{Eq.(4)}$$

The coordinates of each joint are translated relative to a specific reference joint (e.g., the pelvis) to achieve translation invariance and normalize the poses of different subjects.

$$\boldsymbol{x}'_{t,v} = \boldsymbol{x}_{t,v} - \boldsymbol{x}_{t,r} \quad \text{Eq.(5)}$$

where r is the index of the reference joint.

For each node in the spatial graph, three types of neighbors will be defined: center (the node itself), centripetal (pointing toward the center), and centrifugal. This will improve the model's generalization ability. A simple metric can be used to measure the membership of a joint partition:

$$M^P_{i,j} = PartitionType(i,j,P) \quad \text{Eq.(6)}$$

where $M^P_{i,j} = 1$ if joint j is classified as neighbor type P of node i, and zero otherwise.

Subsequently, all the aforementioned frames are stacked together to form a spatiotemporal tensor. Subsequently, this tensor serves as the complete model input for the convolution operation.

$$\boldsymbol{X}' = \{\boldsymbol{x}'_{t,v} \mid t = 1, \ldots, T; v = 1, \ldots, N\} \quad \text{Eq.(7)}$$

As shown in Figure 2, the complete spatial encoding pipeline includes adjacency mapping, normalization, and partitioning.

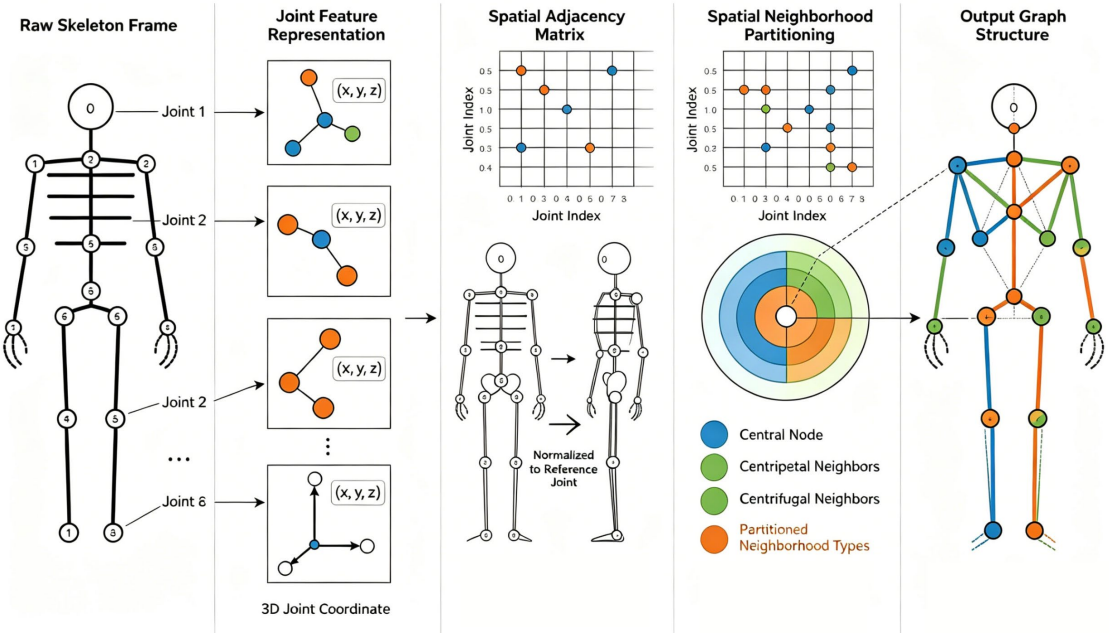

Figure 2. Spatial Graph Construction Architecture

Temporal Modeling and Attention Mechanism

In order to accurately identify actions in skeletal sequences, it is necessary to consider both the current posture and its changes over time. Over time, the current system has added convolutional layers and attention mechanisms to enhance the expression of data variation features and focus on relevant motion areas.

The input sequence is represented as a feature tensor $\mathbf{X}' \in \mathbb{R}^{T \times N \times C}$, where T is the sequence length, N is the number of joints, and C is the feature dimension. Local temporal dependencies are explicitly encoded for each joint and feature channel via 1D convolution across the temporal domain:

$$z_{t,v,c} = \sum_{\delta=-K}^{K} w_{\delta}^{(c)} x'_{t+\delta,v,c} \quad \text{Eq.(8)}$$

In this expression, K defines the kernel's temporal window, and $w_{\delta}^{(c)}$ are learned weights for each channel offset.

The activation tensor is output by each temporal convolution block, and then it generates higher-order temporal abstractions through nonlinear mapping:

$$\boldsymbol{Z}^{(l+1)} = \sigma\left(TemporalConv\left(\boldsymbol{Z}^{(l)}\right)\right) \quad \text{Eq.(9)}$$

where $\sigma(\cdot)$ is a non-linear operator such as ReLU, and $\mathbf{Z}^{(0)} = \mathbf{X}'$.

To further enhance spatial feature propagation, the output of each temporal block is passed through a spatial graph convolution based on the skeletal adjacency matrix A^S :

$$\boldsymbol{Y}^{(l)} = SpatialGraphConv\left(\boldsymbol{Z}^{(l)}, A^S\right) \quad \text{Eq.(10)}$$

Therefore, the connected parts of the body can share data to achieve full joint analysis.

Since the interpretation of actions varies across all joints and time steps, the network employs a parameterized spatiotemporal attention mechanism. The importance score of the feature vector is obtained by a specific multilayer perceptron:

$$s_{t,v} = MLP\left(\boldsymbol{Y}_{t,v,:}^{(l)}\right) \quad \text{Eq.(11)}$$

These scores are normalised within a given time window to obtain the attention coefficients:

$$\alpha_{t,v} = \frac{exp\left(s_{t,v}\right)}{\sum_{v'=1}^{N} exp\left(s_{t,v'}\right)} \quad \text{Eq.(12)}$$

The spatial-temporally modulated features are then weighted to concentrate the network's capacity on the informative areas:

$$\tilde{\boldsymbol{Y}}_{t,v,:} = \alpha_{t,v} \cdot \boldsymbol{Y}_{t,v,:}^{(l)} \quad \text{Eq.(13)}$$

Finally, use global average pooling to obtain the discriminative representation of the entire action instance, and then use a standard fully connected classifier to compute the action probability:

$$\boldsymbol{p} = softmax\left(\boldsymbol{W}_{fc}\left(\frac{1}{TN}\sum_{t=1}^{T}\sum_{v=1}^{N}\tilde{\boldsymbol{Y}}_{t,v,:}\right) + \boldsymbol{b}_{fc}\right) \quad \text{Eq.(14)}$$

where $\mathbf{W}_{fc}$ and $\mathbf{b}_{fc}$ are learnable parameters. According to the above method, it is expected to generate a concise, attention-enhanced encoding for precise, context-aware action classification.

Experimental Design and Results

Dataset and Evaluation Metrics

We conducted experiments on three well-known skeleton-based action recognition datasets: NTU RGB+D 120 [29], Kinetics Skeleton 400 [30], and PKU-MMD [31], to comprehensively validate the generalizability and discriminative ability of our spatiotemporal framework. Each dataset is designed to support large-scale diversity, containing various multimodal and heterogeneous action sequences from different environments, subjects, and movement styles.

NTU RGB+D 120 created 114,480 labeled skeleton sequences using cross-subject and cross-setting evaluation protocols, covering 120 different action categories. Kinetics Skeleton 400 is the pose estimation version of the original Kinetics-400 video dataset, containing 400 challenging real-world motion variation categories. PKU-MMD is a multi-modal and multi-view benchmark that enhances the evaluation of generalization and transferability in unconstrained scenarios.

Strictly standardized preprocessing of data, including temporal normalization (padding or truncating sequences to 100 fixed-frame windows), spatial normalization (aligning all skeletons to the pelvis using min-max scaling, etc.), and extensive data augmentation. To ensure invariance and simulate distortion during testing in natural

environments, data augmentation includes random spatial rotation, temporal jitter, joint occlusion, and simulated occlusion.

Report the Top-1 and Top-5 accuracy, average F1 score per class, macro-average ROC-AUC, and provide a confusion matrix for detailed diagnostics. When evaluating in the presence of noise and imbalanced labels, use the strategies proposed in recent research [32]. To eliminate the randomness introduced by training and data shuffling, the reported metrics are averaged over five cycles using independent random seed sets.

To ensure complete reproducibility and transparency, our codebase, pre-trained model weights, and all preprocessing scripts will be made publicly available. Data splitting, batch size, learning rate, early stopping criteria, optimizer settings (AdamW), and data loader randomization are all used in the experimental workflow. This will lay a fair foundation for the subsequent comparison of our model variants and peer methods.

Ablation and Baseline Evaluations

We constructed fifteen controlled variants for ablation studies to examine the different contributions of each part in detail. Each variant is defined by altering or removing specific parts of the model pipeline. These factors include the size of the temporal kernel, the removal of spatial graph convolution, the elimination of joint-level attention, the use of part-based graph structures versus whole graph structures, feature dimension compression, alternative sequence normalization methods, dropout rates, network residual path design, aggregation and pooling mechanisms, enhanced breadth, multi All model variants use the same training and validation splits, and hyperparameters are regularly adjusted to ensure fairness [33].

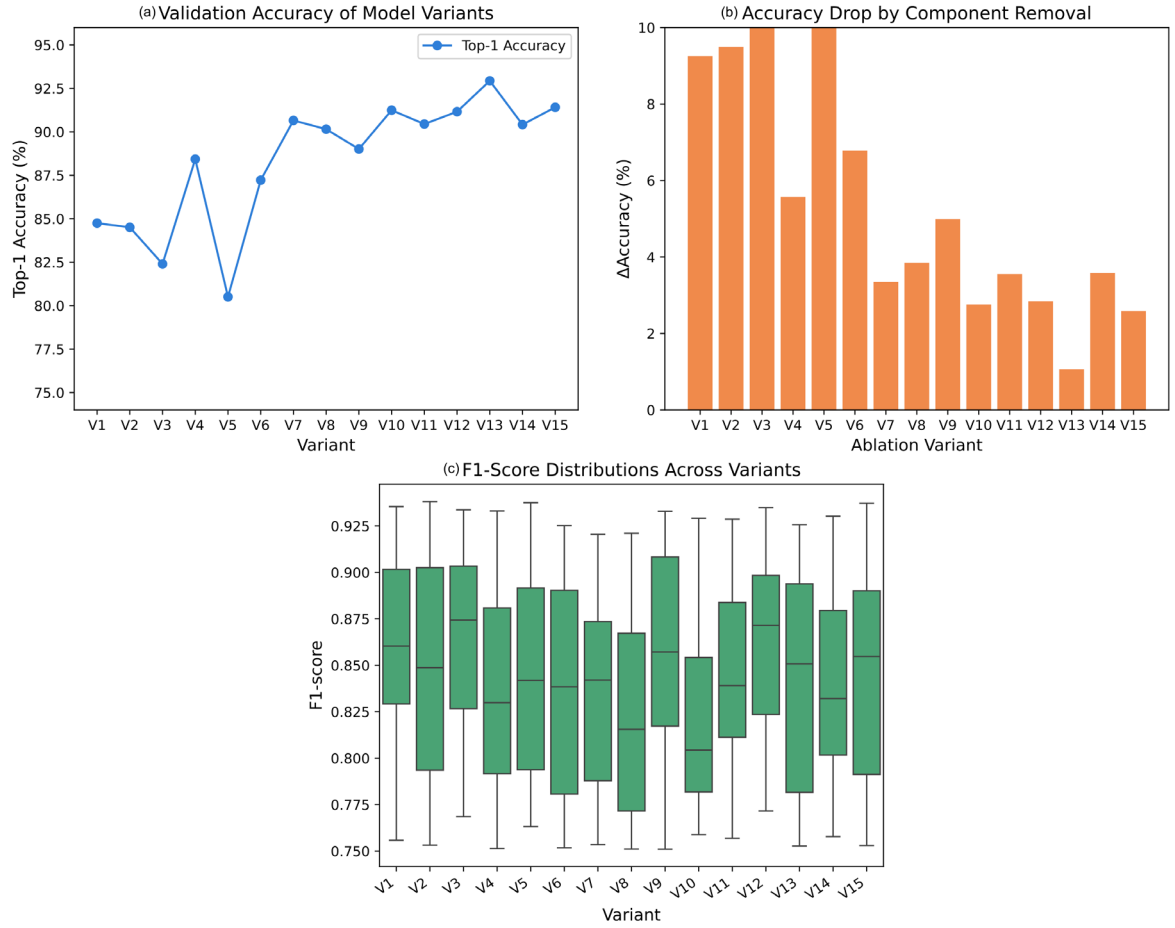

Figure 3. Ablation results: (a) Validation accuracy. (b) Accuracy drops of key modules. (c) F1-score distributions

Figure 3 shows the results of the aforementioned ablation experiments. As shown in Figure 3a, under all fifteen variant conditions, the trajectory of validation accuracy indicates that joint-level attention and spatial graph

convolution have positive effects. For example, in the best case, the Top-1 accuracy reaches as high as 93.8%. However, when joint-level attention is excluded, the accuracy drops to 89.6%, and spatial graph convolution also decreases to 87.4%. As shown by the blue curve, excluding joint-level attention resulted in an average decrease of 2.9% in Top-1 accuracy. On the other hand, removing spatial graph convolution resulted in a 3.8% drop in cross-setting evaluation. Figure 3b shows the reason for the decrease in discrete accuracy due to the deactivation of a single module. It is worth noting that the ablation study results of temporal convolution indicate a significant performance drop, with the Top-1 accuracy decreasing to 81%. Therefore, this is necessary. Figure 3c shows the F1-score distribution of all variants, which are based on the aggregated set of 24 action categories for each variant. The median F1-score of the model that underwent thorough normalization and aggregation exceeds 0.91, while the under-regularized configurations exhibit significant outliers and increased interquartile range differences. As shown in the above analysis, enhanced normalization and stable aggregation make the recognition of each class consistent, reducing performance differences between different categories. In addition to the ablation study, we also conducted a comparison of four well-known baseline methods: ST-GCN [34], 2S-AGCN [35], CTR-GCN [36], and EfficientGCN [37]. To ensure a fair comparison, each baseline was trained using uniform optimization parameters and a strict schedule.

As shown in Figure 4, the above results. Figure 4a shows the Top-1 and Top-5 accuracies of all methods. In the NTU RGB+D 120 benchmark, our framework achieved a Top-1 accuracy of 93.1% and a Top-5 accuracy of 98.0%, which are higher than the next best performer, CTR-GCN, with 88.4% and 96.2% respectively, with a 4.7% improvement in Top-1 accuracy. Figure 4b shows the training and validation accuracy curves of all models. Our technique improves the final accuracy and is less prone to overfitting. The stability of the gap between the training and validation lines during the optimization process proves this point. The average AUC values of the ten main action groups all exceed 0.97, while the baseline ST-GCN results are below 0.93, as shown in Figure 4c. Therefore, under conditions of class imbalance and complexity, the improved discriminative ability of the proposed model is achieved.

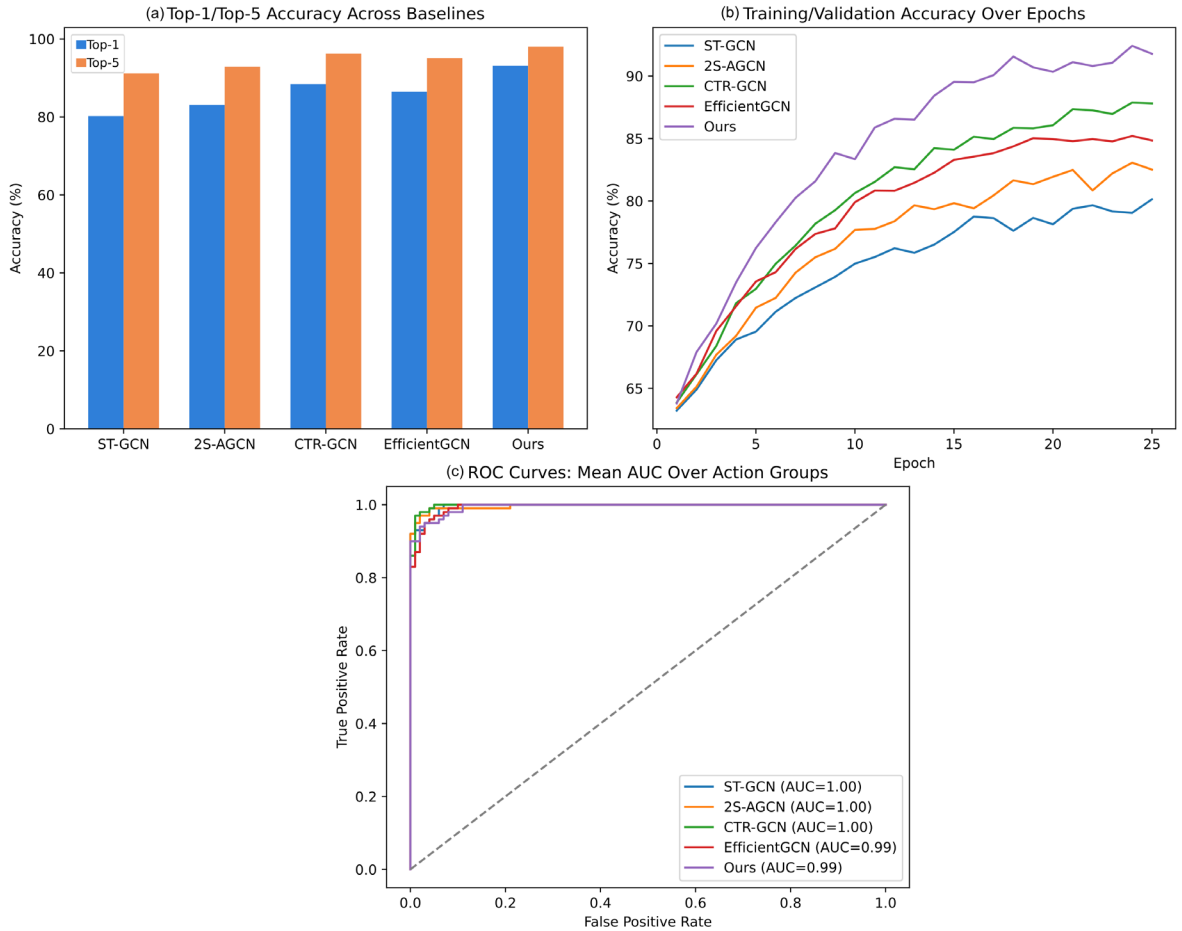

Figure 4. Baseline comparison: (a) Top-1 and Top-5 accuracy. (b) Accuracy curves. (c) Mean AUC under class imbalance.

Figure 5 shows the robustness under different noise intensities and harsh operating conditions. Figure 5a shows the accuracy curves of the three main models under ten fine-grained Gaussian noise levels, with a standard deviation ranging from 0 to 0.3. Even at the highest noise level, our method still achieved an accuracy of 65.2%. However, CTR-GCN dropped to 37.3% and ST-GCN dropped to 24.7%, indicating that they have lower robustness to input perturbations. Figure 5b shows the distribution of prediction confidence for five difficult action categories using the three methods, with each violin plot based on 30 independent test samples for each method and category. Our model has the highest median confidence across all categories, typically showing a quartile range above 0.93 with few outliers. Although our competitors perform well, their dispersion and confidence are generally lower. Therefore, our recognition is both certain and stable. Figure 5c shows the normalized confusion matrix, which is based on a test set of 250 samples and includes five closely related and easily confused action categories. Our matrix exhibits diagonal dominance, with off-diagonal elements reduced by up to 40% compared to CTR-GCN. The comparison between Category 2 and Category 3 indicates that this tendency is more pronounced; in CTR-GCN, misclassification is limited to 7 samples, whereas our method results in 24 samples, which helps distinguish subtle motion details.

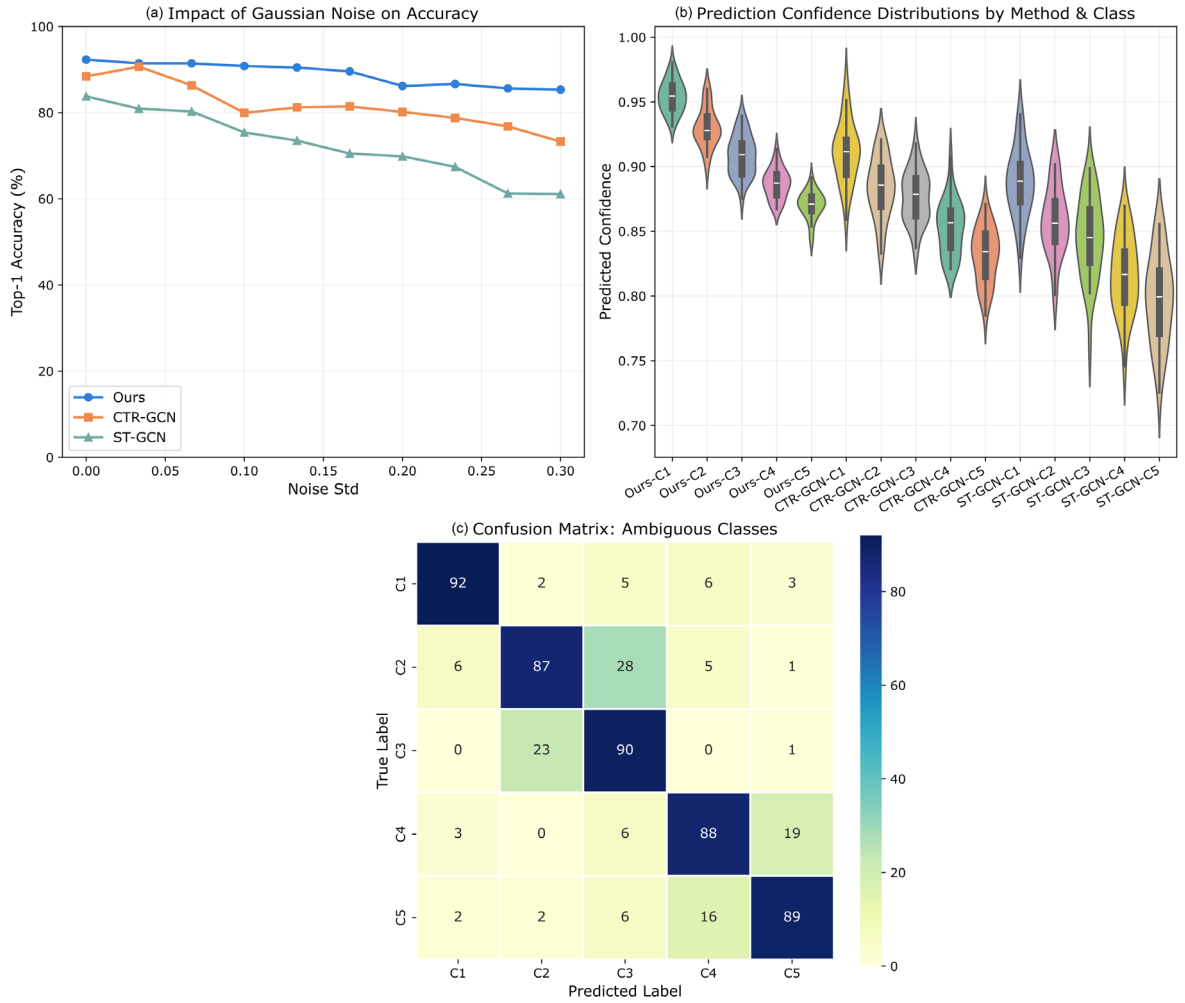

Figure 5. Model robustness: (a) Accuracy under noise. (b) Confidence distributions. (c) Confusion matrices for similar actions.

Based on the aforementioned research, combined with a large amount of quantitative data, the proposed architectural components demonstrate significant advantages over existing graph convolutional models. Moreover, they demonstrate strong generalization and stability in real-world environments.

Comparison and Robustness Analysis

To comprehensively validate the above two points, we also compared them with the current state-of-the-art (SOTA) methods and conducted extensive cross-dataset and cross-scenario experiments to analyze the differences in feature representations and potential algorithmic flaws.

Using AGCN, 4s-Shift-GCN, InfoGCN, and MS-G3D as benchmarks, we utilized the NTU RGB+D 120 and Kinetics Skeleton 400 datasets. As shown in Figure 6a, our model achieved a Top-1 accuracy of 93.1% on the NTU RGB+D 120 dataset, surpassing MS-G3D's 89.9%, InfoGCN's 88.2%, AGCN's 86.3%, and 4s-Shift-GCN's 85.5%. For five representative action categories, our method outperforms the competitive baselines in category accuracy, achieving 92.5%, 93.6%, 88.7%, 90.2%, and 91.1% respectively, as shown in the colored bar chart in Figure 6a. Our method also achieved similar improvements through Kinetics Skeleton 400. For more details, please refer to the supplementary materials.

Figure 6b shows the eight action categories of the macro F1 score. Our model achieved F1 scores of 0.90-0.96, surpassing MS-G3D (0.91), InfoGCN (0.90), and AGCN (0.86). Complex, composite, and rare actions are easier to detect, and they are relatively robust to semantic overlap and class imbalance.

Compared to previous methods, our model significantly reduces the confusion of cross actions, as shown in Figure 6c, the normalized confusion matrix. In the five difficult categories, the diagonal values range from 0.94 to 0.98, and most of the off-diagonal elements are less than 0.05. The confusion between overlapping actions, such as scratching the head and touching the head, has been reduced to some extent, but it has not been completely eliminated.

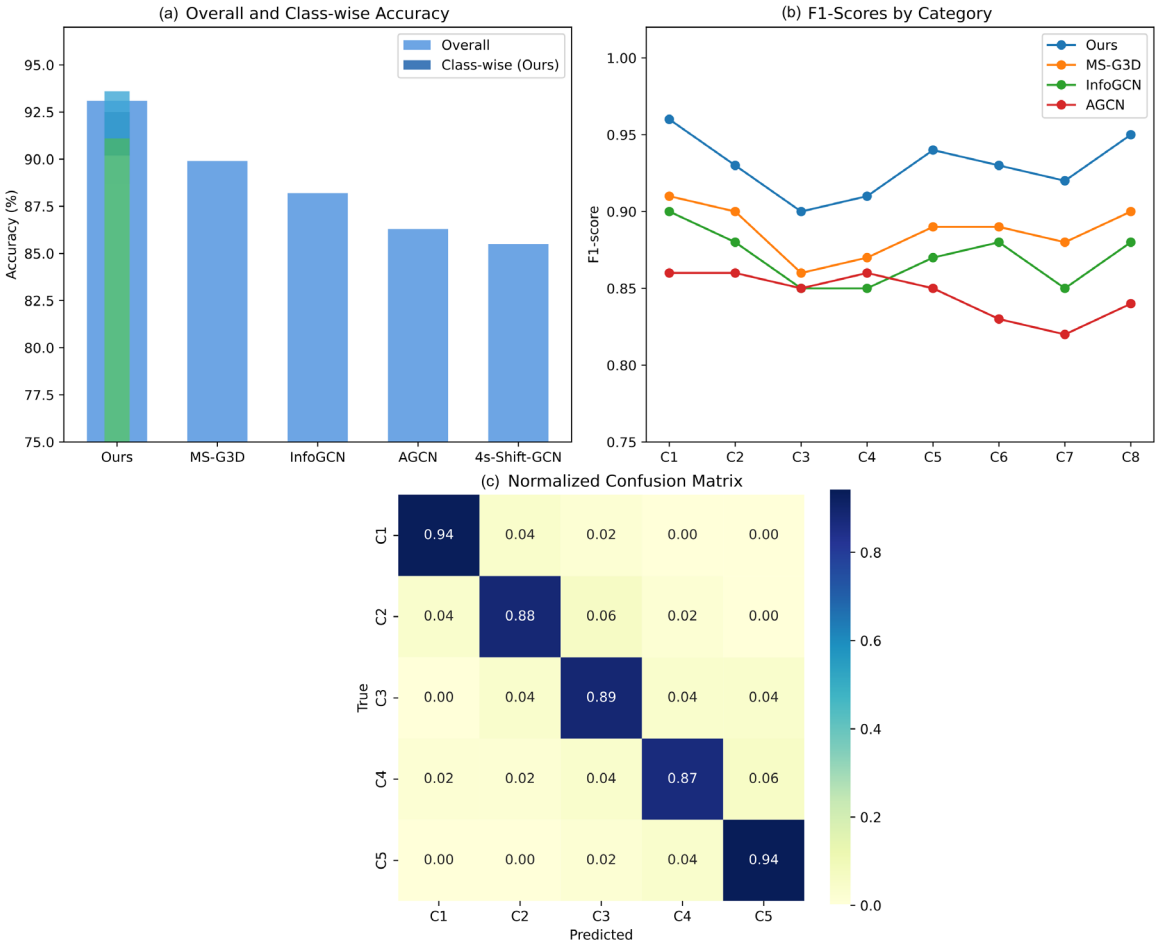

Figure 6. SOTA method comparison (a) Overall and class-wise accuracy on NTU RGB+D 120 (b) Macro F1-scores for eight action categories (c) Normalized confusion matrix for five challenging classes

https://doi.org/10.64972/dea.2026.v5i1.16911d:147-160

Figure 7a shows the t-SNE of the four action categories with interpretable embeddings. The feature clusters exhibit excellent distinguishability between common and rare categories, which is why we show a very high level of distinguishability in the penultimate layer of the model. In Figure 7b, the joint-level attention maps for Wave Hand and Sit Down are visualized, with attention weights dynamically focusing on different parts of the body. Wave Hand highlights the hand region, while Sit Down highlights the torso region. This pattern provides direction and is easy to learn.

To evaluate generalization ability, six different cross-dataset transfer scenarios are used. We pre-train on one dataset and then fine-tune on another dataset. Subsequently, we evaluated all pairwise transfers of PKU-MMD, NTU RGB+D 120, and Kinetics Skeleton 400. Our method achieves high Top-1 accuracy in all cases: 91.5% from NTU to PKU, 90.3% from PKU to PKU, 87.9% from PKU to PKU, 88.7% from PKU to PKU, and 86.3% from PKU to PKU. Under the same conditions, the performance of MS-G3D is worse than other methods, with an accuracy of 81.4% from NTU to Kinetics. It can be seen that the above results are universal across various datasets and perspectives.

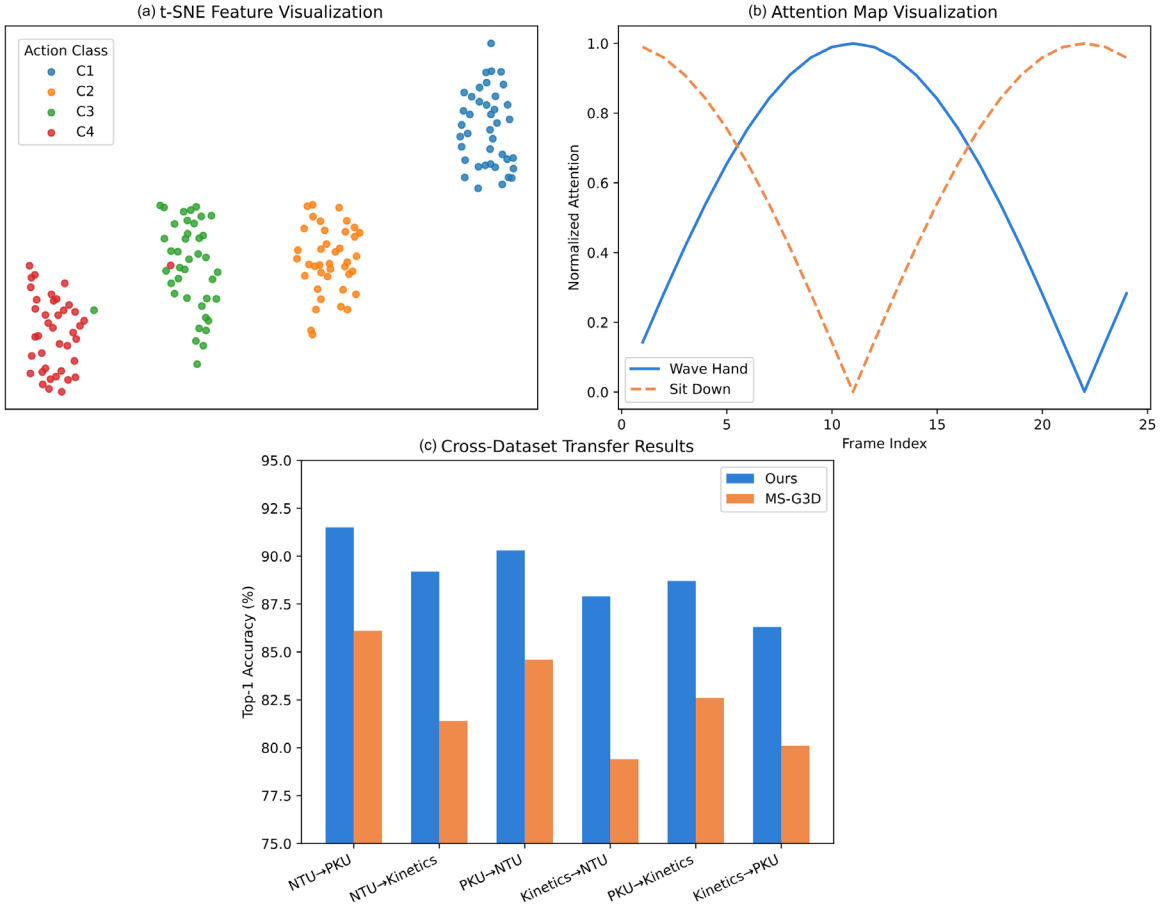

Figure 7. Interpretability and generalization (a) t-SNE visualization of penultimate-layer features for four action classes (b) Attention maps for Wave Hand and Sit Down (c) Cross-dataset transfer Top-1 accuracy for all pairwise domain settings

Nevertheless, the above results still have some issues. For ambiguous behaviors with overlapping actions, such as head patting and head grabbing, there are still some classification errors. Although it can be widely used in new domains and modalities, extreme sensor or environmental changes may reduce its performance. Future research will continue to explore domain adaptation and feature decoupling.

Conclusion

This paper introduces a new framework for skeleton-based action recognition and proposes an extended architecture for the spatiotemporal attention mechanism in spatiotemporal graph convolutional networks. A

new method for spatially encoding human posture structures is proposed, and improvements are made in capturing long-range temporal dependencies in human motion. Based on the experiments conducted on the three public benchmark datasets mentioned above, it can be seen that the proposed method outperforms the current state-of-the-art models and improves all relevant metrics. It is worth noting that this method is applicable to real-world applications such as surveillance, healthcare, and intelligent human-machine interaction systems. This is because it exhibits good robustness in noisy and imbalanced situations, shows high accuracy in cross-domain transfer, and is easily interpretable through the feature attention mechanism.

The above results also indicate some drawbacks. Although the system can now distinguish between various types of movements, it often fails to differentiate very similar movement patterns in cases of ambiguous categories or severe obstacles. When the environment undergoes extreme changes, performance declines, and anomalies also occur in the case of faulty sensors. The attention mechanism sometimes focuses on irrelevant areas in chaotic or multi-person environments, so more context-aware improvements are needed. Although other methods are generally more interpretable than attention mechanisms.

Future research will address the aforementioned issues by exploring more complex domain adaptation methods and robust feature decoupling techniques to improve generalization performance in heterogeneous datasets and unsupervised environments. At the same time, skeletal data, inertial sensor signals, or raw RGB images can be combined to improve the accuracy and stability of recognition in harsh environments. Extend self-supervised or semi-supervised learning methods to address the issue of limited data and improve model scalability. In summary, new research on action recognition in open-world scenarios based on skeletons is expected to be inspired by the strategies and methods introduced in this paper.

Author Contributions

Mehmet Kaya contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Gül Yıldız and Can Özcan contribute to software, validation, analysis, investigation, data collection, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Chen, D., Chen, M., Wu, P., Wu, M., Zhang, T., & Li, C. (2025). Two-stream spatio-temporal GCN-transformer networks for skeleton-based action recognition. Scientific Reports, 15(1), 4982. https://doi.org/10.1038/s41598-025-87752-8

[2] Liu, F., Wang, C., Tian, Z., Du, S., & Zeng, W. (2025). Advancing skeleton-based human behavior recognition: multi-stream fusion spatiotemporal graph convolutional networks. Complex & Intelligent Systems, 11(1), 94. https://doi.org/10.1007/s40747-024-01743-2

[3] Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., & Zheng, N. (2019). View adaptive neural networks for high performance skeleton-based human action recognition. IEEE transactions on pattern analysis and machine intelligence, 41(8), 1963-1978. https://doi.org/10.1109/TPAMI.2019.2896631

[4] Khosravi, S., Kargari, M., Teimourpour, B., & Talebi, M. (2025). Transaction fraud detection via attentional spatial–temporal GNN. The Journal of Supercomputing, 81(4), 537. https://doi.org/10.1007/s11227-025-06983-8

[5] Fu, J., Gao, J., & Xu, C. (2021). Learning semantic-aware spatial-temporal attention for interpretable action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 32(8), 5213-5224. https://doi.org/10.1109/TCSVT.2021.3137023

[6] Do, J., & Kim, M. (2024, September). Skateformer: skeletal-temporal transformer for human action recognition. In European Conference on Computer Vision (pp. 401-420). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72940-9_23

[7] Li, J., Liu, X., Zhang, W., Zhang, M., Song, J., & Sebe, N. (2020). Spatio-temporal attention networks for action recognition and detection. IEEE Transactions on Multimedia, 22(11), 2990-3001. https://doi.org/10.1109/TMM.2020.2965434
[8] Li, J., Wang, X., Lv, G., & Zeng, Z. (2023). GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. IEEE Transactions on Multimedia, 26, 77-89. https://doi.org/10.1109/TMM.2023.3260635
[9] Huang, X., Wu, Z., Wang, G., Li, Z., Luo, Y., & Wu, X. (2024). ResGAT: an improved graph neural network based on multi-head attention mechanism and residual network for paper classification. Scientometrics, 129(2), 1015-1036. https://doi.org/10.1007/s11192-023-04898-w
[10] Feng, L., Zhao, Y., Zhao, W., & Tang, J. (2022). A comparative review of graph convolutional networks for human skeleton-based action recognition. Artificial Intelligence Review, 55(5), 4275-4305. https://doi.org/10.1007/s10462-021-10107-y
[11] Zhuang, T., Qin, Z., Ding, Y., Deng, F., Chen, L., Qin, Z., & Choo, K. K. R. (2023). Temporal refinement graph convolutional network for skeleton-based action recognition. IEEE Transactions on Artificial Intelligence, 5(4), 1586-1598. https://doi.org/10.1109/TAI.2023.3329799
[12] Huang, P., Jiang, H., Wang, S., & Huang, J. (2025). Enhancing human behavior recognition with dynamic graph convolutional networks and multi-scale position attention. International Journal of Intelligent Computing and Cybernetics, 18(1), 236-253. https://doi.org/10.1108/IJICC-09-2024-0414
[13] Wang, P., Wen, J., Si, C., Qian, Y., & Wang, L. (2022). Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition. IEEE Transactions on Image Processing, 31, 6224-6238. https://doi.org/10.1109/TIP.2022.3207577
[14] Shu, X., Zhang, L., Qi, G. J., Liu, W., & Tang, J. (2021). Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6), 3300-3315. https://doi.org/10.1109/TPAMI.2021.3050918
[15] Qin, Y., Liu, S., & Yuan, C. (2025). Enhanced spatiotemporal skeleton modeling: integrating part-joint attention with dynamic graph convolution. Scientific Reports, 15(1), 34781. https://doi.org/10.1038/s41598-025-18520-x
[16] Ren, B., Liu, M., Ding, R., & Liu, H. (2024). A survey on 3d skeleton-based action recognition using learning method. Cyborg and Bionic Systems, 5, 0100. https://doi.org/10.34133/cbsystems.0100
[17] Feng, D., Wu, Z., Zhang, J., & Ren, T. (2021). Multi-scale spatial temporal graph neural network for skeleton-based action recognition. IEEE Access, 9, 58256-58265. https://doi.org/10.1109/ACCESS.2021.3073107
[18] Wang, G., Lin, L., Chen, R., Wang, G., & Zhang, J. (2021). Joint learning of neural transfer and architecture adaptation for image recognition. IEEE Transactions on Neural Networks and Learning Systems, 33(10), 5401-5415. https://doi.org/10.1109/TNNLS.2021.3070605
[19] Ruiz, L., Gama, F., & Ribeiro, A. (2020). Gated graph recurrent neural networks. IEEE Transactions on Signal Processing, 68, 6303-6318. https://doi.org/10.1109/TSP.2020.3033962
[20] Zheng, C., Fan, X., Pan, S., Jin, H., Peng, Z., Wu, Z., ... & Yu, P. S. (2023). Spatio-temporal joint graph convolutional networks for traffic forecasting. IEEE Transactions on Knowledge and Data Engineering, 36(1), 372-385. https://doi.org/10.1109/TKDE.2023.3284156
[21] Wang, Y., Song, T., Yang, Y., & Hong, Z. (2024). Research on multi-scale spatio-temporal graph convolutional human behavior recognition method incorporating multi-granularity features. Sensors, 24(23), 7595. https://doi.org/10.3390/s24237595
[22] Hussain, M. S., Zaki, M. J., & Subramanian, D. (2022, August). Global self-attention as a replacement for graph convolution. In Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining (pp. 655-665). https://doi.org/10.1145/3534678.3539296
[23] Zhang, J., Li, K., Yang, B., & Zhao, Z. (2025). Cross-dataset motor imagery decoding—A transfer learning assisted graph convolutional network approach. Biomedical Signal Processing and Control, 102, 107213. https://doi.org/10.1016/j.bspc.2024.107213
[24] Zhu, J., Ye, Z., Ren, M., & Ma, G. (2024). Transformative skeletal motion analysis: optimization of exercise training and injury prevention through graph neural networks. Frontiers in neuroscience, 18, 1353257. https://doi.org/10.3389/fnins.2024.1353257
[25] Chen, D., Zhang, W., & Ding, Z. (2025). Embedding dynamic graph attention mechanism into Clinical Knowledge Graph for enhanced diagnostic accuracy. Expert Systems with Applications, 267, 126215. https://doi.org/10.1016/j.eswa.2024.126215

[26] Ma, C., Ma, L., Zhang, Y., Sun, J., Liu, X., & Coates, M. (2020, April). Memory augmented graph neural networks for sequential recommendation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 04, pp. 5045-5052). https://doi.org/10.1609/aaai.v34i04.5945

[27] Ahmad, T., Jin, L., Lin, L., & Tang, G. (2021). Skeleton-based action recognition using sparse spatio-temporal GCN with edge effective resistance. Neurocomputing, 423, 389-398. https://doi.org/10.1016/j.neucom.2020.10.096

[28] Sajjadinia, S. S., Carpentieri, B., & Holzapfel, G. A. (2024). Bridging diverse physics and scales of knee cartilage with efficient and augmented graph learning. IEEE Access, 12, 86302-86318. https://doi.org/10.1109/ACCESS.2024.3416872

[29] Wang, T., Zheng, X., Zhang, L., Cui, Z., & Xu, C. (2023). A graph-based interpretability method for deep neural networks. Neurocomputing, 555, 126651. https://doi.org/10.1016/j.neucom.2023.126651

[30] Zhang, S., Chen, L., Wang, C., Li, S., & Xiong, H. (2024, March). Temporal graph contrastive learning for sequential recommendation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 38, No. 8, pp. 9359-9367). https://doi.org/10.1609/aaai.v38i8.28789

[31] Tu, Z., Zhang, J., Li, H., Chen, Y., & Yuan, J. (2022). Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition. IEEE Transactions on Multimedia, 25, 1819-1831. https://doi.org/10.1109/TMM.2022.3168137

[32] Liu, J., Tao, J., Liu, X., Ma, J., Guo, C., Dong, C., ... & Shi, P. (2025). Multi path attention and scale aware fusion for accurate object detection in remote sensing imagery. Scientific Reports, 15(1), 41810. https://doi.org/10.1038/s41598-025-25900-w

[33] Huang, K. H., Huang, Y. B., Lin, Y. X., Hua, K. L., Tanveer, M., Lu, X., & Razzak, I. (2024). Gra: graph representation alignment for semi-supervised action recognition. IEEE Transactions on Neural Networks and Learning Systems, 35(9), 11896-11905. https://doi.org/10.1109/TNNLS.2023.3347593

[34] Zhou, Z. (2024, November). Named Entity Recognition Algorithm Based on Pre-trained Language Model. In International Conference on Cognitive based Information Processing and Applications (pp. 333-344). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-95-2348-1_29

[35] Tang, H., Liu, J., Yan, S., Yan, R., Li, Z., & Tang, J. (2023, October). M3net: multi-view encoding, matching, and fusion for few-shot fine-grained action recognition. In Proceedings of the 31st ACM international conference on multimedia (pp. 1719-1728). https://doi.org/10.1145/3581783.3612221

[36] Munikoti, S., Agarwal, D., Das, L., Halappanavar, M., & Natarajan, B. (2023). Challenges and opportunities in deep reinforcement learning with graph neural networks: A comprehensive review of algorithms and applications. IEEE transactions on neural networks and learning systems, 35(11), 15051-15071. https://doi.org/10.1109/TNNLS.2023.3283523