Graph Deep Learning Project - Reproducibility challenge: MS-G3D

GDL 2024
Group id: 7
Project id: G3D

Dyuman Bulloni, Krunal Rathod
{dyuman.bulloni, krunal.rathod}@usi.ch

Abstract

Skeleton-based action recognition has earned significant attention in recent years due to its applicability in various fields such as human-computer interaction, surveillance, and healthcare. In this study, we tried to repeat the findings from the paper "Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition." The proposed methodology introduces novel techniques, including Disentangled Multi-Scale Aggregation, Disentangling Neighborhoods, and G3D, to effectively model spatial and temporal dependencies within skeletal data. We conducted experiments using benchmark datasets, including NTU RGB+D 60, NTU RGB+D 120, and Kinetics Skeleton 400, and implemented the methodology using the author's provided code. Our reproduction efforts yielded results comparable to those reported in the original paper, demonstrating the effectiveness of the proposed methodology in achieving state-of-the-art performance in skeleton-based action recognition tasks. [a]

[a] This paper contributes to the reproducibility and validation of existing research findings, providing insights into the robustness and generalization capabilities of the proposed methodology

1 Introduction

Skeleton-based action recognition has emerged as a crucial area of research within computer vision, driven by its wide-ranging applications in fields such as human-computer interaction, surveillance, and healthcare monitoring. Unlike traditional methods that rely on RGB images or videos, skeleton-based approaches analyze the spatial and temporal patterns of human skeletal movements, providing a more robust and interpretable representation of human actions. In recent years, there has been a surge of interest in developing advanced models for skeleton-based action recognition, fueled by the availability of large-scale datasets and advancements in deep learning techniques. These models aim to automatically learn discriminative features from skeletal data, enabling accurate and real-time recognition of human actions in various scenarios.

However, despite the progress made in this field, several challenges remain. Traditional approaches often rely on handcrafted features or heuristic rules, which may not generalize well across different datasets or action classes. Additionally, modeling spatial and temporal dependencies within skeletal data presents its own set of challenges, including the need to capture long-range temporal dependencies and complex joint correlations effectively. To address these challenges, researchers have turned to graph neural networks (GNNs), which offer a promising framework for modeling complex relationships in graph-structured data. By representing skeletal data as graphs, where nodes correspond to joints and edges capture spatial relationships between joints, GNNs enable the development of more robust and interpretable models for skeleton-based action recognition.

One significant contribution in this direction is the paper titled "Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition" by [1]. This paper proposes a novel methodology that leverages GNNs to model spatial and temporal dependencies within skeletal data, introducing several innovative techniques such as Disentangled Multi-Scale Aggregation, Disentangling Neighborhoods, and G3D. These techniques aim to capture complex joint correlations across spacetime, addressing the limitations of traditional methods and enabling more accurate and interpretable action recognition. Before delving into the details of the proposed methodology, it is essential to understand the evolution of skeleton-based action recognition and the challenges faced by previous approaches. Early methods in this field often relied on handcrafted features or rule-based systems to recognize actions from skeletal data. These methods typically suffered from limited scalability and generalization, as they struggled to capture the underlying dynamics of human motion effectively.

With the advent of deep learning, researchers began exploring more data-driven approaches to skeleton-based action recognition. One of the pioneering works in this area is the Spatial Temporal Graph Convolutional Network (ST-GCN) proposed by [2]. ST-GCN employs graph convolutions along with temporal convolutions to model spatial-temporal dependencies within skeletal data, achieving state-of-the-art performance on benchmark datasets.

While methods like ST-GCN represented a significant advancement in skeleton-based action recognition, they still faced limitations in capturing long-range temporal dependencies and complex joint correlations effectively. To address these challenges, subsequent works introduced innovative techniques such as multi-scale graph convolutions and attention mechanisms.

For example, [3] proposed Actional-Structural Graph Convolutional Networks (AS-GCN), which employ graph convolutions with higher-order polynomials of the adjacency matrix to capture multi-scale information from skeletal data. Similarly, [1] introduced Disentangled Multi-Scale Aggregation, which aims to capture complex joint correlations across different scales by disentangling the neighborhood information of skeletal data.

Despite these advancements, the field of skeleton-based action recognition continues to evolve, with researchers exploring new techniques and methodologies to improve the accuracy and robustness of action recognition systems. The paper by [1] represents a significant contribution in this direction, offering a unified framework for modeling spatial and temporal dependencies within skeletal data.

In the following sections of this report, we will delve deeper into the methodology proposed by [1], discussing the various techniques introduced in the paper and their implications for skeleton-based action recognition. We will also present the results of our reproduction efforts, aiming to validate the effectiveness of the proposed methodology on benchmark datasets such as NTU RGB+D 60, NTU RGB+D 120, and Kinetics Skeleton 400.

Through our reproduction efforts, we seek to contribute to the reproducibility and validation of existing research findings in skeleton-based action recognition, providing insights into the scalability and generalization capabilities of the proposed methodology. Additionally, we aim to identify any potential limitations or areas for future research, further advancing the state-of-the-art in this exciting field.

In summary, skeleton-based action recognition represents a vital area of research within computer vision, with significant implications for various real-world applications. The paper by [1] offers a novel methodology that addresses many of the challenges faced by previous approaches, providing a promising framework for more accurate and interpretable action recognition systems.

2 Related works

Skeleton-based action recognition has witnessed significant advancements in recent years, driven by the availability of large-scale datasets and advancements in deep learning techniques. In this section, we provide an overview of the existing methods in skeleton-based action recognition, highlighting their key contributions and limitations.

Neural Networks on Graphs:

The use of graph neural networks (GNNs) has emerged as a promising approach for modeling spatial and temporal dependencies within skeletal data. GNNs enable the representation of skeletal data as graphs, where nodes correspond to joints, and edges capture spatial relationships between joints. Spectral GNNs, such as those proposed by [4] and [5], convolve the input graph signals with learned filters in the graph Fourier domain. However, these methods are limited in terms of computational efficiency and generalizability to new graphs due to the requirement of eigendecomposition and fixed adjacency matrices.

Spatial GNNs, on the other hand, perform layer-wise updates for each node by selecting neighbors with a neighborhood function, merging features from selected neighbors, and applying an activation function. Among spatial GNNs, the Graph Convolutional Network (GCN) proposed by [6] has gained significant attention for its simplicity and effectiveness in modeling localized spectral convolutions. Subsequent works, such as Graph Attention Networks (GAT) by [7] and Graph Isomorphism Network (GIN) by [8], have further improved the performance of spatial GNNs by incorporating attention mechanisms and adaptively updating node features.

Multi-Scale Graph Convolutions:

To capture features from non-local neighbors and model multi-scale information in skeletal data, researchers have explored multi-scale graph convolutions. These methods typically raise the adjacency matrix to higher powers to aggregate features from long-range neighbor nodes. For example, [9] proposed the use of higher-order polynomials of the graph adjacency matrix to capture multi-scale information in graph-structured data. Similarly, [10] introduced the Truncated Block Krylov network, which employs dense feature concatenation from different hidden layers to capture multi-scale information. Despite their effectiveness, these methods may suffer from weighting bias and computational inefficiency due to the reliance on adjacency powering.

Skeleton-Based Action Recognition

Early approaches to skeleton-based action recognition primarily relied on handcrafted features or rule-based systems to recognize actions from skeletal data. These methods typically extracted features such as joint angles, velocities, or accelerations and fed them into downstream classifiers. For example, [11] proposed a method based on Hidden Markov Models (HMMs) to recognize actions from skeletal data, achieving promising results on benchmark datasets.

With the advent of deep learning, researchers began exploring more data-driven approaches to skeleton-based action recognition. One of the pioneering works in this area is the Spatial Temporal Graph Convolutional Network (ST-GCN) proposed by [2]. ST-GCN employs graph convolutions along with temporal convolutions to model spatial-temporal dependencies within skeletal data, achieving state-of-the-art performance on benchmark datasets such as NTU RGB+D and Kinetics.

Subsequent works have built upon the foundations laid by ST-GCN, introducing innovative techniques to improve the accuracy and robustness of action recognition systems. For example,[3] proposed Actional-Structural Graph Convolutional Networks (AS-GCN), which employ higher-order polynomials of the adjacency matrix to capture multi-scale information from skeletal data. Similarly, [12] introduced Disentangled Multi-Scale Aggregation, which aims to capture complex joint correlations across different scales by disentangling neighborhood information.

Despite these advancements, challenges remain in skeleton-based action recognition, particularly in capturing long-range temporal dependencies and modeling complex joint correlations effectively. Recent works have focused on addressing these challenges by introducing attention mechanisms, cross-spacetime connectivity, and joint-bone fusion techniques [12], [13], [14], [15]. These methods have significantly improved the accuracy and robustness of action

recognition systems, paving the way for more effective applications in real-world scenarios.

MS-G3D Successors:
Different works [16, 17] analyze the MS-G3D to use it as baseline for further improvements, solving different criticity. The architecture proved particularly useful for the healthcare field.
[16, 17] show that MS-G3D adopt shared weight in the time dimension, therefore limiting the capabilities of correctly representing cross-spacetime features. Their CST-GCN tackle the issue by using unshared weights in spatial and temporal dimensions.

3 Methodology

The methodology proposed in the paper addresses the challenge of skeleton-based action recognition by introducing a novel approach for disentangling and unifying graph convolutions. This section provides a detailed description of the methodological aspects presented in the original paper. The approach consists of several key components, including Disentangled Multi-Scale Aggregation, Disentangling Neighborhoods, G3D (Unified Spatial-Temporal Modeling), and Model Architecture. Each component plays a crucial role in capturing complex joint correlations and long-range temporal dependencies inherent in human actions.

3.1 GCNs

The Graph Convolutional Nets (GCNs) operate on skeleton inputs represented by features X and graph structure A. The layer-wise update rule of GCNs can be applied to features at time t as follows:

$$X_t^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X_t^{(l)}\Theta^{(l)}\right) \tag{1}$$

Here: - $\tilde{A} = A + I$ represents the skeleton graph with added self-loops to preserve identity features. - $\tilde{D}$ is the diagonal degree matrix of $\tilde{A}$. - $\sigma(\cdot)$ denotes an activation function. - The term $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X^{(l)}$ can be interpreted as an approximate spatial mean feature aggregation from the direct neighborhood followed by an activated linear layer. This equation describes how information from neighboring nodes in the skeleton graph is aggregated and transformed using learnable weights $\Theta^{(l)}$ at each layer l. This process allows the model to capture spatial relationships and learn meaningful representations from the skeleton data.

3.2 Disentangled Multi-Scale Aggregation

The biased weighting problem in existing approaches arises from the spatial aggregation framework, where higher-order polynomials of the adjacency matrix are utilized to aggregate multi-scale structural information at time t. This process is represented mathematically as:

$$X^{(l+1)} = \sigma\left(\sum_{k=0}^{K}\hat{A}^{(k)}X_t^{(l)}\Theta_k^{(l)}\right) \tag{2}$$

Here, $X^{(l+1)}$ denotes the features at the $(l+1)$-th layer, $X^{(l)}$ represents the features at the l-th layer, $\Theta^{(l)}$ is the learnable weight matrix at layer l, and σ signifies the activation function. The term $A^{(k)}$ aggregates features across different scales, controlled by the parameter K.

However, this approach leads to biased weighting, where distant neighborhoods exert disproportionate influence on closer ones during the aggregation process. This bias affects the model's ability to effectively capture multi-scale information, as closer neighborhoods may be overshadowed by the influence of distant ones. Thus, there is a need to address this bias to improve the model's performance in capturing complex relationships within the graph structure.

To overcome this issue, we propose a disentangled formulation for the k-adjacency matrix $\tilde{A}(k)$, which captures the dependencies of distant neighborhood's weighting on closer neighborhoods. The k-adjacency matrix $\tilde{A}(k)$ is defined such that it assigns a weight of 1 if there exists a direct edge between nodes v_i and v_j with a distance of k hops, and 0 otherwise.

Mathematically, the k-adjacency matrix $\tilde{A}(k)$ is represented as:

$$\tilde{A}(k) = \begin{cases} 1 & \text{if } d(v_i, v_j) = k, \\ 1 & \text{if } i = j, \\ 0 & \text{otherwise}, \end{cases} \tag{3}$$

where $d(v_i, v_j)$ denotes the shortest distance in terms of the number of hops between nodes v_i and v_j.

Utilized disentangled k-adjacency matrix $\tilde{A}(k)$ ensure that the aggregation of multi-scale structural information is performed in a non-redundant manner, with distant neighborhoods' weighting influencing closer neighborhoods appropriately. This disentanglement of neighborhoods helps in improving the effectiveness and efficiency of skeleton-based action recognition algorithms.

3.3 G3D: Unified Spatial-Temporal Modeling

Unlike existing methods that treat skeleton actions as sequences of disjoint graphs and extract features separately through spatial-only (e.g., GCNs) and temporal-only (e.g., TCNs) modules, G3D aims to capture complex spatial-temporal joint relationships more effectively.

The key insight behind G3D is that strong connections between nodes should incorporate a significant portion of each other's features during layer-wise propagation. However, existing methods weaken signals as they propagate across spacetime through local aggregators like GCNs and TCNs, leading to the aggregation of redundant information from an increasingly larger spatial-temporal receptive field. Additionally, GCNs do not perform weighted aggregation to distinguish each neighbor, exacerbating the problem.

To address these issues, G3D introduces cross-spacetime skip connections, modeled with cross-spacetime edges in a spatial-temporal graph. Specifically, a sliding temporal window of size τ over the input graph sequence, resulting in a spatial-temporal subgraph $G(\tau)$ at each step. Each node within $G(\tau)$ is densely connected to itself and its 1-hop spatial neighbors across all τ frames. Equation 4 defines the construction of the block adjacency matrix A^{τ} for a sliding temporal window of size τ over the input graph sequence. It is represented as:

$$A^{\tau} = \begin{bmatrix} A^{\tau} & \cdots & A^{\tau} \\ \vdots & \ddots & \vdots \\ A^{\tau} & \cdots & A^{\tau} \end{bmatrix} \in \mathbb{R}^{\tau N \times \tau N} \tag{4}$$

Here, each submatrix A^{τ} in the block matrix represents the adjacency matrix for the spatial-temporal subgraph $G(\tau)$. The construction involves tiling the adjacency matrix A^{τ} into the block matrix A^{τ} to capture the temporal connectivity between nodes across the sliding temporal window. This formulation allows for the representation of spatial-temporal relationships within the window by extending the spatial connectivity A^{τ} over time, enabling the modeling of complex spatiotemporal dynamics in skeleton-based action recognition tasks.

The unified spatial-temporal graph convolutional operator for the t-th temporal window is defined as:

$$\left[X_{\tau}^{(l+1)}\right]_t = \sigma\left(\tilde{D}_{\tau}^{-1/2}\tilde{A}_{\tau}\tilde{D}_{\tau}^{-1/2}\left[X_{\tau}^{(l)}\right]_t \Theta^{(l)}\right) \tag{5}$$

where X_{τ} represents the feature tensor for the τ-th temporal window, D_{τ} is the diagonal degree matrix, A_{τ} is the block adjacency matrix, $\Theta^{(l)}$ denotes the learnable weight matrix at layer l, and σ is the activation function. This equation describes the propagation of node features through the spatial-temporal graph convolutional layer, where features are aggregated from neighboring nodes within the τ temporal window according to the spatial-temporal connectivity defined by A_{τ}. The resulting features are then passed through an activation function to introduce non-linearity. G3D facilitates the integration of spatial and temporal information in a unified framework, enabling more effective modeling of complex spatial-temporal joint relationships for improved action recognition performance.

The concept of dilated windows in the context of spatial-temporal modeling allows for the creation of non-adjacent frames within a temporal window, expanding the temporal receptive field without increasing the size of the spatial-temporal structure. By selecting frames at intervals determined by the dilation rate d within a window of τ frames, and reusing the same spatial-temporal structure $\tilde{A}(\tau)$, dilated windows enable the extraction of node features $X(\tau, d)$ over a larger temporal range. Mathematically, this can be represented as:

$$X_{(\tau,d)} \in \mathbb{R}^{T \times \tau \times N \times C}$$

Where T represents the number of frames, N denotes the number of nodes, and C signifies the dimensionality of the feature vectors.

Integration of the disentangled multi-scale aggregation scheme (Eq. 5) into the G3D framework allows for multi-scale reasoning directly within the spatial-temporal domain. This results in the derivation of the MS-G3D module from Eq. 6, represented as:

$$X_{(\tau)}^{(l+1)} = \sigma\left(\sum_{k=0}^{K} \tilde{D}_{(k,\tau)}^{-1/2}\tilde{A}_{(k,\tau)}\tilde{D}_{(k,\tau)}^{-1/2}[X_{\tau}^{(l)}]_t \Theta_{(k)}^{(l)}\right) \tag{6}$$

Where $\tilde{A}_{(k,\tau)}$ and $\tilde{D}_{(k,\tau)}$ are defined analogously to $\tilde{A}_{(k)}$ and $\tilde{D}_{(k)}$ respectively. This formulation ensures that the MS-G3D module incorporates disentangled multi-scale aggregation while leveraging the unified spatial-temporal graph convolutional operator.

3.4 Model Architecture

The final model architecture consists of a hierarchical arrangement of spatial-temporal graph convolutional (STGC) blocks designed to extract discriminative features from skeleton sequences for action recognition. Each STGC block incorporates two primary pathways to capture complex spatial-temporal relationships and long-range dependencies effectively.

The first pathway, known as the G3D pathway, constructs spatial-temporal windows and applies disentangled multi-scale graph convolutions. This pathway is crucial for capturing intricate regional spatial-temporal joint correlations. By utilizing multi-scale graph convolutions, the model can effectively aggregate information across different scales, allowing it to capture both local and global patterns in the skeleton data. The resulting features are then processed using fully connected layers for feature readout.

In parallel, the factorized pathway complements the G3D pathway by incorporating long-range, spatial-only, and temporal-only modules. This pathway starts with a multi-scale graph convolutional layer capable of modeling the entire skeleton graph with different scales. Subsequently, it includes multi-scale temporal convolution layers to capture extended temporal contexts.

To address the limitations of traditional graph convolutional layers, the model introduces adaptive graphs. Learnable graph residual masks, denoted as A_{res}, are incorporated into each $\tilde{A}^{(k)}$ and $\tilde{A}^{(\tau,k)}$. These masks dynamically adjust the edge weights in the graph convolutional layers, allowing for flexible feature propagation and improved model adaptability.

Furthermore, the model incorporates joint-bone two-stream fusion to leverage both joint and bone features for action recognition. This fusion framework utilizes a separate model with an identical architecture, trained using bone features initialized as vector differences of adjacent joints. The final prediction scores are obtained by summing the softmax scores from both joint and bone models.

The overall architecture can be summarized by the final equation:

$$X_{(t)}^{(l+1)} = \sigma\left(\sum_{k=0}^{K} \tilde{D}_{(k)}^{-1/2}(\tilde{A}_{(k)} + A_{(k)}^{\text{res}})\tilde{D}_{(k)}^{-1/2}X_{(t)}^{(l)}\Theta_{(k)}^{(l)}\right) \tag{7}$$

where $X_{(t)}^{(l+1)}$ represents the updated features at time t in the $l+1$-th layer, σ denotes the activation function, $\tilde{A}_{(k)}$ is the disentangled multi-scale adjacency matrix, $A_{(k)}^{\text{res}}$ is the learnable graph residual mask, $\tilde{D}_{(k)}$ is the diagonal degree matrix of $\tilde{A}_{(k)}$, $X_{(t)}^{(l)}$ is the input features, and $\Theta_{(k)}^{(l)}$ represents the learnable weight matrix at layer l of the network for the k-th scale.

This architecture enables the model to effectively extract discriminative features from skeleton sequences, capturing both local and global spatial-temporal patterns while addressing the limitations of traditional graph convolutional

layers through adaptive graph structures and incorporating joint-bone fusion for improved performance.

4 Implementation

The pipeline procedure we used is consistent with what the authors defined. After retrieving the relevant datasets we preprocessed accordingly to be compatible with their models.
We tested the pre-trained weights and have a look on trained models anew to verify their efforts and achievements.
For each dataset, we generate the test scores, which is a heavy computational task, in order to perform all ensembles at once.

4.1 Datasets

In the paper, authors have used three recognized benchmark datasets in the field of skeleton-based action recognition: NTU RGB+D 60, NTU RGB+D 120, and Kinetics Skeleton 400.
To verify their results, we considered the same datasets.
NTU RGB+D 60 Skeleton: Dataset retrieved as paper suggestion from the authors' GitHub repo.[18]. The NTU RGB+D 60 dataset comprises RGB and depth videos capturing human actions performed by multiple subjects across diverse indoor environments. It encompasses 60 action classes, including common activities such as walking, jumping, and waving. Each action instance is recorded from three camera viewpoints, resulting in a total of 60,000 action samples. The dataset is characterized by various challenges, including variations in viewpoint, subject identity, clothing, and environmental conditions, making it well-suited for evaluating the robustness and generalization capabilities of action recognition algorithms.
NTU RGB+D 120 Skeleton: Dataset retrieved as paper suggestion from the authors' GitHub repo.[18]. Building upon the NTU RGB+D 60 dataset, the NTU RGB+D 120 dataset extends the action classes to 120, providing a more comprehensive evaluation framework. With a total of 114,480 action samples, it offers a larger and more diverse set of action instances, covering a broader spectrum of human activities. Similar to its predecessor, NTU RGB+D 120 includes challenges such as occlusions, variations in lighting conditions, and complex background clutter, ensuring a rigorous evaluation of action recognition models across a wider range of scenarios.
Kinetics Skeleton 400: The Kinetics Skeleton 400 dataset is derived from the Kinetics dataset[19], which originally consists of RGB videos depicting human actions in various contexts. By applying advanced pose estimation algorithms to the RGB videos, skeletal sequences representing human movements are extracted, resulting in the Kinetics Skeleton 400 dataset. It comprises 400 action classes, encompassing a diverse array of actions ranging from simple gestures to complex activities. The dataset provides a unique perspective on action recognition by focusing solely on skeletal information, facilitating a deeper understanding of human actions and behaviors.

The NTU RGB+D 60, NTU RGB+D 120, and Kinetics Skeleton 400 datasets provide diverse and challenging datasets for training and evaluating action recognition models. These datasets encompass a wide range of human actions performed in different contexts and environments, making them ideal benchmarks for assessing the performance of skeleton-based action recognition algorithms.
No data augmentation is used, in order to achieve fair performance comparison.

4.2 Hyperparameters

We set the hyperparameters accordingly to the paper reproduced to verify their results.
All models have been trained using SGD and a momentum of 0.9, a batch size of 32 and an initial learning rate of 0.05, and a step LR decay with a factor of 0.1 at epochs 30, 40 for NTU+D 60 Dataset, 30,50 for NTU+D 120 Dataset and 45,55 for Kinetics Skeleton 400. Weight decay is set to 0.0005 for the final models.
All skeleton sequences are padded to T = 300 frames by replaying the actions.
Inputs are pre-processed with normalization and translation following the reference paper.

4.3 Experimental setup

The code used for running the experiments are available on the author's GitHub repository link .

Hardware	Specifications
CPU	12 x Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz
RAM	16 GB DDR4 SDRAM
SSD	2TB SATA 1Gb
GPU	NVIDIA GeForce RTX 2070 Super with Max-Q

Table 1. Experimental setup for the project

Using a different setup from the authors, time is not comparable. Furthermore, as mentioned in their GitHub repository , Out Of Memory (OOM) errors can happen on different machines and setup with respect of PyTorch/CUDA.

4.4 Computational requirements

Throughout our experiments, we monitored the computational resources consumed, including CPU/GPU hours, memory usage, and runtime.
GPU Hours: 67h.
Memory Requirements: 8GB.
Runtime:
Pre-Trained Evaluation Joints

- NTU 60 XSub: 2h 37'
- NTU 60 XView: 8h 14'
- NTU 120 XSub: 4h 35'
- NTU 120 XSet: 9h 28'
- Kinetics Skeleton 400: 2h 37'

Pre-Trained Evaluation Bones

- NTU 60 XSub: 4h 20'
- NTU 60 XView: 4h 12'
- NTU 120 XSub: 9h 30'
- NTU 120 XSet: 11h 18'
- Kinetics Skeleton 400: 8h 41'

5 Results

The following results are taken from the original paper to comparison purposes.
The paper refers to different models in the skeleton-based action recognition task to represent the efficiency of their proposed method compared to the at the time current-state-of-the-art, using the three already mentioned benchmark datasets.
By our reproducibility study, we supported similar results to the original findings.

Methods	X-Sub (%)	X-Set (%)
ST-LSTM[20]	55.7	57.9
GCA-LSTM[21]	61.2	63.3
RotClips+MTCNN[22]	62.2.	61.8
BodyPoseEvolutionMap[23]	64.6	66.9
2s-AGCN[24]	82.9	84.9
MS-G3D Net[1]	86.9	88.4
MS-G3D Net (Ours)	95.0	85.9

Table 2. Classification accuracy comparison against state-of-the-art methods on the NTU RGB+D 120 Skeleton dataset.

Methods	X-Sub (%)	X-View (%)
IndRNN[25]	81.8	88.0
HCN[26]	86.5	91.1
ST-GR[27]	86.9	92.3
AS-GCN[3]	86.8	94.2
2s-ACGCN[24]	88.5	95.1
AGC-LSTM[28]	89.2	95.0
DGNN[29]	89.9	96.1
GR-GCN[30]	87.5	94.3
MS-G3D Net (Joints)[1]	89.4	95.0
MS-G3D Net (Bone Only)[1]	90.1	95.3
MS-G3D Net[1]	91.5	96.2
MS-G3D Net (Ours)	89.3	95.0

Table 3. Classification accuracy comparison against state-of-the-art methods on the NTU RGB+D 60 Skeleton dataset.

Table 2 compares non-graph [20, 21, 22, 23] and graph-based methods [24]. Table 3 compares non-graph methods [25, 26], graph-based methods with spatial edges [3, 24, 27, 28, 29] and with spatial-temporal edges [30]. Table 4 compares single-stream [2, 3] and multi-stream [24, 27, 29] methods.

Our reproduction effort yielded results closely aligned with those reported in the original paper, affirming the robustness and reliability of the MS-G3D Net architecture across multiple benchmark datasets. The slight difference with the results is attributed to the setup, as mentioned in subsection 4.3. In order to run properly without OOM,

Methods	Top-1 (%)	Top-5 (%)
ST-GCN[2]	30.7	52.8
AS-GCN[3]	34.8	56.5
ST-GR[27]	33.6	56.1
2s-AGCN[24]	36.1	58.7
DGNN[29]	36.9	59.6
MS-G3D Net[1]	38.0	60.9
MS-G3D Net (Ours)	37.0	62.0

Table 4. Classification accuracy comparison against state-of-the-art methods on the Kinetics Skeleton 400 dataset.

we used Apex O2 as suggested by the authors. The classification accuracy comparison against state-of-the-art methods, as presented in Tables Table 2, Table 3, and Table 4, demonstrates the superior performance of our implementation.

In Table 2, our MS-G3D Net implementation achieved remarkable classification accuracies of 95.0% for X-Sub and 85.9% for X-Set on the NTU RGB+D 120 Skeleton dataset, surpassing the performance of all other methods considered. Similarly, on the NTU RGB+D 60 Skeleton dataset (Table 3), our implementation demonstrated competitive accuracy, achieving 89.3% for X-Sub and 95.0% for X-View, outperforming several state-of-the-art approaches. Furthermore, Table 4 illustrates the effectiveness of our implementation on the challenging Kinetics Skeleton 400 dataset, where our MS-G3D Net achieved a top-1 accuracy of 37.0% and a top-5 accuracy of 62.0%, showcasing its capability to handle diverse action recognition tasks.

It is worth noting that our reproduction results not only validate the claims made in the original paper but also highlight the consistency and reliability of the proposed methodology across different datasets and experimental conditions.

6 Discussion and conclusion

The contribute of MS-G3D Net is undoubted in the state-of-the-art for skeleton-based action recognition tasks, and recent works demonstrated its applicability especially on the healthcare field.
The paper shows the potential of using to the model advantage the knowledge of the body structure.
In this work, we reiterate their two different methods that builds the MS-G3D architecture: the disentangled multi-scale aggregation scheme for graph convolutions, which removes redundant dependencies between neighborhoods, and G3D, which models spatial-temporal dependencies of from the skeleton graphs.
While our work is limited to reproducibility, we had the opportunity to test their achievements and further learn from more recent papers, which point out different criticalities, as the limitation of cross-spacetime features.
Further studies could consider different hyperparameters to fine-tune the models, and apply the model to other datasets to prove its applicability.

References

[1] Zhenghao Chen Zhiyong Wang Wanli Ouyang Ziyu Liu1, Hongwen Zhang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 143–152, 2020.

[2] Xiong Y. Lin D. Yan, S. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI Conference on Artificial Intelligence*, 2018.

[3] Li W. Cook C. Zhu C. Gao Y. Li, S. Actional-structural graph convolutional networks for skeleton-based action recognition. In *IEEE International Conference on Computer Vision*, 2019.

[4] Zaremba W. Szlam A. LeCun Y. Bruna, J. Spectral networks and locally connected networks on graph. International Conference on Learning Representations, 2014.

[5] Bresson X. Vandergheynst P. Defferrard, M. Convolutional neural networks on graphs with fast localized spectral filtering. Advances in Neural Information Processing Systems, 2016.

[6] Welling M. Kipf, T. N. Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations, 2017.

[7] Cucurull G. Casanova A. Romero A. Lio P. Bengio Y. Velickovic, P. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[8] Hu W. Leskovec J. Jegelka S. Xu, K. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

[9] Boscaini D. Masci J. Rodola E. Svoboda J. Bronstein M. M. Monti, F. Geometric deep learning on graphs and manifolds using mixture model cnns. IEEE, 2017.

[10] Pan S. Chen F. Long G. Zhang C. Yu P. S. Wu, Z. A comprehensive survey on graph neural networks. In *IEEE Transactions on Neural Networks and Learning Systems*, pages 4–24, 2019.

[11] Ryoo M. S. Aggarwal, J. K. Human activity analysis: A review. ACM Computing Surveys, 2011.

[12] Chen Z. Wang H. Yeung D. Y. Shi, X. Skeleton-based action recognition with directed graph neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[13] Shen L. Sun G. Hu, J. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[14] Nie S. Liu Z. He Y. Liu, Y. Skeleton-based action recognition with convolutional neural networks. In *IEEE International Conference on Multimedia and Expo.*, 2018.

[15] Xiong Y. Wang Z. Qiao Y. Lin D. Tang X. Van Gool L. Wang, L. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, 2017.

[16] Q. Liu, M. Cai, D. Liu, and et al. Ess ms-g3d: extension and supplement shift ms-g3d network for the assessment of severe mental retardation. *Complex Intelligent Systems*, 10:2401–2419, 2024. doi: 10.1007/s40747-023-01275-1.

[17] H. Tian, H. Li, W. Jiang, X. Ma, X. Li, H. Wu, and Y. Li. Cross-spatiotemporal graph convolution networks for skeleton-based parkinsonian gait mds-updrs score estimation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:412–421, 2024. doi: 10.1109/TNSRE.2024.3352004. Epub 2024 Jan 19.

[18] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[19] Karen Simonyan Brian Zhang Chloe Hillier Sudheendra Vijayanarasimhan Fabio Viola Tim Green Trevor Back Paul Natsev Mustafa Suleyman Andrew Zisserman Will Kay, Joao Carreira. The kinetics human action video dataset. Computer Vision and Pattern Recognition, 2017.

[20] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, pages 816–833. Springer, 2016.

[21] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, 27 (4):1586–1599, 2017.

[22] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. Learning clip representations for skeleton-based 3d action recognition. *IEEE Transactions on Image Processing*, 27(6):2842–2855, 2018.

[23] Mengyuan Liu and Junsong Yuan. Recognizing human actions as the evolution of pose estimation maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1159–1168, 2018.

[24] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2019.

[25] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[26] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.

[27] Bin Li, Xi Li, Zhongfei Zhang, and Fei Wu. Spatio-temporal graph routing for skeleton-based action recognition. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[28] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1227–1236, 2019.

[29] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7912–7921, 2019.

[30] Jiaxiang Tang Jiaying Liu Xiang Gao, Wei Hu and Zong ming Guo. Optimized skeleton-based action recognition via sparsified graph regression. In *27th ACM International Conference on Multimedia*, pages 601—610. ACM, 2019.