# Skeleton-Based Actions Recognition with Significant Displacements

Chengming Liu[1[0000-0002-8650-4271]], Jiahao Guan[2[0009-0000-0574-0305]] and Haibo Pang[*[ 0000-0001-8832-1500]]

[1,2,*] Zhengzhou University, Zhengzhou, China
panghbzzu@163.com

**Abstract.** In the realm of human skeleton-based action recognition, the graph convolutional networks have proven to be successful. However, directly storing coordinate features into the graph structure presents challenges in achieving shift, scale, and rotation invariance, which is crucial for actions with significant displacements. Such as figure skating, due to the significant displacements of athletes relative to the camera and the inherent perspective effects, leading to variations in scale, position, and rotation-related features. Significant displacements and perspective effects in actions video result in variations in scale, position, and rotation-related features. To address this, drawing inspiration from leveraging high-order information, we propose a novel cosine stream. This stream utilizes the bending angle of human joints for action recognition based on human skeleton. Furthermore, we introduce a new keyframe downsampling algorithm that significantly improves model performance. Notably, our approach does not necessitate any modifications to the backbone. Through extensive experiments on three datasets—FSD-10, FineGYM, and NTU RGB+D, our approach demonstrates improved recognition of actions with significant displacement compared to current mainstream methods.

**Keywords:** Action Recognition, Skeleton, Angle, Figure skating.

## 1 Introduction

Action recognition has become an active research area in recent years, as it plays a significant role in video understanding. Prior investigations have explored various modalities for feature representation, such as RGB frames, optical flows, audio waves, and human skeletons. Among these modalities, skeleton-based action recognition has garnered heightened interest in recent years due to its action-focusing nature and robustness against complicated background. Among the various techniques for skeleton-base action recognition, Graph Convolutional Networks *(GCN)* have been one of the most popular approaches. Yan [10] were pioneers in applying GCN along with temporal convolution to recognize human skeleton-based action. To bolster the capabilities of GCN, recent approaches [8], [11], [12], [13] have aimed to acquire more fitting topologies. However, these GCN based approaches fall short in addressing the challenge of achieving shift, scale, and rotation invariance. When it comes to recognizing actions such as those of motor vehicle drivers and figure skaters, the execution of actions by drivers and skaters may introduce velocities that are not directly correlated with the actions themselves. These velocities can result in translational shifts and scale changes of the performers in the video frames, as illustrated in **Fig. 1**, resulting in variations in the coordinate features of human key points. When the model lacks shift, scale, and rotation invariance, these coordinate changes can interfere with the model's recognition. In this work, our primary focus is on recognizing figure skating actions. Inspired by domain knowledge in figure skating, we leverage novel high-order information extracted from skeleton data to quantify the joint range of motion between two bones. The joint range of motion, typically measured in degrees, provides a valuable metric for assessing joint flexibility and mobility. Inherently linked to posture and movement, it emerges as an inherently discriminative feature for action recognition tasks. We represent the joint range of motion through cosine similarities between pairs of bone vectors, forming the foundation of our cosine stream. By feeding these cosine similarities into a graph convolutional network, we make predictions for action labels. Simultaneously, the cosine stream integrates with the joint-bone two-stream network, giving rise to the development of a comprehensive three-stream network.

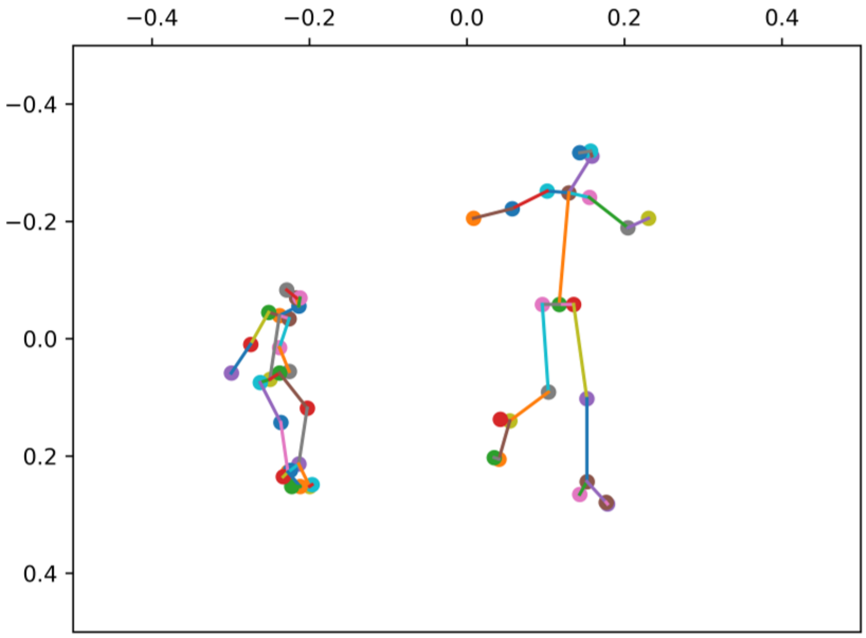

**Fig. 1.** The left side depicts the human body's skeleton at the 35th frame, and the right side shows the skeleton at the 220th frame of the same sample. The original data, normalized without further processing, presents both frames on the same canvas. The smaller appearance of the left skeleton suggests a greater distance from the camera, while the larger right skeleton exhibits noticeable changes in body joint position and rotation. This emphasizes the presence of shift, rotation, and scale variations in the sequence data.

In data processing, contemporary methods encompass two primary steps for preprocessing the provided skeleton sequence for model input. Firstly, in the spatial dimension, the initial frame is designated as the reference, with the skeleton's center point set as the origin, thereby aligning the skeleton's spine with the z-axis. Secondly, addressing the temporal dimension involves implementing diverse solutions to manage inconsistent sequence lengths within the dataset. A recent work by Duan [20] introduces a uniform sampling technique, evenly dividing sequences into N non-overlapping segments with an equal number of frames. One frame is then randomly selected from each segment and aggregated to form a new sub-sequence. While effective, this method overlooks considerations for keyframes. Building upon ideas from Liu [12] and TSN [9], we enhance the approach by simplifying the keyframe selection strategy and integrating it with the original uniform sampling. Various combination strategies have been explored, resulting in particularly significant improvements in the joint stream and bone stream.

Our contributions are summarized as follows:

- We propose a cosine stream, which quantifies the joint range of motion between two bones in degrees, to assist in action recognition with significant displacement.
- We have enhanced the existing downsampling algorithm by integrating the keyframe concept. This enhancement yields substantial improvements in both the joint and bone streams.
- The experimental results indicate that our method can enhance the accuracy performance of the model without necessitating modifications to the network structure itself.

## 2 Related Work

GCN [1], [2], [3], [4], [5], [6] is widely adopted in skeleton-based action recognition. It models human skeleton sequences as spatiotemporal graphs. Yan introduced ST-GCN [10], a widely recognized baseline for GCN-based approaches, which integrates spatial graph convolutions and temporal convolutions to model spatiotemporal data.

Shi [11] bring in adaptive topology of the graph and propose bone stream integrated with joint stream in a two-stream network. Additionally, they proposed extracting motion information using the coordinate differences of the joints and the bones between two consecutive frames, and then combining them into a multi-stream network [8].

Liu [12] have proposed a disentangled multi-scale aggregation scheme aimed at removing redundant dependencies between vertex features from various neighborhoods.

They also introduced a three-dimensional graph convolution operator that facilitates direct information flow across space and time. Chen [13] have proposed a channel-wise topology refinement graph convolution, which dynamically models channel-wise topologies in a refinement approach, leading to flexible and effective correlation modeling.

## 3 Method

### 3.1 Cosine stream

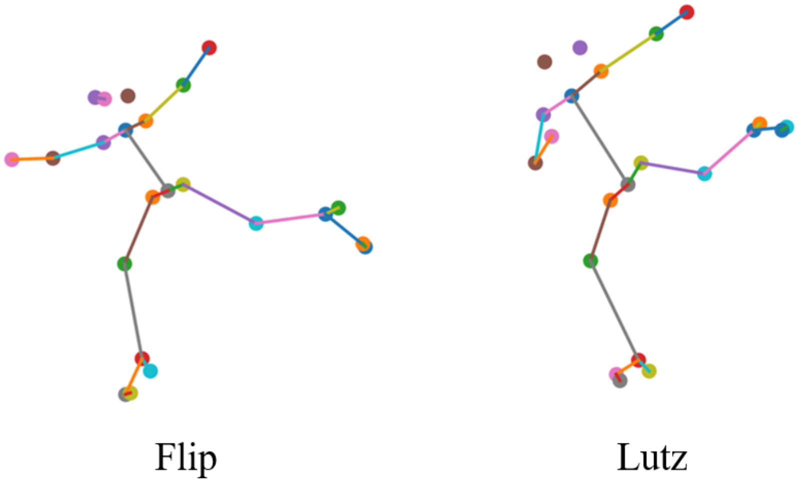

**Fig. 2.** The takeoff phase involves two jumping techniques in figure skating: the flip and the lutz. In the depiction of the flip jump on the left, the skater positions the left foot on the inner edge of the skate blade, shifting the overall body weight to the inside of the blade, with the left arm naturally extended. On the right, showcasing the lutz jump, the only difference is the skater placing the left foot on the outer edge of the skate blade. This results in a relatively outward shift of the body weight. For stability and enhanced takeoff power, skaters typically choose to naturally curve the left hand towards the right side.

In figure skating, athletes often maintain a consistent skating speed during the execution of actions, leading to significant displacement and variations in position features. The angle of the ice skate blade relative to the ice surface, distinguishing between the inside and outside edges, is a crucial factor in action classification (see **Fig. 2**). To preserve blade clarity, slight differences in body joint angles occur, which tend to remain relatively stable during displacement compared to coordinates. Joint angle changes are typically induced by specific actions, making them more discriminative. We aim to input human body joint angles as raw features into the network in the form of cosine similarities. The cosine similarity $\cos v_i$ is calculated using Eq. (1) is the neighborhood of $v_i$,

$A^2_{|N(v_i)|}$is the number of permutations, and $\overrightarrow{e_{ij}}$ is the vector from $v_i$ to $v_j$. Values for vertices in the cosine stream graph are generated, and an empty cosine similarity with a value of 0 is added to the outermost vertices, ensuring consistency in the design of the graph and network of cosine with that of joints and bones.

$$cosv_i = \frac{1}{A^2_{|N(v_i)|}} \sum_{j\in N(v_i)} \sum_{k\in N(v_i)-j} \frac{\overrightarrow{e_{ij}}}{\|\overrightarrow{e_{ij}}\|} \frac{\overrightarrow{e_{ik}}}{\|\overrightarrow{e_{ik}}\|} \quad (1)$$

## 3.2 Keyframe downsampling algorithm

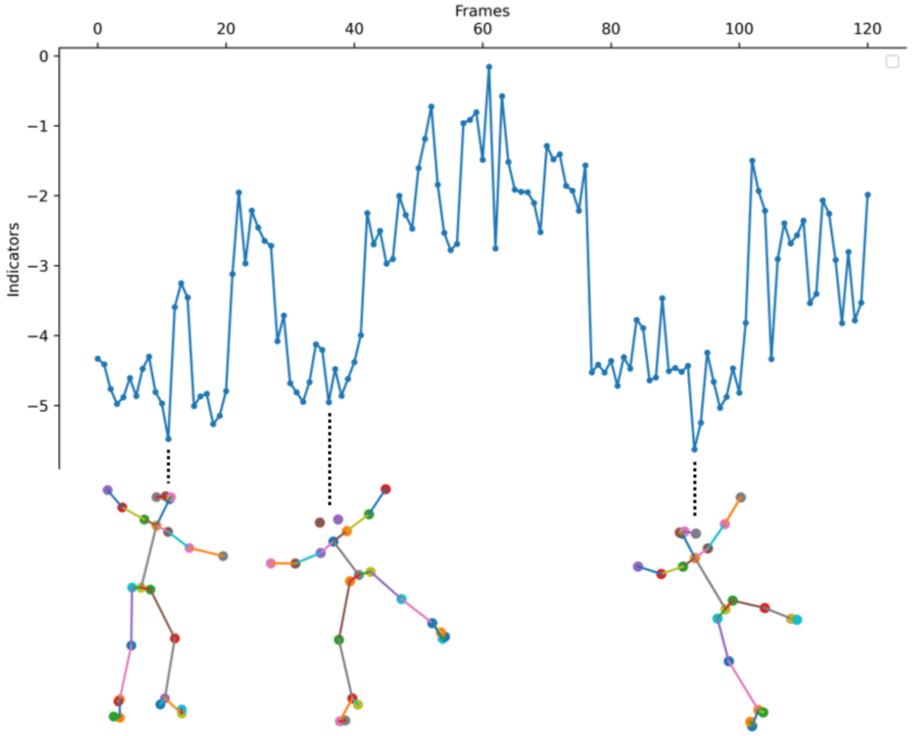

**Fig. 3.** Applying the method introduced in Section 3.2, we obtained downsampled indicators for each frame, along with the skeleton diagrams for frames 11, 36, and 93, corresponding to the minimum values of the indicators. Observing the images, it becomes evident that this downsampling method efficiently highlights frames where the human skeleton is more extended, aiding the model in making accurate assessments of movements.

Sampling keyframes is a crucial aspect of video analysis in figure skating, ensuring that selected frames encapsulate the most discriminative information within a video. In the figure skating task, the fast-changing motion frames are distinctly important for jump action. Through Eq. (1), we transformed the original joint stream into a cosine stream, with attribute values stored in the vertices ranging from -1 to 1. This reflects the joint's range of motion from 180 degrees to 0 degrees, eliminating the need for normalization.

We then sum the angles for all joints in the body to obtain the downsampling indicator, representing the extent of limb extension in each frame. A smaller value indicates a larger sum of angles for various joints in the entire body, implying greater limb extension, as illustrated in **Fig. 3**. Hence, it is immensely beneficial in identifying key frames within sequences. For instance, in a jumping sequence, the indicator during the takeoff phase is smaller than that during the mid-air spinning phase. This suggests that recognizing the takeoff is more crucial than identifying the posture during mid-air action, aligning with the focus on judging actions in figure skating sports. We decided to incorporate Uniform sampling, as introduced by [14], [15], [20], into our proposed keyframe-based downsampling approach, resulting in various fusion strategies:

1. Sort video frames based on the keyframe selection indicator. Choose frames with the smallest indicator to create a new subsequence of M frames, appending it to the N-frame subsequence from uniform sampling.
2. Divide the sequence into M non-overlapping substrings. Select the frame with the smallest indicator from each substring to form a new M-frame subsequence. Connect it to the N-frame subsequence from uniform sampling.
3. Building upon 1, rearrange the generated N+M frames chronologically to create a new downsampling sequence.
4. Building upon 2, rearrange the generated N+M frames chronologically to create a new downsampling sequence.

## 4 Experiments

### 4.1 Datasets

Our work mainly focuses on figure skating actions, corresponding to the FSD-10 dataset. Nevertheless, we conducted experiments on two additional widely used datasets, FineGYM and NTU RGB+D, to evaluate the method's generalization capacity.

**FSD-10.** The Figure Skating Dataset (FSD-10) [7] is a challenging dataset in competitive sports, featuring 1484 figure skating videos labeled with 10 actions. It includes 989 training and 495 testing videos, segmented from around 80 hours of global figure skating championships (2017-2018).

**FineGYM.** FineGYM [16] is a high-quality action recognition dataset with 29k videos and 99 fine-grained gymnastic action classes. Human poses are extracted using GT bounding boxes (provided by [14]).

**NTU RGB+D.** The NTU RGB+D dataset [17] is a human action recognition dataset with 56,880 skeleton sequences from 40 volunteers, categorized into 60 classes. It suggests two evaluation setups: (1) Cross-subject (X-sub), with training from 20 subjects and testing from the remaining 20; (2) Cross-view (X-view), training from views 2 and 3, and testing exclusively from view 1. In our experiments, 2D human poses are estimated using HRNet [18] (provided by [15]).

### 4.2 Implementation Details

All experiments are conducted on one RTX 3090 GPU with PYSKL [15] and MMaction2[19]. Except for downsampling configurations, we used the default hyperparameter settings provided by PYSKL. We employed UniformSampling to sample 100 frames in the FSD-10 dataset and 50 frames in the GYM dataset. Additionally, Keyframesampling was used to sample 25 frames in both datasets.

### 4.3 Ablation study

In this section, we analyze the proposed cosine stream and keyframe sampling algorithm on the FSD-10 dataset. For the cosine stream, we selected three latest and widely recognized skeleton-based action recognition models—AGCN [11], MSG3D [12], and CTRGCN [13]—as baselines. No modifications are required to the network structure for the cosine stream.

For the keyframe downsampling algorithm, we chose the current state-of-the-art model CTRGCN [13] as the baseline and demonstrated its effectiveness on the joint, bone, and cosine streams. As publicly available results for these methods on the FSD-10 dataset were not found, we conducted our experiments on this dataset using networks successfully reproduced from the PYSKL [15] toolbox. The experiments were conducted with the same hyperparameter settings to ensure fairness and consistency in the evaluation.

**Table 1.** Improvement of the Cosine Stream Across Different Models on FSD-10 Dataset, where 'j' and 'b' represent the 'joint stream' and 'bone stream', respectively, 'c' signifies the cosine stream.

| Acc(%) | AGCN [11] | | | | |
|---|---|---|---|---|---|
| | j | b | c | j&b | j&b&c |
| Mean Class | 90.2 | 91.5 | 87.7 | 91.7 | **92.6 ↑** |
| Top1 | 88.9 | 90.4 | 87.1 | 90.8 | **91.5 ↑** |
| Acc(%) | MSG3D [12] | | | | |
| | j | b | c | j&b | j&b&c |
| Mean Class | 90.1 | 90.5 | 89.4 | 90.9 | **91.7 ↑** |
| Top1 | 90.1 | 89.2 | 88.5 | 90.6 | **90.8 ↑** |

**Effectiveness of cosine stream.** We evaluated the cosine stream against three widely used skeleton-based methods (**Table 1**). Despite slightly lower performance compared to joint and bone streams, the cosine stream employs 1D cosine similarity data, while the others use 2D coordinate data. However, the aggregated three-stream model consistently outperforms two-stream methods in Mean Class Accuracy and Top1 Accuracy. In particular, due to the prevalent use of motion features in current state-of-the-art methods, we additionally incorporated experiments involving joint motion, bone motion, and our proposed cosine stream extended to include cosine motion in the temporal dimension in our experiments with CTRGCN. As shown in **Table 2**, we observed that both joint motion and bone motion, as well as cosine motion, performed worse than the spatial dimension feature streams. Furthermore, the fusion of joint motion and bone motion with the original joint-bone dual-stream model only resulted in a modest increase of 0.7% in Mean Class accuracy and 0.5% in Top1 accuracy, while the inclusion of cosine motion brought about negligible improvement. We hypothesize that this may be due to the fact that the actions in the dataset generally involve certain speeds during execution, which are not significantly correlated with the actions themselves, thereby resulting in unsatisfactory performance of the motion features. Therefore, we did not utilize motion features in subsequent experiments.

**Table 2.** Improvement of the Cosine Stream with CTRGCN on FSD-10 Dataset, where 'j', 'b', 'jm', and 'bm' represent the 'joint stream', 'bone stream', 'joint motion stream', and 'bone motion stream', respectively, while 'c' and 'cm' signify the cosine stream and cosine motion stream.

| Acc(%) | CTRGCN [13] | | | | | |
|---|---|---|---|---|---|---|
| | j | jm | b | bm | c | cm |
| Mean Class | 90.9 | 89.5 | 91.5 | 89.6 | 90.1 | 88.6 |
| Top1 | 90.1 | 88.2 | 90.8 | 88.7 | 90.4 | 87.3 |

| Acc(%) | CTRGCN [13] | | | |
|---|---|---|---|---|
| | j&b | j&b&jm&bm | j&b&c | j&b&c&jm&bm&cm |
| Mean Class | 92.5 | 93.2 | **93.5 ↑** | **93.6 ↑** |
| Top1 | 92.0 | 92.5 | **93.2 ↑** | 93.2 |

**Effectiveness of Keyframesampling.** We explored various keyframe sampling and uniform sampling strategies (Section3.2), with results shown in **Table 3**. Strategy analysis reveals that merely increasing downsampled frames (Origin and Strategy 0) may lead to a slight improvement in model performance. Strategy 1 demonstrates that simply adding keyframes can hardly enhance performance. Meanwhile, Strategy 2 introduces segment-wise keyframe selection, leading to enhanced temporal modeling. Strategies 3 and 4 combine uniform sampling with keyframe sampling, differing in their integration approach. Compared to simple concatenation, inserting keyframes from downsampling into the subsequence generated by uniform sampling proves more conducive to modeling, evident in the joint and bone streams. Strategy 4 notably improves Mean Class Accuracy by 1.5% and Top1 Accuracy by 1.4% in the joint stream, and by 2.8% and 2.4% in the bone stream, respectively. If compared to the original uniform sampling algorithm of N+M frames, the improvements in the joint stream are 1.0% and 1.4%, and the improvements in the bone stream are 2.2% and 1.9%.

**Table 3.** Results of CTRGCN on FSD-10, showcasing different keyframe sampling strategies for each stream. 'Origin' and the number 0 denote no keyframe sampling, while uniform sampling frames are set at N and N+M. Numbers 1-4 correspond to strategies detailed in the section 3.2, which combine uniform sampling for N frames and keyframe sampling for M frames.

| Acc(%) | Strategies | | | | | |
|---|---|---|---|---|---|---|
| | j_origin | j0 | j1 | j2 | j3 | j4 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Mean Class | 90.9 | 91.4 | 89.4 | 90.7 | **91.6 ↑** | **92.4 ↑** |
| Top1 | 90.1 | 90.1 | 88.2 | **90.6 ↑** | **90.4 ↑** | **91.5 ↑** |
| Acc(%) | Strategies | | | | | |
| | b_origin | b0 | b1 | b2 | b3 | b4 |
| Mean Class | 91.5 | 92.1 | 89.8 | 91.7 | **92.9 ↑** | **94.3 ↑** |
| Top1 | 90.8 | 91.3 | 89.9 | 91.1 | **92.7 ↑** | **93.2 ↑** |
| Acc(%) | Strategies | | | | | |
| | c_origin | c0 | c1 | c2 | c3 | c4 |
| Mean Class | 90.1 | 89.4 | 89.5 | **90.9 ↑** | 90.1 | **90.6 ↑** |
| Top1 | 90.4 | 88.7 | 88.7 | 90.4 | 88.7 | 89.4 |
| Acc(%) | Strategies | | | | | |
| | j0&b0 | j4&b4 | j4&b4&c4 | | j4&b4&c_origin | |
| Mean Class | 92.5 | **94.3 ↑** | **94.3 ↑** | | 94.6 | |
| Top1 | 91.8 | **93.2 ↑** | **93.2 ↑** | | **94.4 ↑** | |

Analyzing the third sub-table reveals keyframe sampling doesn't improve the cosine stream. This is attributed to the indicator using raw data from the cosine stream, causing redundancy and a slight performance decrease. This confirms that improvements in the first two sub-tables are due to keyframes rather than increased downsampled frames.

Fusing the streams with keyframe sampling (fourth sub-table) involves Strategy 4 for joint and bone streams. Compared to the original joint-bone two-stream model, there is a 1.8% improvement in Mean Class Accuracy and 1.2% in Top1 Accuracy. However, keyframe sampling does not enhance the performance of the cosine stream. On the other hand, when fused with the cosine stream using the original strategy, it results in a 2.1% increase in Mean Class Accuracy and 2.4% in Top1 Accuracy. Furthermore, we observed that fusing the joint stream and bone stream under the uniform sampling algorithm of N+M frames resulted in poorer results compared to the joint dual-stream model under N frames. This finding suggests that simply adding data does not effectively improve the final result.

## 4.4 Cross-dataset Validations

**Table 4.** Enhanced CTRGCN model performance on FineGYM dataset with cosine stream and keyframe sampling.

| Acc(%) | CTRGCN [13] | | | | |
|---|---|---|---|---|---|
| | j | b | c | j&b | j&b&c |
| Mean Class | 88.7 | 91.4 | 85.0 | 92.0 | **92.5 ↑** |
| Top1 | 91.9 | 93.7 | 89.0 | 94.5 | **94.7 ↑** |
| Acc(%) | CTRGCN [13]& Keyframesampling | | | | |
| | j4 | b4 | c4 | j4&b4 | j4&b4&c |
| Mean Class | **89.5 ↑** | **91.5 ↑** | 84.1 | **92.6 ↑** | **93.0 ↑** |

| | | | | | |
|---|---|---|---|---|---|
| Top1 | **92.6 ↑** | **93.9 ↑** | 88.4 | **94.8 ↑** | **95.1 ↑** |

**Table 4** presents the results of introducing the cosine stream on the FineGYM dataset, showcasing a 0.5% improvement in Mean Class Accuracy and a 0.2% improvement in Top-1 Accuracy for the original joint-bone two-stream model. When employing keyframe sampling strategy 4 for both the joint stream and bone stream, the three-stream model exhibited a 1.0% increase in Mean Class Accuracy and a 0.6% increase in Top-1 Accuracy compared to the original joint-bone two-stream model. The observed smaller enhancement is attributed to dataset differences, where the categorization of action classes in the FineGYM dataset may have less correlation with the angular relationships of joints within the body.

**Table 5.** CTRGCN model performance on NTU RGB+D dataset with cosine stream and keyframe sampling. ‘*’ indicates results derived from the pth model downloaded from [15].

| Acc(%) | CTRGCN [13] on X-sub | | | | |
|---|---|---|---|---|---|
| | j | b | c | j&b | j&b&c |
| Top1 | 90.6* | 92.7* | 87.8* | 93.3 | **93.6 ↑** |
| Top1 | 89.3 | 91.6 | 89.0 | 92.3 | **92.9 ↑** |
| Acc(%) | CTRGCN [13]& Keyframesampling on X-sub | | | | |
| | j | b | c | j&b | j&b&c |
| Top1 | 89.9 | 91.4 | 87.3 | 92.5 | 92.8 |
| fanyiAcc(%) | CTRGCN [13] on X-view | | | | |
| | j | b | c | j&b | j&b&c |
| Top1 | 96.9* | 97.5* | 87.0 | 98.4 | 98.4 |
| Top1 | 96.2 | 96.1 | 87.0 | 97.3 | **97.4 ↑** |
| Acc(%) | CTRGCN [13]& Keyframesampling on X-view | | | | |
| | j4 | b4 | c4 | j4&b4 | j4&b4&c |
| Top1 | 95.7 | 95.7 | 87.5 | 97.2 | 97.2 |

In **Table 5**, the results indicate that the enhancements of the model with the cosine stream and keyframe sampling on the NTU RGB+D dataset are not as promising. This holds true for both our experimental results and the data results cited in other works. We attribute this observation to two factors: (1) In terms of the inherent categorization of the dataset, NTU RGB+D places less emphasis on human joint angles. (2) The dataset is collected in a controlled lab environment where subjects do not exhibit significant displacement relative to the camera, leading to a loss of the advantageous shift, scale, and rotation invariance of angle features.

## 5 Conclusion

In this work, we introduce the cosine stream, using joint angles represented by bone vectors as cosine similarity for action prediction. The cosine stream significantly enhances performance in environments sensitive to joint angles or subject movement relative to the camera. Additionally, we propose a keyframe-based downsampling algorithm using cosine values across joints to measure body stretching, improving performance in such environments. The aggregated multi-stream model also benefits from this enhancement.

**Acknowledgments.** This work was supported in part by the national key R&D program of China(2020YFB1712401), the Nature Science Foundation of China (62006210, 62206252), the Key science and technology project of Henan province (221100211200,221100210100), the technological research projects in Henan province (232102210090)

## References

1. Bruna J, Zaremba W, Szlam A, et al. Spectral networks and locally connected networks on graphs[J]. arXiv preprint arXiv:1312.6203, 2013.
2. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering[J]. Advances in neural information processing systems, 2016, 29.
3. Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.
4. Niepert M, Ahmed M, Kutzkov K. Learning convolutional neural networks for graphs[C]//International conference on machine learning. PMLR, 2016: 2014-2023.
5. Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.
6. Duvenaud D K, Maclaurin D, Iparraguirre J, et al. Convolutional networks on graphs for learning molecular fingerprints[J]. Advances in neural information processing systems, 2015, 28.
7. Liu S, Liu X, Huang G, et al. FSD-10: a dataset for competitive sports content analysis[J]. arXiv preprint arXiv:2002.03312, 2020.
8. Shi L, Zhang Y, Cheng J, et al. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks[J]. IEEE Transactions on Image Processing, 2020, 29: 9532-9545.
9. Wang L, Xiong Y, Wang Z, et al. Temporal segment networks for action recognition in videos[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(11): 2740-2755.
10. Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).
11. Shi L, Zhang Y, Cheng J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 12026-12035.

12. Liu Z, Zhang H, Chen Z, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 143-152.
13. Chen Y, Zhang Z, Yuan C, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 13359-13368.
14. Duan H, Zhao Y, Chen K, et al. Revisiting skeleton-based action recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 2969-2978.
15. Duan H, Wang J, Chen K, et al. Pyskl: Towards good practices for skeleton action recognition[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2022: 7351-7354.
16. Shao D, Zhao Y, Dai B, et al. Finegym: A hierarchical video dataset for fine-grained action understanding[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 2616-2625.
17. Shahroudy A, Liu J, Ng T T, et al. Ntu rgb+ d: A large scale dataset for 3d human activity analysis[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1010-1019.
18. Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5693-5703.
19. Contributors M M A. Openmmlab’s next generation video understanding toolbox and benchmark[J]. 2020.
20. Duan H, Wang J, Chen K, et al. Dg-stgcn: Dynamic spatial-temporal modeling for skeleton-based action recognition[J]. arXiv preprint arXiv:2210.05895, 2022.