Class-Aware Contrastive Learning for Fine-Grained Skeleton-Based Action Recognition

Xinyu Bian[1], Dongliang Chang[2⋆], Yuqi Yang[1], Zhongjiang He[3], Kongming Liang[1], and Zhanyu Ma[1]

[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
[2] Department of Automation, Tsinghua University, Beijing, China
[3] China Telecom Artificial Intelligence Technology Co. Ltd, Beijing, China
[1]{bianxinyu, yangyuqi, liangkongming, mazhanyu}@bupt.edu.cn,
[2]changdongliang@pris-cv.cn,[3]hezhongj_1@163.com

Abstract. Graph convolutional networks have significantly advanced skeleton-based action recognition by efficiently processing non-mesh skeleton sequences. However, existing methods struggle with fine-grained action recognition due to the high similarity of samples across categories. In this paper, we propose a class-aware contrastive learning framework designed to emphasize subtle motion feature differences. Our method enhances discriminative capability for fine-grained action recognition by refining negative sample selection in contrastive learning to prioritize samples from similar categories. Furthermore, our framework incorporates global context from multiple sequences during the graph learning process and utilizes memory banks to store rich instance information, enriching cross-sequence context understanding. Our method achieves remarkable performance compared to state-of-the-art methods on the NTU RGB+D, NW-UCLA, and FineGym datasets. Codes are available at: https://github.com/PRIS-CV/Class-Aware-Contrastive-Learning-for-Action-Recognition.

Keywords: Fine-grained action recognition · graph convolutional network · Class-Aware · contrastive learning · cross-sequence

1 Introduction

Skeleton-based action recognition has significant applications and impacts in various fields, including surveillance[25], human-computer interaction[1], augmented reality[10], and healthcare[47]. Traditionally, action recognition relied solely on RGB modalities. However, it faced challenges such as sensitivity to lighting conditions, background clutter, and viewpoint variations. The advent of skeleton-based methods has mitigated these challenges by focusing on the human skeletal structure, providing a more invariant and compact representation of

⋆ indicates corresponding author.

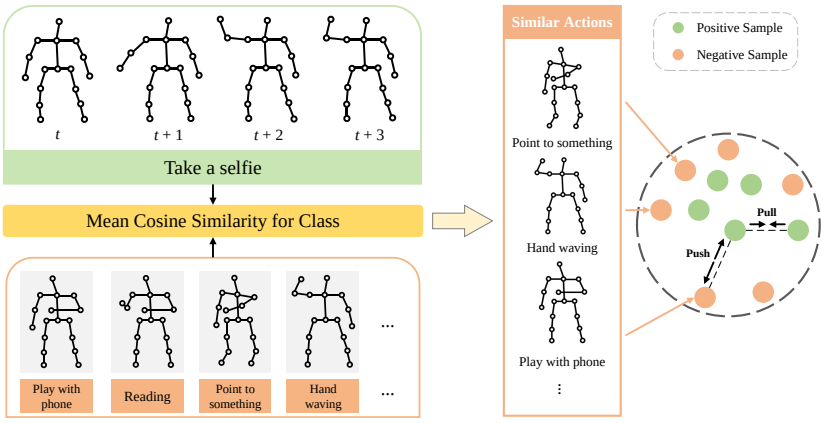

Fig. 1. Our class-aware sampling strategy enhances contrastive learning. By calculating the similarity between classes, we identify the most similar class features, allowing us to select the most effective negative samples for contrastive learning.

human actions. Skeleton-based action recognition offers several advantages over RGB-based methods, including robustness to variations in lighting and viewpoint, as well as reduced computational complexity due to the lower dimensionality of skeletal data. These advantages make skeleton-based action recognition effective in scenarios where consistent and reliable action recognition is essential.

Recently, advancements in graph convolutional networks (GCN) have further propelled the development of skeleton-based action recognition. These networks are adept at modeling the spatial-temporal dependencies inherent in skeletal data, enabling more accurate and efficient analysis of human movements. GCN can effectively capture the complex interactions between different body parts over time by representing the skeleton as a graph, where joints are nodes and bones are edges. This has led to significant improvements in the performance and robustness of skeleton-based action recognition systems.

One of the primary challenges faced by skeleton-based action recognition methods is effectively modeling and capturing spatial and temporal dependencies. Prominent methods such as ST-GCN[45] have pioneered the use of predefined graph structures to capture these relationships. Subsequent research has advanced the extraction of global spatiotemporal features by employing adaptive graphs to dynamically aggregate features within each sequence. However, previous GCN-based methods have inadequately addressed cross-sequence contextual relationships. This limitation hampers their ability to capture subtle differences among similar categories solely through intra-sequence graph learning.

Another major challenge is the issue of class ambiguity arising from insufficient differentiation between similar actions. In fine-grained action recognition tasks, accurately identifying subtle differences between similar categories becomes even more critical. Fine-grained features[11,12,44] often exhibit significant intra-class differences and inter-class similarities, which can lead to mis-

classifications. For example, actions like "reading" and "writing" involve similar movements, making it challenging for models to distinguish between them based solely on local sequence information. Geng *et al*. [17] proposed a self-attention enhanced graph neural network to improve the recognition accuracy of fine-grained actions. Chen *et al*. [3] proposed a novel multi-granular spatiotemporal graph network for skeleton-based action classification that jointly models the coarse- and fine-grained skeleton motion patterns. However, these approaches are limited to the spatio-temporal domain, which can result in the network overlooking crucial distinctions among similar fine-grained actions.

Recently, contrastive learning methods have demonstrated significant potential by focusing on cross-sample relationships within the data. These methods employ skeleton transformation to generate positive and negative samples, thereby optimizing the learning process by bringing positive sample pairs closer together and pushing negative sample pairs further apart. Chen *et al*. [4] proposed SimCLR, which generates positive samples through a series of data augmentation methods such as random cropping, Gaussian blur, and color distortion. On the other hand, He *et al*. [19] applied a memory module that maintains a queue for storing negative samples, with this queue being constantly updated during training. However, these methods are not designed for fine-grained action recognition and face challenges in distinguishing between similar actions.

To tackle these challenges, in this paper, we propose a class-aware contrastive learning framework to leverage cross-sequence context to guide graph learning for skeleton-based fine-grained action recognition. Our method harnesses the power of contrastive learning to enhance the discriminative capability of GCN by explicitly incorporating class-aware information. The core concept is to leverage inter-class differences to guide the learning process, ensuring that representations of similar actions are distinctly separated. Additionally, we utilize a negative sampling strategy that selects samples from significantly different classes, ensuring effective differentiation between similar actions. We also utilize memory banks to store both instance-level and class-level representations, facilitating comprehensive comparisons of actions across the entire dataset and ensuring that the model captures a nuanced and thorough understanding of each action class.

Our contributions are summarized as follows:

(i) We introduce a framework that integrates class-aware information into the contrastive learning process, enhancing the discriminative power of graph convolutional networks for skeleton-based action recognition. Our method utilizes cross-sequence context to enrich representation learning, improving the model's ability to differentiate action categories effectively.

(ii) We propose an innovative negative sampling strategy that selects negative samples from distinctly different classes. This approach ensures robust learning by emphasizing inter-class differences, effectively addressing the challenge of class ambiguity in fine-grained action recognition tasks.

(iii) We integrate memory banks for storing instance-level and class-level representations to comprehensively enrich cross-sequence context, enabling thorough comparisons and ensuring a robust understanding of each action category.

(iv) We extensively experimented on the FineGym fine-grained dataset and two widely used datasets (NTU RGB+D, Northwestern-UCLA) to compare our method with state-of-the-art models. The results demonstrate significant progress in fine-grained action recognition achieved by our method.

2 RELATED WORKS

2.1 Skeleton-Based Action Recognition

Skeleton-based action recognition involves classifying human actions using sequences of key points that represent human joints. Early approaches primarily employed recurrent neural networks (RNN) and convolutional neural networks (CNN). For instance, Du *et al*. [13] and Wang *et al*. [27] applied RNN to extract temporal features from skeleton sequences, but these approaches often struggled to capture spatial dependencies effectively. Conversely, CNN-based methods transformed skeleton data into grid-like representations, simplifying the training process but losing the structured dependencies inherent in the skeleton data. Yan *et al*. [45] pioneered the use of GCN with their ST-GCN model, which employed fixed, heuristically designed graphs to capture spatial-temporal dynamics. This approach spurred further improvements, including multi-scale graph convolutions, channel-decoupled graphs, and adaptive graphs, all aimed at better representing skeleton data over time and space.

Despite these advancements, fine-grained action recognition remains a challenging task due to the subtle differences between actions within the same high-level category. For instance, distinguishing between different types of gymnastic movements requires the model to capture nuanced details often overlooked by coarse-grained action recognition systems. Recent research[29,16] has focused on enhancing feature representation and discrimination capabilities through various means. Chi *et al*. [9] emphasized the importance of focusing on the intrinsic connections between joints, indicating that using only external topology in graph convolution leads to significant inefficiency and information loss in message transmission. Song *et al*. [40] proposed a spatial-temporal attention module that represents action-specific correlations with fewer parameters. Furthermore, Cheng *et al*. [7] introduced a decoupling GCN that enhances the graph modeling ability without incurring additional computational costs.

In contrast to most current GCN methods, which generate graphs based solely on the local context within each sequence and neglect cross-sequence relationships, we propose to explore cross-sequence global context to shape graph representations. The method enables the learned graph to describe distinct features within each sequence while also emphasizing the similarities and differences in motion patterns across sequences.

2.2 Contrastive Learning for Fine-Grained Tasks

Contrastive learning has emerged as a potent technique for acquiring discriminative features by contrasting positive and negative samples, thereby enhancing

intra-class compactness and inter-class separability within the feature space. Initially focused on instance-level discrimination, recent advancements like SimCLR[4] and MoCo[19] have introduced semantic-level memory banks to enrich contextual information and bolster representation learning. These frameworks have demonstrated effectiveness across diverse domains by encouraging models to develop representations that are robust to data augmentations.

Traditional contrastive learning hinges on the judicious selection of positive and negative samples, critical for the efficacy of learned representations. The key idea is to pull together the positive pairs and push away the negative pairs in the feature space. The following approaches[22,43,46,34] learn representations by contrasting positive pairs with negative pairs to ensure that the representations of positive pairs are more similar than those of negative pairs. Typically, positive samples come from the same action class, while negative samples are drawn from different classes. These negative samples are either selected randomly or through hard mining strategies[24,35,23]. To increase the pool of negative samples, a memory bank mechanism was introduced[19], allowing for the storage of a greater number of negative instances. Therefore, selecting informative negative samples is crucial for effective learning, especially for challenging fine-grained action samples.

In the field of skeleton-based action recognition, previous works[33,20,26] have applied contrastive learning to enhance performance. Su *et al*. [41] introduced novel representation learning methods that emphasize motion consistency and continuity. Li *et al*. [28] focused on mining positive pairs within the data space and explored cross-modal distribution relations. Further, Guo *et al*. [18] employed aggressive augmentations to enhance representation universality.

Compared with the above methods, we introduce a class-aware strategy for negative sample sampling. Relative to existing negative sample sampling strategies, we advocate selecting negative samples from classes that exhibit greater similarity. Our strategy directly addresses the primary hurdle in fine-grained action recognition, enabling the model to capture and exploit nuanced differences between similar actions, thereby enhancing overall discriminative capability.

3 METHODOLOGY

3.1 Preliminaries

The human skeleton is represented as a graph with joints as vertices and bones as edges, which is denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the vertex set $\mathcal{V} = \{v_1, v_2, ..., v_N\}$, and N represents the number of vertices. The edge set $\mathcal{E}$ is expressed as the corresponding adjacency matrix $\mathrm{A}_k \in \mathbb{R}^{N \times N}$. For each vertex v_N, the feature dimension is set to C. Therefore, the skeleton feature with T frames can be represented as $\mathrm{X} \in \mathbb{R}^{T \times N \times C}$.

GCN Encoder We use CTR-GCN[5] as the backbone network in practice. The GCN encoder is constructed to extract deep features of X. In general, the graph convolution operation can be expressed as:

$$\mathrm{X}_{out} = \sum_{k=1}^{K_v} \mathrm{A}_k \mathrm{X} \mathrm{W}_k, \tag{1}$$

where $\mathrm{X}_{out} \in \mathbb{R}^{T \times N \times C'}$ represents the output feature with C' channels, $k \in \{1, 2, ..., K_v\}$, K_v represents the number of subgraphs, and $\mathrm{W}_k \in \mathbb{R}^{C \times C'}$ represents the weight matrix that adjusts the number of learnable topological subsets.

The GCN encoder generates two distinct outputs: a feature vector f_c for classification and a graph vector f_g for graph contrastive learning. In the classification stage, the feature vector f_c is projected into a low-dimensional feature space using the MLP layer, resulting in the class prediction. Cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$ is then employed to supervise the category prediction, leveraging the true labels to guide the learning process, as illustrated below:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_i y_i \log \hat{y}_i, \tag{2}$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability for the i^{th} class.

Supervised Contrastive Learning We use a supervised contrastive learning method that utilizes label information to guide the contrastive learning process. It allows each anchor to have multiple positive and negative samples, where the positive samples and the anchor belong to the same class, while the negative samples belong to other categories. We use known labels to construct the positive sample pair $\mathbf{z}$ and $\mathbf{z}_+$ in contrastive learning, and use the class-aware sampling strategy in the subsection 3.3 part to select the negative sample $\mathbf{z}_-$. The negative sample set is represented as N_-. InfoNCE loss is used to bring positive pairs closer in the feature space while pushing negative pairs away, as follows:

$$\mathcal{L}_{\mathrm{NCE}} = -\log \frac{\exp(\mathrm{sim}(\mathbf{z}, \mathbf{z}_+)/\tau)}{\exp(\mathrm{sim}(\mathbf{z}, \mathbf{z}_+)/\tau) + \sum_{\mathbf{z}_- \in N_-} \exp(\mathrm{sim}(\mathbf{z}, \mathbf{z}_-)/\tau)}, \tag{3}$$

where τ is the temperature hyperparameter, $\mathrm{sim}(\cdot,\cdot)$ is the cosine similarity.

3.2 Graph Contrastive Learning

Following the method [21], our approach incorporates graph contrastive learning into both the classification and contrastive learning processes. The overall framework is shown in Figure 2. The GCN encoder outputs a feature vector f_c for classification and a graph f_g for contrastive learning. In the contrastive learning part, we perform feature extraction as shown in Equation 4 and embed the graph into the vector.

$$\mathbf{z}_j = g(f(\mathbf{X}_j)), \tag{4}$$

where $\mathbf{X}_j$ is the j^{th} skeleton sequence, $\mathbf{z}_j$ is the vector that maps the graph to the feature space, g is the graph project head, and f is the GCN encoder.

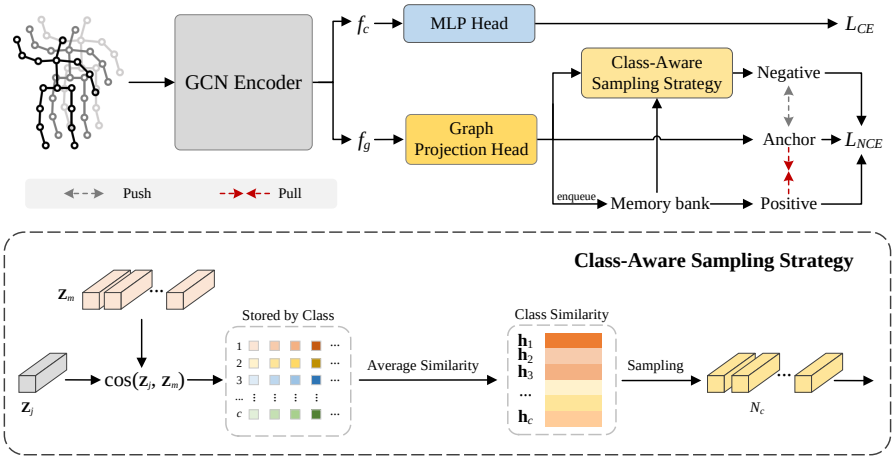

Fig. 2. The framework of our method. The input skeleton sequence is processed by the GCN encoder to generate a feature vector f_c for classification and a graph f_g for graph contrastive learning. The graph projection head embeds f_g into a vector $\mathbf{z}_j$, establishes a memory bank to store instance samples, and selects negative samples of similar categories using the class-aware sampling strategy. In the class-aware strategy, cosine similarity between the input vector $\mathbf{z}_j$ and vectors of other classes $\mathbf{z}_m$ is computed and stored categorically. The class similarity $\mathbf{h}_c$ is derived by averaging these similarities. N_c negative samples are selected from these classes based on their similarity.

Graph Project Head We embed the graph into a vector through the graph projection head. The GCN encoder obtains the graph $f_g \in \mathbb{R}^{K_v \times N \times N \times C}$. The average pooling layer compresses the graph to $\overline{f_g} \in \mathbb{R}^{K_v \times N \times N}$ along the channel dimension. Then, we reduce the dimension of the graph $\overline{f_g}$ to one dimension and finally project the graph into the vector $\mathbf{z}_j$. We use the vector $\mathbf{z}_j$ to guide the subsequent contrastive learning process.

Contrastive Learning Loss To enrich the cross-sequence context, we construct a memory bank to store sample instances. Similar to [19], we design the memory bank $\mathcal{M}$ to store instance features. The features extracted from the data in each batch are stored in the memory bank, and the repository is continuously updated using a first-in-first-out strategy. Meanwhile, referring to Equation 3, InfoNCE loss is used to optimize contrastive learning:

$$\mathcal{L}_{\mathrm{CL}} = -\log \frac{\exp(\mathrm{sim}(\mathbf{z}_j, \mathbf{z}_j)/\tau)}{\exp(\mathrm{sim}(\mathbf{z}_j, \mathbf{z}_{j+})/\tau) + \sum_{j-=1}^{N_c} \exp(\mathrm{sim}(\mathbf{z}_j, \mathbf{z}_{j-})/\tau)}, \tag{5}$$

where $\mathbf{z}_{j+}$ is a positive sample vector with the same label as $\mathbf{z}_j$, $\mathbf{z}_{j-}$ is negative sample vector, N_c is the number of negative samples.

Therefore, our overall loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CL}} + \mathcal{L}_{\mathrm{CE}}. \tag{6}$$

3.3 Class-Aware Sampling strategy

In the process of contrastive learning, the selection of negative samples is crucial. As training proceeds, the contribution of most samples to the model gradually decreases. To effectively handle the problem of small inter-class differences in fine-grained action recognition, we propose a class-aware sampling strategy. This strategy enhances the discriminative ability of the model by selecting category samples with greater similarity as negative samples.

Sample Similarity We construct a set of negative samples from the training data, which are samples that are different from the anchor class. These samples are stored in a memory bank $\mathcal{M}$ to facilitate the comparison of a larger number of sample features. For each anchor sample, we calculate its similarity score with all other class samples. The similarity score can be calculated by the following formula:

$$\text{sim}(\mathbf{z}_j, \mathbf{z}_m) = \frac{\phi(\mathbf{z}_j) \cdot \phi(\mathbf{z}_m)}{\|\phi(\mathbf{z}_j)\| \|\phi(\mathbf{z}_m)\|}, \tag{7}$$

where $\phi(\cdot)$ represents the feature representation of sample, $\mathbf{z}_m$ is the m^{th} other class sample vector, $\cdot$ denotes the dot product, and $\|\cdot\|$ represents the vector norm.

Class Similarity After obtaining the similarity scores for all samples from the other classes, we calculate the average similarity score for each class. The average similarity score $\mathbf{h}_c$ for a class c is calculated as follows:

$$\mathbf{h}_c = \frac{1}{K} \sum_{n=1}^{K} (\text{sim}(\mathbf{z}_j, \mathbf{z}_n)), \tag{8}$$

where K represents the total number of negative samples from class c, and $\mathbf{z}_n$ represents the vector for the n^{th} negative sample of a class c distinct from the anchor class.

High-Similarity Class Sampling We then sort the classes based on their average similarity scores and select the classes with the higher average similarity. From these high-similarity classes, we sequentially select N_c negative samples to conduct contrastive learning. We hope to let the model compare the differences between more similar samples more, and appropriately discard some easily distinguishable samples. Through the implementation of a class-aware sampling strategy in contrastive learning, we promote the closer clustering of positive samples while increasing the separation of negative samples, improving the overall discriminative power of the model.

4 EXPERIMENTS

4.1 Datasets

FineGym FineGym[37] is a large-scale fine-grained action recognition dataset with 29,000 videos of 99 fine-grained gymnastic action categories. Each sequence is annotated as skeletons with 17 joints. In particular, it provides temporal annotations at both action and sub-action levels with a three-level semantic hierarchy.

FineGym collects 10 different event categories in the field of gymnastics and performs fine-grained annotations on four events for women (vault, balance beam, floor exercise, and uneven bars). Based on these four event categories, FineGym defines and screens 15 group categories, and further defines 530 different element categories, of which 354 categories have sub-action data. We follow the method[14] to extract the skeleton data from the 2D pose estimator.

NTU RGB+D NTU RGB+D(NTU60)[36] is a large-scale skeleton-based action recognition dataset, comprising 56,880 action sequences collected from 60 different action classes. Each sequence is annotated as skeletons with 25 joints. The sequences were performed by 40 subjects and captured by three Microsoft Kinect v2 cameras from different views. Generally, two evaluation protocols are employed to assess performances: (1) cross-subject (X-Sub): train data are performed by 20 subjects, and test data are performed by other 20 subjects. (2) cross-view (X-View): train data comes from cameras 2 and 3, and test data comes from camera 1.

Northwestern-UCLA Northwestern-UCLA (NW-UCLA)[42] is a widely used dataset for action recognition, consisting of 1,494 sequences from 10 action categories. Each sequence is annotated as skeletons with 20 joints. The sequences were performed by 10 subjects and captured by three Kinect cameras from different views. We follow the official evaluation protocol, where the train data are obtained from the first two cameras and the test data are collected from the third camera.

4.2 Implementation details

To comprehensively validate the effectiveness of our method, we adopt CTR-GCN[5] as the baseline model. Our method is implemented using the PyTorch deep learning framework. We train our model with SGD, employing a weight decay of 0.0004. The batch size is set to 64 on the FineGym and NTU60 datasets, and 16 on the NW-UCLA dataset. The base learning rate is set to 0.1. The temperature hyperparameter τ is set to 0.8. For all datasets, we set the number of positive samples N_+ to 128 and the number of negative samples N_c to 512 during training.

For FineGYM, the learning rate decays with a factor of 0.1 at epoch 75 and 115 for 180 epochs. For NTU60, the learning rate decays with a factor of 0.1 at

Table 1. Comparison of classification accuracy with state-of-the-art methods on the FineGym and NTU60 datasets.

Methods	Publication	FineGym	NTU60 X-Sub	NTU60 X-View
ST-GCN[45]	AAAI 2018	36.4	81.5	88.3
2s-AGCN[38]	CVPR 2018	-	88.5	95.1
Shift-GCN[8]	CVPR 2020	-	90.7	96.5
MS-G3D[31]	CVPR 2020	92.0	91.5	96.2
DC-GCN+ADG[7]	ECCV 2020	-	90.8	96.6
MST-GCN[6]	AAAI 2021	-	91.5	96.6
CTR-GCN[5]	ICCV 2021	91.9	92.4	96.8
EfficientGCN-B4[40]	TPAMI 2022	-	91.7	95.7
InfoGCN[9]	CVPR 2022	92.0	92.7	**96.9**
FR-Head[48]	CVPR 2023	-	92.8	96.8
Ske2Grid[2]	ICML 2023	91.8	90.1	95.1
Ours		**94.0**	**92.8**	96.7

Table 2. Comparison of classification accuracy with state-of-the-art methods on NW-UCLA dataset.

Methods	Publication	NW-UCLA		
		Joint	Bone	Joint+Bone
2s-AGCN[38]	CVPR 2018	92.0	92.2	95.0
AGC-LSTM[39]	CVPR 2019	93.3	-	-
Shift-GCN[8]	CVPR 2020	92.5	-	94.2
DC-GCN+ADG[7]	ECCV 2020	-	-	95.3
CTR-GCN[5]	ICCV 2021	94.6	91.8	94.2
InfoGCN[9]	CVPR 2022	94.0	**95.3**	**96.3**
TD-GCN[30]	TMM 2023	94.8	93.5	-
Ours		**94.8**	92.5	**96.3**

epoch 35 and 55 for 65 epochs. For NW-UCLA, the learning rate decays with a factor of 0.1 at epoch 50 for 65 epochs. All experiments are conducted using a single NVIDIA 1080 GPU.

Comparsion methods We compare our method with the state-of-the-art methods on the FineGym, NTU60 and NW-UCLA datasets. To demonstrate the effectiveness of our method on spatial-temporal joints and large-scale classes, we conduct fine-grained tasks on the FineGym dataset. To show the generality of our method, we perform experiments on the NTU60 and NW-UCLA datasets.

Following most mainstream methods[45,5,15,9], we evaluate the model using four modalities: joint, bone, joint motion, and bone motion. The 4-stream ensemble denotes the simultaneous use of these four modalities. We use the 4-stream ensemble on the NTU60 dataset to obtain the final experimental results.

4.3 Results and Analysis

Performance on FineGym Dataset In Table 1, our class-aware contrastive learning method outperforms most GCN methods and achieves state-of-the-art performance on the FineGYM dataset. **Ours** method improves by 2.1% over the baseline, demonstrating its superior capability in handling fine-grained action recognition tasks and effectively addressing the challenges of fine-grained actions.

Performance on NTU60 Dataset The results on the NTU60 dataset are presented in Table 1. **Ours** method demonstrates a significant improvement in the X-Sub setting, outperforming most previous approaches and achieving a 0.4% increase over the baseline. We also achieve good results in the X-View setting; however, our improvements are limited as our method primarily focuses on fine-grained features. As a solution primarily designed for fine-grained action recognition, our method significantly enhances the ability to distinguish fine-grained actions while also maintaining robust performance on coarse-grained datasets. This balance underscores the generalization capability and effectiveness of our method across different datasets.

Performance on NW-UCLA Dataset Table 2 shows the results of different modalities on the NW-UCLA dataset. **Ours** method achieves a significant improvement using only joint modality, outperforming the state-of-the-art method by 0.2%. Our performance on the bone modality is slightly lower than that of InfoGCN. It can be attributed to the introduction of attention-based graph convolution in InfoGCN, which enables it to capture more detailed bone information. In contrast, our method emphasizes the differences between features. Additionally, when combining joint and bone modalities, **Ours** method achieves a 2.1% improvement over the baseline, consistent with state-of-the-art results. These experiments demonstrate the strong generalizability of our method.

Overall, on all datasets, our method outperforms all existing methods under nearly all evaluation benchmarks. Our model achieves state-of-the-art performance on the fine-grained dataset FineGym, while also maintaining good performance on the coarse-grained dataset.

5 Future Analysis

5.1 Visualization Results

To qualitatively evaluate our proposed method, we apply t-SNE [32] to visualize the embedding distribution of our method compared to the baseline across three datasets. Figure 3 shows the comparison of the results between them. It demonstrates that our method effectively increases the distance between samples from different classes. This effect is particularly pronounced on the FineGym dataset. It can be seen there are similar samples that are difficult to distinguish in the baseline visualization. Compared with ours, it can not only accurately distinguish

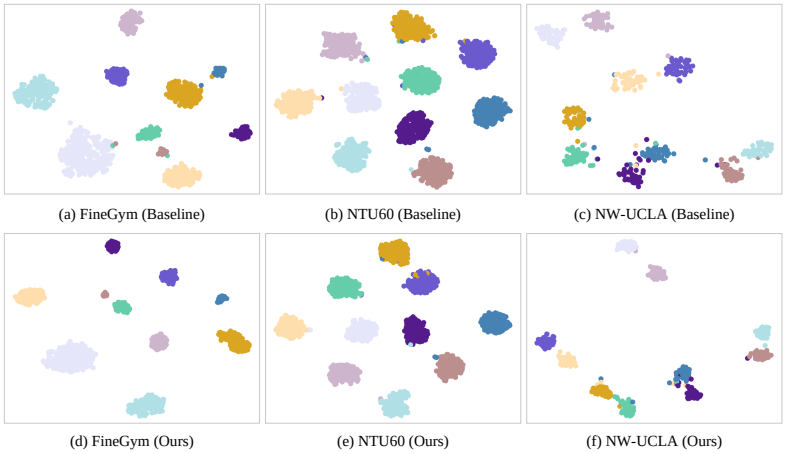

Fig. 3. Visualization of features by t-SNE on the FineGym, NTU60, and NW-UCLA datasets. Different colors represent different classes.

similar categories, but also enhance the clustering effect. The above visualization shows that our method achieves more discriminative representations.

5.2 Ablation Study

Table 3 shows the results of our method on the FineGym dataset under different strategies. We only use the joint input modality in the experiments with CTR-GCN as the GCN encoder.

Contrastive Learning Sampling Strategy In our experiments, we first evaluate the effectiveness of combining contrastive learning methods. Using contrastive learning methods helps to improve the performance of the baseline, with a performance improvement of 1.7%. Next, we compare the impact of different negative sample sampling strategies on the experiment. "R" and "H" are random sampling and hard sampling, respectively, and "CA" is the class-aware sampling we proposed. The experimental results show that the class-aware sampling strategy is more effective in negative sample selection and is superior to the other two sampling methods. Moreover, random sampling involves inherent randomness, and we find that incorporating a portion of random samples into our training process is meaningful for recognition. Therefore, we combine hard sampling and class-aware sampling with random sampling, respectively, and observe that random sampling and class-aware sampling achieve the best result.

Impact of the size of N_c In Table 4, we investigate the impact of selecting N_c negative samples. Given the crucial role of negative samples in contrastive learning, their quantity significantly influences training effectiveness. We opt to

Table 3. Comparison of classification accuracy when different negative sample sampling strategies are applied to the baseline.

Methods	Accuracy(%)
Baseline (w/o Contrast)	91.9
Contrast w/ R	93.6
Contrast w/ H	93.6
Contrast w/ CA	93.8
Contrast w/ R+H	93.7
(Ours) Contrast w/ R+CA	**94.0**

Table 4. Comparison of classification accuracy with different numbers of negative samples.

Sampling	N_+	N_c	Accuracy(%)
CA	128	128	76.2
		256	93.7
		512	**93.8**

utilize only the class-aware sampling and select 128 positive samples and evaluate N_c values of 128, 256, and 512. The results indicate that incorporating additional negative samples can enhance action recognition performance to a certain extent. To achieve the best performance, we set N_c to 512.

5.3 Cross-Semantic Analysis of Action Recognition

According to the experimental results, the method has demonstrated strong capabilities in fine-grained action recognition. We aim to further analyze the applicability of the method in coarse-grained action recognition. Considering the structural complexity of the FineGym dataset, we conduct experiments at various semantic levels to further evaluate the coarse-grained action recognition capabilities of our method. Specifically, we segment the dataset into event and group categories based on the semantic hierarchy within FineGym. These experiments are designed to rigorously test the effectiveness of our method across different levels of semantic granularity.

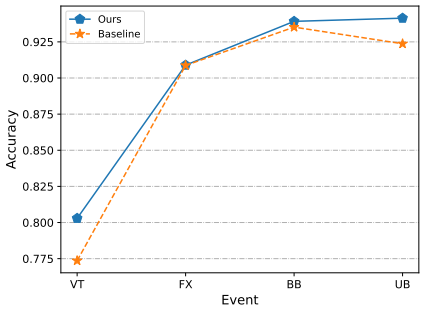

Fig. 4. Comparison of action recognition accuracy with different events on FineGym dataset with the baseline. "VT" stands for vault, "FX" for floor exercise, "BB" for balance beam, and "UB" for uneven bars.

Analysis of Event Categories As shown in Figure 4, we divide the dataset into 4 event categories. The proposed method demonstrates outstanding performance across these different event categories. Remarkably, for vault and uneven bars events, which include more similar sub-actions, such as "stretched salto backward with 1 turn off" and "stretched salto backward with 2 turn off", our method significantly outperforms the baseline.

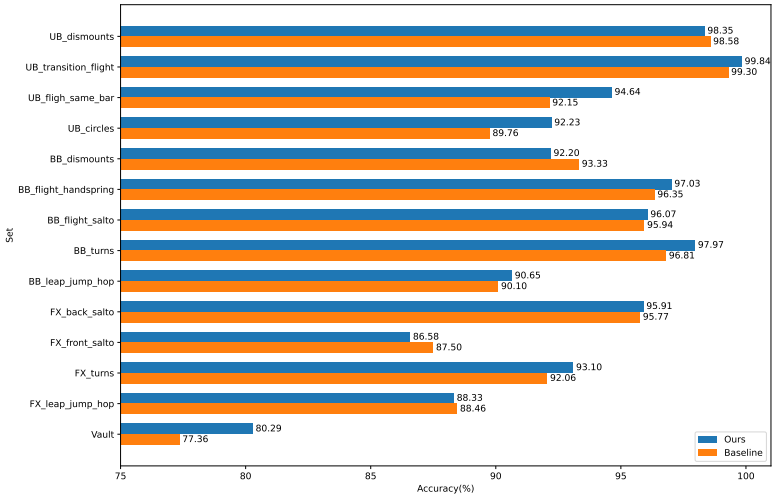

Fig. 5. Comparison of action recognition accuracy with different sets on FineGym dataset with the baseline.

Analysis of Set Categories In Figure 5, we divide the dataset into 14 set categories. Our method achieves significant improvements in most sets and maintains consistent performance in a few others. These results demonstrate the strong competitiveness of our method for coarse-grained action recognition.

6 Conclusion

This paper proposes a class-aware contrastive learning method for fine-grained skeleton action recognition. The method leverages contrastive learning to explore the rich semantic context across sequences and uses a category-aware negative sampling strategy to identify differences between fine-grained actions. This expands the distance between similar classes and enhances the extraction of subtle features. Experimental results demonstrate the significant performance of our method and confirm its effectiveness.

Acknowledgement

This work was supported by the National Nature Science Foundation of China (Grant 62225601, U23B2052, 62406171, 62476029), in part by the Beijing Natural Science Foundation Project No. L242025, in part by the Youth Innovative Research Team of BUPT No. 2023YQTD02, in part by the China Postdoctoral Science Foundation No. 2023M741961, and in part by the Postdoctoral Fellowship Program of CPSF No. GZB20240359.

References

1. Bandi, C., Thomas, U.: Skeleton-based action recognition for human-robot interaction using self-attention mechanism. In: FG. pp. 1–8 (2021)
2. Cai, D., Kang, Y., Yao, A., Chen, Y.: Ske2grid: skeleton-to-grid representation learning for action recognition. In: ICML. pp. 3431–3441 (2023)
3. Chen, T., Zhou, D., Wang, J., Wang, S., Guan, Y., He, X., Ding, E.: Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In: ACM MM. pp. 4334–4342 (2021)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607 (2020)
5. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: ICCV. pp. 13359–13368 (2021)
6. Chen, Z., Li, S., Yang, B., Li, Q., Liu, H.: Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In: AAAI. vol. 35, pp. 1113–1122 (2021)
7. Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J., Lu, H.: Decoupling gcn with dropgraph module for skeleton-based action recognition. In: ECCV. pp. 536–553 (2020)
8. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-based action recognition with shift graph convolutional network. In: CVPR. pp. 183–192 (2020)
9. Chi, H.g., Ha, M.H., Chi, S., Lee, S.W., Huang, Q., Ramani, K.: Infogcn: Representation learning for human skeleton-based action recognition. In: CVPR. pp. 20186–20196 (2022)
10. Cormier, M., Schmid, Y., Beyerer, J.: Enhancing skeleton-based action recognition in real-world scenarios through realistic data augmentation. In: WACV. pp. 290–299 (2024)
11. Du, R., Xie, J., Ma, Z., Chang, D., Song, Y.Z., Guo, J.: Progressive learning of category-consistent multi-granularity features for fine-grained visual classification. IEEE TPAMI **44**(12), 9521–9535 (2021)
12. Du, R., Yu, W., Wang, H., Lin, T.E., Chang, D., Ma, Z.: Multi-view active fine-grained visual recognition. In: ICCV. pp. 1568–1578 (2023)
13. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: CVPR. pp. 1110–1118 (2015)
14. Duan, H., Wang, J., Chen, K., Lin, D.: Pyskl: Towards good practices for skeleton action recognition. In: ACM MM. pp. 7351–7354 (2022)
15. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: CVPR. pp. 2969–2978 (2022)
16. Fu, M., Zheng, Y., Chang, D., Li, W., Ma, Z.: Multi-frequency feature enhancement for multi-granularity visual classification. In: APSIPA. pp. 484–489. IEEE (2023)
17. Geng, P., Lu, X., Hu, C., Liu, H., Lyu, L.: Focusing fine-grained action by self-attention-enhanced graph neural networks with contrastive learning. IEEE TCSVT (2023)
18. Guo, T., Liu, H., Chen, Z., Liu, M., Wang, T., Ding, R.: Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In: AAAI. vol. 36, pp. 762–770 (2022)
19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020)

20. Hu, J., Hou, Y., Guo, Z., Gao, J.: Global and local contrastive learning for self-supervised skeleton-based action recognition. IEEE TCSVT (2024)
21. Huang, X., Zhou, H., Wang, J., Feng, H., Han, J., Ding, E., Wang, J., Wang, X., Liu, W., Feng, B.: Graph contrastive learning for skeleton-based action recognition. arXiv preprint arXiv:2301.10900 (2023)
22. Isola, P., Zoran, D., Krishnan, D., Adelson, E.H.: Learning visual groups from co-occurrences in space and time. arXiv preprint arXiv:1511.06811 (2015)
23. Kalantidis, Y., Sariyildiz, M.B., Pion, N., Weinzaepfel, P., Larlus, D.: Hard negative mixing for contrastive learning. Advances in neural information processing systems **33**, 21798–21809 (2020)
24. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
25. Kim, S., Yun, K., Park, J., Choi, J.Y.: Skeleton-based action recognition of people handling objects. In: WACV. pp. 61–70 (2019)
26. Li, D., Tang, Y., Zhang, Z., Zhang, W.: Cross-stream contrastive learning for self-supervised skeleton-based action recognition. Image and Vision Computing **135**, 104689 (2023)
27. Li, L., Zheng, W., Zhang, Z., Huang, Y., Wang, L.: Skeleton-based relational modeling for action recognition. arXiv preprint arXiv:1805.02556 **1**(2), 3 (2018)
28. Li, L., Wang, M., Ni, B., Wang, H., Yang, J., Zhang, W.: 3d human action representation learning via cross-view consistency pursuit. In: CVPR. pp. 4741–4750 (2021)
29. Li, X., Yu, L., Cao, J., Chang, D., Ma, Z., Liu, N.: Small-sample image classification method of combining prototype and margin learning. In: APSIPA. pp. 91–95. IEEE (2019)
30. Liu, J., Wang, X., Wang, C., Gao, Y., Liu, M.: Temporal decoupling graph convolutional network for skeleton-based gesture recognition. IEEE TMM **26**, 811–823 (2023)
31. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: CVPR. pp. 143–152 (2020)
32. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
33. Mao, Y., Zhou, W., Lu, Z., Deng, J., Li, H.: Cmd: Self-supervised 3d action representation learning with cross-modal mutual distillation. In: ECCV. pp. 734–752 (2022)
34. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
35. Robinson, J., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. arXiv preprint arXiv:2010.04592 (2020)
36. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: CVPR. pp. 1010–1019 (2016)
37. Shao, D., Zhao, Y., Dai, B., Lin, D.: Finegym: A hierarchical video dataset for fine-grained action understanding. In: CVPR. pp. 2616–2625 (2020)
38. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: CVPR. pp. 12026–12035 (2019)
39. Si, C., Chen, W., Wang, W., Wang, L., Tan, T.: An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: CVPR. pp. 1227–1236 (2019)

40. Song, Y.F., Zhang, Z., Shan, C., Wang, L.: Constructing stronger and faster baselines for skeleton-based action recognition. IEEE TPAMI **45**(2), 1474–1488 (2022)
41. Su, Y., Lin, G., Wu, Q.: Self-supervised 3d skeleton action representation learning with motion consistency and continuity. In: ICCV. pp. 13328–13338 (2021)
42. Wang, J., Nie, X., Xia, Y., Wu, Y., Zhu, S.C.: Cross-view action modeling, learning and recognition. In: CVPR. pp. 2649–2656 (2014)
43. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR. pp. 3733–3742 (2018)
44. Xu, S., Chang, D., Xie, J., Ma, Z.: Grad-cam guided channel-spatial attention module for fine-grained visual classification. In: MLSP. pp. 1–6. IEEE (2021)
45. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI (2018)
46. Ye, M., Zhang, X., Yuen, P.C., Chang, S.F.: Unsupervised embedding learning via invariant and spreading instance feature. In: CVPR. pp. 6210–6219 (2019)
47. Yin, J., Han, J., Wang, C., Zhang, B., Zeng, X.: A skeleton-based action recognition system for medical condition detection. In: BioCAS. pp. 1–4. IEEE (2019)
48. Zhou, H., Liu, Q., Wang, Y.: Learning discriminative representations for skeleton based action recognition. In: CVPR. pp. 10608–10617 (2023)