Part Aware Contrastive Learning for Self-Supervised Action Recognition

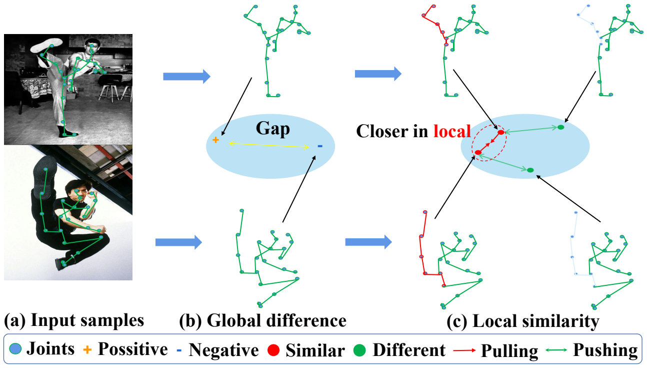
Yilei Hua[1] , **Wenhan Wu**[2] , **Ce Zheng**[3] , **Aidong Lu**[2] , **Mengyuan Liu**[4] , **Chen Chen**[3] , **Shiqian Wu**[1]
[1]School of Information Science and Engineering, Wuhan University of Science and Technology
[2]University of North Carolina at Charlotte
[3]Center for Research in Computer Vision, University of Central Florida
[4]Peking University, Shenzhen Graduate School
{hyl1997, shiqian.wu}@wust.edu.cn, {wwu25, aidong.lu}@uncc.edu, cezheng@knights.ucf.edu,
chen.chen@crcv.ucf.edu, nkliuyifang@gmail.com

Abstract

In recent years, remarkable results have been achieved in self-supervised action recognition using skeleton sequences with contrastive learning. It has been observed that the semantic distinction of human action features is often represented by local body parts, such as legs or hands, which are advantageous for skeleton-based action recognition. This paper proposes an attention-based contrastive learning framework for skeleton representation learning, called SkeAttnCLR, which integrates local similarity and global features for skeleton-based action representations. To achieve this, a multi-head attention mask module is employed to learn the soft attention mask features from the skeletons, suppressing non-salient local features while accentuating local salient features, thereby bringing similar local features closer in the feature space. Additionally, ample contrastive pairs are generated by expanding contrastive pairs based on salient and non-salient features with global features, which guide the network to learn the semantic representations of the entire skeleton. Therefore, with the attention mask mechanism, SkeAttnCLR learns local features under different data augmentation views. The experiment results demonstrate that the inclusion of local feature similarity significantly enhances skeleton-based action representation. Our proposed SkeAttnCLR outperforms state-of-the-art methods on NTURGB+D, NTU120-RGB+D, and PKU-MMD datasets.The code and settings are available at this repository: https://github.com/GitHubOfHyl97/SkeAttnCLR

1 Introduction

With the advancements in human pose estimation algorithms [Cao *et al.*, 2017; Sun *et al.*, 2019; Zheng *et al.*, 2023; Zheng *et al.*, 2022], skeleton-based human action recognition has emerged as an important field in computer vision. However, traditional supervised learning methods [Zhang *et al.*, 2020; Plizzari *et al.*, 2021] require extensive labeled data, resulting in significant human effort. Thus, self-supervised learning has gained attention due to its ability to learn representations from unlabeled data. Self-supervised learning has shown success in natural language [Kim *et al.*, 2021] and vision [He *et al.*, 2022], leading researchers to explore self-supervised learning pre-training for human skeleton-based action recognition.



(a) Input samples (b) Global difference (c) Local similarity

Figure 1: The motivation of our method. Although input skeletons (a) belong to the same movement category, there exists a significant gap between their skeleton sequences in feature space in (b). However, by considering local feature similarity, as shown by the red points in (c), the distance between two actions of the same semantic category becomes shorter in feature space. Therefore, we aim to extend local similarity-based contrastive learning to complement global contrastive learning, in order to bring samples with similar local features closer in feature space. Through the attention mechanism, the local similarity is discriminated and the network focuses on the changes in human action parts, such as leg movements (as highlighted in red). This approach is expected to be more conducive to learning local representations that are beneficial for accurate action recognition.

The current skeleton-based contrastive learning framework has been developed from image-based methods, with researchers exploring cross-view learning [Li *et al.*, 2021a] and data augmentation [Guo *et al.*, 2022; Zhang *et al.*, 2022]. In this work, we focus on instance discrimination via contrastive learning for self-supervised representation learning. Our proposed method emphasizes learning the representation of local features and generating more contrastive pairs from various local actions within a sample to improve contrastive learning performance. As depicted in Figure 1, such action samples are often challenging due to significant global representation differences, resulting in a considerable gap in the feature space. However, the distance between local features in the same action category is closer in the feature space, leading to semantic similarity. Local actions often determine semantic categories, making it desirable to consider local similarities in contrastive learning. In this study, we propose to improve previous works by addressing the following: 1) How to learn the relationship between local features and global features of human actions in skeleton-based self-supervised learning? 2) How to ensure that the contrastive learning network learns features with local semantic action categories?

In this paper, we propose SkeAttnCLR, a contrastive framework for self-supervised action recognition based on the attention mechanism. It addresses the issues of learning the relationship between local and global features of human actions and ensures that the contrastive learning network learns features with local semantic information. Motivated by Fig. 1, the proposed scheme consists of two parts: global contrastive learning and local contrastive learning. The global contrastive learning follows the spirit of SkeletonCLR [Li *et al.*, 2021a], which is used to learn the global structure information of the human skeleton. The local contrastive learning is developed to learn local action features with discriminative semantic information. Specifically, the attention mask module based on a multi-head self-attention mechanism [Vaswani *et al.*, 2017] (MHSAM) is used to explore local features. This module divides skeleton action features into salient and non-salient areas at the feature level. Contrastive pairs are constructed for salient and non-salient features, as well as negative contrastive pairs between them to represent their oppositions in the contrastive learning model. This allows the network to learn key features with semantic distinction, regardless of whether they are embodied in salient or non-salient features.

The proposed method, SkeAttnCLR, presents a novel contrastive learning architecture that effectively learns the overall structure of human skeletal actions through global contrastive learning while also extracting key action features through local contrastive learning. As our SkeAttnCLR performs feature-level attention without interfering with the encoder structure, it can be applied to different encoder types, making it generalizable in extracting better action representations for downstream tasks. Our contributions are summarized as follows:

- A novel contrastive learning architecture is presented, in which the overall structure of human skeletal actions are learned through global contrastive learning, and key action features are extracted through local contrastive learning.
- A global-local contrastive learning framework SkeAttnCLR that leverages the attention mechanism with local similarity for skeleton-based models is proposed.
- We develop the Multi-Heads Attention Mask module to improve contrastive learning performance by generating ample contrastive pairs. This is achieved via salient and non-salient features.
- The proposed method outperforms the state-of-the-art methods in most evaluation metrics and especially achieves an overall lead in comprehensive comparison with the baseline, which employs only global features.

2 Related Work

2.1 Self-Supervised Representation Learning

Self-supervised learning is a method of learning data representation in a large amount of unlabeled data by setting a specific pretext task. There are many kinds of pretext tasks in self-supervised learning, which can be divided into two types: generative and discriminative. The examples, such as jigsaw puzzles [Wei *et al.*, 2019], data restoration [Pathak *et al.*, 2016], and Mask-based methods [Bao *et al.*, 2021; He *et al.*, 2022] that have been successful in image and language fields are generative self-supervised learning. Instance discrimination is a class of discriminative self-supervised learning pretext tasks. Instance discrimination based on contrastive learning [He *et al.*, 2020] is a class of discriminative self-supervised learning pretext tasks. In recent years, the contrastive learning method of the Moco series [He *et al.*, 2020; Chen *et al.*, 2020b] stores the key vector by constructing a memory bank queue, and updates the encoder K through the momentum update mechanism. SimCLR [Chen *et al.*, 2020a] improves the performance of contrastive learning by adding additional MLP modules and embedding calculations with large batch sizes. In addition, BYOL [Richemond *et al.*, 2020], contrastive clustering [Li *et al.*, 2021b], DINO [Caron *et al.*, 2021], and SimSiam [Chen and He, 2021] have also achieved promising results. In the recent work LEWEL [Huang *et al.*, 2022], the importance of local feature contrastive learning is emphasized for the first time in image tasks. The work of predecessors laid the foundation for our SkeAttnCLR, allowing us to go further on this basis.

2.2 Skeleton-Based Action Recognition

Earlier human skeleton action recognition models based on deep learning are mainly designed based on RNN [Hochreiter *et al.*, 2001] and CNN [Ke *et al.*, 2017; Li *et al.*, 2017]. In recent years, due to the development of graph networks, human action skeleton recognition has begun to use GRU [Shi *et al.*, 2017; Su *et al.*, 2020] or GCN-based [Li *et al.*, 2019; Liang *et al.*, 2019] models. At the same time, due to the recent success of the Transformer model in images and natural language, there have been many attempts to design a Transformer-based [Shi *et al.*, 2020; Plizzari *et al.*, 2021] human skeleton action recognition model. ST-GCN [Yan *et al.*, 2018] is a widely used GCN-based human skeleton recognition model in recent years. It models the skeleton data structure from the perspective of Spatial-Temporal. In the experiment of this paper, we mainly use ST-GCN as the backbone encoder. In addition, in order to demonstrate the generalizability of our method, we also use the GRU-based BIGRU [Su *et al.*, 2020] and Transformer-based DSTA [Shi *et al.*, 2020] to conduct comparison experiments.

2.3 Contrastive Learning for Skeleton-Based Models Pre-training

SkeleonCLR [Li *et al.*, 2021a] is a simple contrastive learning framework designed on the basis of MocoV2 [Chen *et al.*, 2020b]. On this basis, CrossCLR [Li *et al.*, 2021a] was proposed for multi-view contrastive learning to achieve cross-view consistency. AimCLR [Guo *et al.*, 2022] and HiCLR

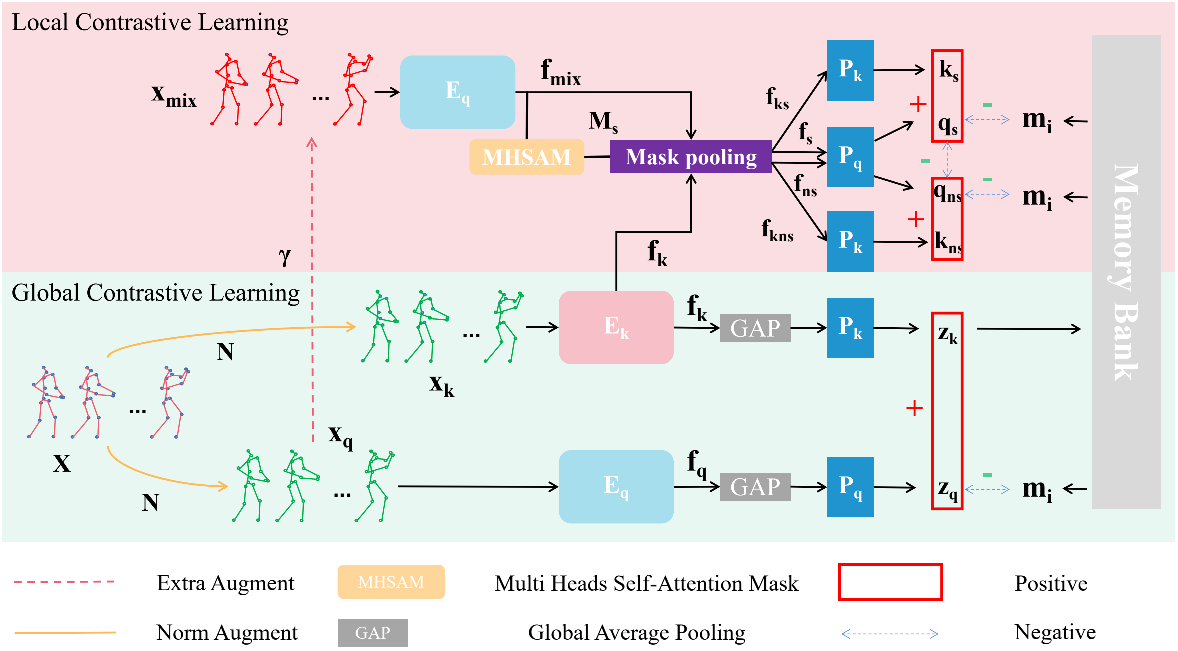


Figure 2: Architecture of the proposed SkeAttnCLR. First, in the global contrastive learning, original data X first obtains two different data augmented versions x_q and x_k after normal data augmentation N in Section 3.1. Then, x_q and x_k input their respective encoders E_q and E_q to obtain feature embeddings f_q and f_k for the next step of calculation. Meanwhile, the x_q is converted into x_{mix} after extra data augmentation γ in Section 3.2. x_{mix} gets its feature embedding f_{mix} through E_q. Finally, f_q and f_k are calculated for global contrastive learning as described in Section 3.1. At the same time, f_{mix} and f_k obtain local feature embedding under the action of the soft mask M_s generated by the MHSAM module and perform local contrastive learning calculations according to Section 3.2. As described in Section 3.1, at each iteration, the z_k of the current round is stored in the memory bank M_k. Then, according to the storage order, m_i is generated to construct negative pairs.

[Zhang *et al.*, 2022] aim to expand more contrastive pairs under stronger data augmentation conditions to improve single-view contrastive learning performance. SkeleMixCLR [Chen *et al.*, 2022] relies on the unique skeleton mixing data augmentation to design a targeted contrastive learning framework. It is noted that there is a lack of consideration of how to use the local features of human motion in the existing skeleton-based contrastive learning methods. The data augmentation method in the SkeleMixCLR locally mixes real human action parts to explore local feature combinations. However, this method needs to be marked at the feature level according to the Spatial-Temporal position of the data mixture, which requires the data to maintain Spatial-Temporal consistency after downsampling. Hence SkeleMixCLR is not conducive to extending to other backbones. It is desirable to propose a simple and generalizable local contrastive learning method.

3 SkeAttnCLR

As aforementioned, the local information of human motion has not been fully mined and emphasized. In this study, We attempt to extend local contrastive learning based on feature-level local similarity to global contrastive learning. We also use the attention mask generated by MHSAM to divide our defined attention salient features and non-salient features at the feature level. Then, contrastive pairs are constructed using the relations within local features, and between local and global features for the pretext task of instance discrimination.

SkeAttnCLR is shown in Fig. 2, which is a method built on a single view. In the global part, we follow the basic design of SkeletonCLR [Li *et al.*, 2021b]. The input to the local contrastive learning part comes from further data augmentation of the global part query input. In the local contrastive learning part, we divide the feature vectors obtained from encoder E_q or encoder E_k into attention-salient and non-salient embeddings through the MHSAM module. Finally, local-to-local, global-to-global, and local-to-global contrastive pairs are constructed between the global and local embeddings.

3.1 Global Contrastive Learning

In this section, we introduce the specific details of global contrastive learning, laying the foundation for the subsequent introduction of local contrastive learning.

Data Augmentation. For the input data of global contrastive learning, we adopt Shear and Crop[Li *et al.*, 2021a] as the augmentation strategy. We refer to this part of the data augmentation combination as normal data augmentation N. N randomly converts the read skeleton sequence X into two different data-augmented versions x_q and x_k as positive pairs.

Global Contrastive Learning Module. As shown in Fig. 2, the two encoders E_q and E_k respectively embed x_q and x_k into the feature space: $f_q = \mathrm{E}_q\left(x_q; \theta_q\right)$ and $f_k = \mathrm{E}_k\left(x_k; \theta_k\right)$, where $f_q, f_k \in \mathbb{R}^{n \times C_f}$. Among them, E_k follows E_q to up-

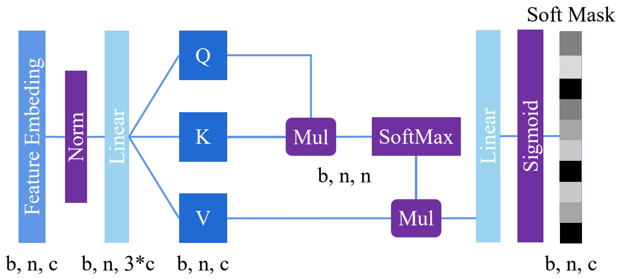


Figure 3: The Multi Head Self-attention Mask Module (MHSAM), whose input is feature-level embedding, and the output is a soft mask, which is also regarded as an attention-weight matrix.

date the parameters θ_k through the momentum update mechanism: $\theta_k \leftarrow M\theta_k + (1 - M)\theta_q$, where M is a momentum coefficient. We apply a dynamic momentum coefficient that changes according to the number of global iterations following BYOL [Richemond *et al.*, 2020], the formula is as follows: $M = 1 - (1 - M_0) \cdot (\cos(\pi \cdot iter/iter_{max}) + 1)/2$.

Then, after f_q and f_k are processed by global average pooling, they are respectively input into predictor P_q and P_k to obtain the output embedding z_q and z_k: $z_q = \mathrm{P}_q(f_q)$ and $z_k = \mathrm{P}_k(f_k)$, where $z_q, z_k \in \mathbb{R}^{C_z}$. p_k is the momentum updated version of p_q.

To construct negative pairs, a queue $M_k = [m_i]_{i=1}^{K}$ named memory bank is used to store previous embeddings z_k. At each iteration, M_k provides a large number of negative pairs for contrastive learning. In each iteration, z_k is calculated by the previous samples stored in M_k which is dequeued according to the storage order to obtain m_i, forming a large number of negative pairs with the newly calculated z_q.

In the global contrastive learning part, we adopt InfoNCE [Oord *et al.*, 2018] for global instance discrimination:

$$L_{\text{Info}} = -\log\left(\frac{\exp(z_q \cdot z_k/\tau)}{\exp(z_q \cdot z_k/\tau) + Neg(z_q)}\right) \qquad (1)$$

z_q, z_k and m_i are all normalized. τ is a hyperparameter (set to 0.2 in experiments). $Neg(z) = \sum_{i=1}^{K} \exp(z \cdot m_i/\tau)$ denotes the similarity between embedding z and m_i from memory bank. In this loss function, the global distance between samples is calculated by the dot product.

3.2 Local Contrastive Learning

Local contrastive learning builds upon the concept of global contrastive learning by specifically focusing on mastering the structural features of human action sequences. Extra data augmentation can help the model explore more local feature combinations. The samples generated by the extra data augmentation get new feature embeddings through E_q, and we use these feature embeddings to construct positive pairs with z_k, which locally brings similar samples closer. Conversely, the negative pairs of embeddings obtained from local features and m_i stored in the memory bank will pull away samples that lack local similarities. In addition, we construct an opposite relationship between attention-salient features and non-attention-salient features through two methods of masking and setting up negative pairs. It not only expands the number of negative pairs in contrastive learning but also allows the network to learn to focus on parts with key action semantics.

In this section, we first perform extra data augmentation on the basis of x_q, then we use MHSAM to divide the feature embeddings of the local contrastive learning, and finally, we construct rich contrastive pairs between the divided local embeddings in pursuit of better local feature exploration.

Extra Augmentation. According to AimCLR [Guo *et al.*, 2022], stronger data augmentation is beneficial for learning human action features under certain methods. Therefore, in the part of local contrastive learning, we apply an extra data augmentation γ on the basis of normal data augmentation results to explore the possibility of more human action features. Since data mixing augmentation [Chen *et al.*, 2022] can randomly combine parts of different human motion samples, it is more conducive to exploring local features than other noise addition and filtering methods. Therefore, we apply data mixing data augmentation here as our extra augmentation. In the part of local comparison learning, x_q is randomly transformed into x_{mix} after data mixing augmentation through γ.

Multi Head Self Attention Mask. With the success of transformer [Vaswani *et al.*, 2017] in various tasks, the multi-head self-attention mechanism has also attracted much attention as a part of the transformer. Due to its unique query-matching mechanism, the multi-head self-attention mechanism is able to estimate the correlation between features from multiple angles, and mine the connection between local features of actions at the feature level. The MHSAM module is shown in Fig. 3. The input to this module is a tensor of size $b \times n \times c$, where b represents the batch size, n is the length of a single data, and c is the number of data channels.

MHSAM is designed to embed the encoder output at the feature level. Since the Encoder in the local comparative learning shares parameters with the E_q of the global comparative learning, so we have $f_{mix} = \mathrm{E}_q(x_{mix}; \theta_q)$, where $f_{mix} \in \mathbb{R}^{n \times C_f}$. The Q, K and V of the multi-head self-attention mechanism are calculated by the following formula:

$$\mathrm{Q},\ \mathrm{K},\ \mathrm{V} = \mathrm{Linear}(f_{mix}) = f_{mix}w_q, f_{mix}w_k, f_{mix}w_v \qquad (2)$$

w_q, w_k and w_v is the parameter matrix of the linear network. Then attention feature is described by the softmax function:

$$x_{attn} = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (3)$$

where $\frac{1}{\sqrt{d_k}}$ is the normalization scaling factor. After calculation, x_{attn} is put into a simple linear network projection for adjustment. Finally, there is a Sigmoid function to generate a soft mask M_s. The formula is as follows:

$$M_s = \mathrm{Sigmoid}(\lambda \cdot \mathrm{proj}(x_{attn})) \qquad (4)$$

where λ is a hyperparameter that adjusts the tolerance of neutral features. The larger the value of λ, the easier the value of the mask tends to be polarized.

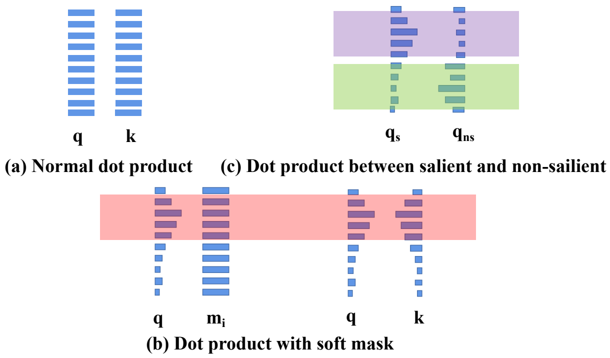

q
k
(a) Normal dot product
q_s
q_ns
(c) Dot product between salient and non-sailient
q
m_i
q
k
(b) Dot product with soft mask

Figure 4: (a) The normal dot product in the calculation of L_{Info}. (b) and (c) show the attention affected dot product in the calculation of L_s and L_{ns}.

Local Contrastive Learning Module. After obtaining the mask, we use it to separate the feature-level embedding extracted by the encoder into salient and non-salient features. Following the mask pooling procedure, inputs f_{mix} and f_k are transformed in both salient and non-salient ways to get f_s, f_{ns}, f_{ks} and f_{kns}. The formula is as follows:

$$M_{ns} = I - M_s \tag{5}$$

$$f_s = \left(\sum^n f_{mix} \cdot M_s\right)/n, f_{ks} = \left(\sum^n f_k \cdot M_s\right)/n \tag{6}$$

$$f_{ns} = \left(\sum^n f_{mix} \cdot M_{ns}\right)/n, f_{kns} = \left(\sum^n f_k \cdot M_{ns}\right)/n \tag{7}$$

where I is the identity matrix, and n is the data length of the dimension that needs to be pooled.

After getting the embeddings from mask pooling, we have $q_s = \mathrm{P}_q(f_s)$, $q_{ns} = \mathrm{P}_q(f_{ns})$, $k_s = \mathrm{P}_k(f_{ks})$ and $k_{ns} = \mathrm{P}_k(f_{kns})$. In local contrastive learning, we make q_s and k_s a positive pair, q_{ns} and k_{ns} a positive pair, and set q_s and q_{ns} a negative pair. Then, the loss functions of salient features and non-salient features are extended by Section 3.1 as:

$$L_s = -\log\left(\frac{\exp(q_s \cdot k_s/\tau)}{\exp(q_s \cdot k_s/\tau) + \exp(q_s \cdot q_{ns}/\tau) + Neg(q_s)}\right) \tag{8}$$

$$L_{ns} = -\log\left(\frac{\exp(q_{ns} \cdot k_{ns}/\tau)}{\exp(q_{ns} \cdot k_{ns}/\tau) + \exp(q_{ns} \cdot q_s/\tau) + Neg(q_{ns})}\right) \tag{9}$$

In the estimation of L_s and L_{ns}, We utilize attention to leverage the calculation of local similarity as shown in Fig. 4. We assume that the general dot product is shown in Fig. 4 (a). The mask generated by MHSAM highlights the parts that are concerned at the feature level so that the result of the dot product operation tends to highlight the local similarity in Fig. 4 (b). The dot product between q_s and q_{ns} is shown in Fig. 4 (c), which represents the opposition between them. Besides, We still use the memory bank M_k that stores global feature embeddings to provide negative pairs in local contrastive learning, thus it is worth noting that the dot product of local-to-global also emphasizes the local similarity as local-to-local does, as shown in Fig. 4 (b). Finally, the total loss function for local contrastive learning can be expressed as:

$$L_{\text{local}} = \mu L_s + (1-\mu)L_{ns} \tag{10}$$

where $\mu \in (0,1)$.

3.3 The Overall Objective of SkeAttenCLR

SkeAttnCLR estimates the distance of global features in the feature space between samples through global contrastive learning for global instance discrimination and also computes the distance of local features through local contrastive learning for instance discrimination. Combining local and global contrastive learning, SkeAttnCLR can be optimized by the following loss function:

$$L = L_{\text{global}} + L_{\text{local}} = L_{\text{info}} + L_{\text{local}} \tag{11}$$

4 Experiments

4.1 Datasets

We select the most widely used NTU series dataset for experimental evaluation. The NTU dataset contains a wide range of human action categories and has a unified and standardized data processing code in use for many years, which ensures fairness when compared with previous methods.

NTU-RGB+D 60 (NTU-60). NTU-60 [Shahroudy *et al.*, 2016] is a large-scale skeleton dataset for human skeleton-based action recognition, containing 56,578 videos with 60 actions and 25 joints for each human body. The dataset includes two evaluation protocols: the Cross-Subject (X-Sub) protocol, which divides data by subject with half used for training and half for testing, and the Cross-View (X-View) protocol, which uses different camera views for training. The testing samples are captured by cameras 2 and 3 for training, and samples from camera 1 are used for testing.

NTU-RGB+D 120 (NTU-120). NTU-120 [Liu *et al.*, 2019] is an expansion dataset of NTU-60, containing 113,945 sequences with 120 action labels. It also offers two evaluation protocols, the Cross-Subject (X-Sub) and Cross-Set (X-Set) protocols. In X-Sub, 53 subjects are used for training and 53 subjects are used for testing, while in X-Set, half of the setups are used for training (even setup IDs) and the remaining setups (odd setup IDs) are used for testing.

PKU Multi-Modality Dataset (PKU-MMD). PKU-MMD [Liu *et al.*, 2020] is a substantial dataset that encompasses a multi-modal 3D comprehension of human actions, containing around 20,000 instances and 51 distinct action labels. It is split into two subsets for varying levels of complexity: Part I is designed as a simpler version, while Part II offers a more challenging set of data due to significant view variation.

4.2 Experimental Settings

Our experiments mainly use the SGD optimizer [Ruder, 2016] to optimize the model. For all contrastive learning training, we use a learning rate of 0.1, a momentum of 0.9, and a weight decay of 0.0001 for a total of 300 epochs for training, and adjust the basic learning rate to one-tenth of the original at the 250th epoch. In addition, our data processing employs

Method	Stream	NTU-60		NTU-120		PKU-MMD	
		Xsub	Xview	Xsub	Xset	Part I	Part II
Baseline	J	68.3	76.4	-	-	-	-
Baseline*	J	72.0 +3.7	79.0 +2.6	51.0	61.7	80.3	39.1
Ours	J	**80.3** +12.0	**86.1**+9.7	**66.3**+15.1	**74.5**+12.8	**87.3**+7.0	**52.9**+23.8
Baseline	M	53.3	50.8	-	-	-	-
Baseline*	M	56.5+3.2	57.2+6.4	46.1	43.8	66.5	14.0
Ours	M	**63.9**+10.6	**58.7**+7.9	**49.9**+3.8	**59.3**+15.5	**72.2**+5.7	**32.7**+18.7
Baseline	B	69.4	67.4	-	-	-	-
Baseline*	B	66.0-3.4	69.0+1.6	51.1	56.3	79.7	21.5
Ours	B	**76.2**+6.8	**76.0**+8.6	**63.0**+11.9	**67.3**+11.0	**87.3**+7.6	**37.7**+16.2
Baseline	3S	75.0	79.8	-	-	-	-
Baseline*	3S	75.9 +0.9	79.8 0	65.0	65.9	85.3	38.8
Ours	3S	**82.0**+6.1	**86.5**+6.7	**77.1**+12.1	**80.0**+14.1	**89.5**+4.2	**55.5**+16.7

Table 1: Linear evaluation comparisons with the baseline using the same backbone, where J, M, and B indicate joint, motion, and bone, 3S means three streams fusion, $*$ indicates that results obtained with our settings.

human skeleton action sequences with a length of 64 frames, and the batch size is 128. We choose ST-GCN [Yan *et al.*, 2018] as the main backbone of our experiments for a fair comparison, as it is the most widely adopted method in existing skeleton-based action recognition approaches. Meanwhile, We also provide the detailed settings for all backbones in the experiments and t-Distributed Stochastic Neighbor Embedding (t-SNE) [Van der Maaten and Hinton, 2008] visualization results in the **Appendix**.

KNN Evaluation Protocol. During the contrastive learning training process, we directly use a KNN cluster every 10 epochs to cluster the feature embeddings extracted by the encoder and evaluate the clustering accuracy on the test set. Finally, the model with the highest KNN result is selected to participate in other experiments.

Linear Evaluation Protocol. Linear evaluation is the most commonly used evaluation method for downstream classification tasks. Its usual practice is to freeze the parameters of the backbone encoder trained by self-supervised learning, and use supervised learning to train a linear fully connected layer classifier in the test. In the experiment, we use the SGD optimizer with a learning rate of 3 to train for 100 epochs and adjust the base learning rate at 60th epoch.

Finetune Evaluation Protocol. Different from the linear evaluation, finetune evaluation does not freeze the parameter update of the encoder. We use SGD with an initial learning rate of 0.05 for optimization. In this experiment, the dynamic learning rate adjustment that comes with PyTorch [Paszke *et al.*, 2019] is applied, which automatically adjusts when the loss does not converge.

Semi-finetune Evaluation Protocol. The difference between semi-finetune and finetune evaluation is that the former only uses a few labeled data for training. We experimented with 1% labeled data and 10% labeled data respectively, and the optimizer settings follow finetune evaluation.

4.3 Result Comparison

To validate the effectiveness of our method, we compare it with other methods of linear evaluation, KNN evaluation, finetune evaluation, and semi-finetune evaluation protocols. To facilitate fair comparisons, we mainly choose similar methods that use ST-GCN (backbone network) and achieve SOTA in recent years for comparison.

Comparisons with Baseline. Our method is compared with the baseline(SkeletonCLR[Li *et al.*, 2021a]) on NTU-60, NTU-120 and PKU-MMD datasets when ST-GCN is used as the backbone encoder. The results are shown in Table 1. In order to demonstrate the generalizability of our method, we additionally use BIGRU [Su *et al.*, 2020] and transformer (DSTA) [Shi *et al.*, 2020] as the backbone encoder on the NTU-60 dataset for comparison with the baseline. The results are shown in Table 2. As we can see from Table 1 and Table 2, our method has a comprehensive improvement compared to the baseline on different datasets and different backbone encoders. The experiments demonstrate the effectiveness of our method adding local contrastive learning on a global basis.

Method	Stream	ST-GCN	BIGRU	Transformer
		Xsub	Xsub	Xsub
Baseline	J	68.3	-	-
Baseline*		72.0+3.7	64.8	54.5
Ours		**80.3**+12.0	**72.7**+7.9	**71.3**+16.8
Baseline	M	53.3	-	-
Baseline*		56.5+3.2	58.4	45.0
Ours		**63.9**+10.6	**74.9**+16.5	**52.3**+7.3
Baseline	B	69.4	-	-
Baseline*		66.0 -3.4	63.0	51.0
Ours		**76.2** +6.8	**69.2** +6.2	**73.7** +22.7

Table 2: Linear evaluation comparisons with different backbones on NTU-60 dataset. J, M, and B indicate joint, motion, and bone.

Comparisons with Previous Works. For a fair comparison, we mainly select works that also mainly use ST-GCN for experiments in recent years and achieve SOTA to compare with our method. The experimental results are shown in Table 3, our method is in an advantageous position in most comparisons. Especially in the comparison of three-stream results under the NTU-120 dataset, we have achieved a comparative advantage of more than 7%. In the next analysis of KNN results, we combine the results of the linear evaluation to analyze the

Method	Stream	NTU-60		NTU-120	
		Xsub	Xview	Xsub	Xset
SkeletonCLR	J	68.3	76.4	56.8	55.9
CrossCLR	J	72.9	79.9	-	-
AimCLR	J	74.3	79.7	63.4	63.4
HiCLR	J	77.6	82.4	-	-
SkeleMixCLR	J	79.6	84.4	67.4	69.6
Ours	J	80.3	86.1	66.3	74.5
SkeletonCLR	3S	77.8	83.4	67.9	66.7
AimCLR	3S	78.9	83.8	68.2	68.8
HiCLR	3S	80.4	85.5	70	70.4
SkeleMixCLR	3S	81	85.6	69.1	69.9
Ours	3S	82.0	86.5	77.1	80.0

Table 3: Linear evaluation comparisons with other methods using the same backbone, J indicates joint, 3S means three streams fusion.

reason why our xsub single-stream in NTU-120 does not reach the best. Notably, we have a considerable improvement in the NTU-120 dataset with three streams of data ensemble, which indicates that SkeAttnCLR performs better with the fusion of joint-bone-motion streams.

KNN Evaluation Results. As shown in Table 4, in the KNN evaluation comparison with similar methods, our method achieves SOTA in most indicators. Based on the results of the linear evaluation, we speculate that SkeAttnCLR performs worse in xsub due to skeleton captured by xsub is not as good as that of xset, which leads to bias in local similarity.

Method	NTU-60		NTU-120	
	Xsub	Xview	Xsub	Xset
SkeletonCLR	60.7	64.8	42.9	41.9
AimCLR	63.7	71.0	47.3	48.9
HiCLR	67.3	75.3	-	-
SkeleMixCLR	65.5	72.3	48.3	49.3
Ours	69.4	76.8	46.7	58.0

Table 4: KNN evaluation results on NTU-RGB+D dataset.

Semi-finetune Evaluation Results. The experimental results of semi-finetune are shown in Table 5, which shows that our method is not limited by the amount of labeled data.

Method	Label	NTU-60	
		Xsub	Xview
3S-CrossCLR	1%	51.1	50
3S-AimCLR		54.8	54.3
3S-SkeleMixCLR		55.3	55.7
Ours		59.6	59.2
3S-CrossCLR	10%	74.4	77.8
3S-AimCLR		78.2	81.6
3S-SkeleMixCLR		79.9	83.6
Ours		81.5	83.8

Table 5: Semi-supervised evaluation results.

Finetune Evaluation Results. From our experimental results in Table 6, our method has surpassed the recent methods and backbone which trained from scratch. From the overall effect, our method provides the most effective pre-training parameters for supervised fine-tuning.

Method	Stream	NTU-60		NTU-120	
		Xsub	Xview	Xsub	Xset
SkeletonCLR	J	82.2	88.9	73.6	75.3
AimCLR		83.0	89.2	77.2	76.0
SkeleMixCLR		84.5	91.1	75.1	76.0
Ours		87.3	92.8	77.3	87.8
ST-GCN	3S	85.2	91.4	77.2	77.1
SkeletonCLR		86.2	92.5	80.5	80.4
AimCLR		86.9	92.8	80.1	80.9
HiCLR		88.3	93.2	82.1	83.7
SkeleMixCLR		87.8	93.9	81.6	81.2
Ours		89.4	94.5	83.4	92.7

Table 6: Fully finetune evaluation results, J means joint, 3S indicates three streams.

4.4 Ablation Study

Ablation studies are conducted on NTU-60 dataset, and the related evaluation protocol is introduced in Section 4.2.

Negative pair: q_s vs q_{ns}	L_s	L_{ns}	Xsub	Xview
×	√	√	80.9	84.1
×	√	×	75.5	79.2
√	√	×	76.3	80.4
√	√	√	80.3	86.1

Table 7: Ablation study of loss function designs on NTU-60 dataset Joint level.

Ablation Study of Local Contrastive Loss Function Designs. In Section 3.2, we introduce L_{local} to optimize local contrastive learning, which is composed of two mirrored loss functions for the salient area and the non-salient area after weighting. In the experiment, we verified the effect of adding q_s and q_{ns} as a hard-negative contrastive pair, and the necessity of mirror loss function design for the salient and the non-salient feature area. Then, the results are shown in Table 7. In addition, we also conduct ablation experiments for the parameter μ, and verify that the optimal weight μ of L_s and L_{ns} is 0.5, which also shows the mirror image relationship of L_s and L_{ns}. The experimental results are shown in the **Appendix** together with the parameter tuning experiments of the MHSAM module in Section 3.2.

5 Conclusion

In this work, we propose SkeAttnCLR, a novel attention-based contrastive learning framework for self-supervised 3D skeleton action representation learning aimed at enhancing the acquisition of local features. The proposed method emphasizes the importance of learning local action features by leveraging attention-based instance discrimination to bring samples with similar local features closer to the feature space. Experimental results demonstrate that SkeAttnCLR achieves significant improvements over the baseline approach that only relies on global learning. Particularly, our framework achieves outstanding results in various evaluation metrics based on the NTU-60 and NTU-120 datasets.

References

[Bao *et al.*, 2021] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[Cao *et al.*, 2017] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, 2017.

[Caron *et al.*, 2021] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *CVPR*, pages 9650–9660, 2021.

[Chen and He, 2021] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021.

[Chen *et al.*, 2020a] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[Chen *et al.*, 2020b] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[Chen *et al.*, 2022] Zhan Chen, Hong Liu, Tianyu Guo, Zhengyan Chen, Pinhao Song, and Hao Tang. Contrastive learning from spatio-temporal mixed skeleton sequences for self-supervised skeleton-based action recognition. *arXiv preprint arXiv:2207.03065*, 2022.

[Guo *et al.*, 2022] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *AAAI*, volume 36, pages 762–770, 2022.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.

[Hochreiter *et al.*, 2001] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, J urgen Schmidhuber, and Corso Elvezia. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. *A Field Guide to Dynamical Recurrent Neural Networks*, pages 237–244, 2001.

[Huang *et al.*, 2022] Lang Huang, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, and Toshihiko Yamasaki. Learning where to learn in cross-view self-supervised learning. In *CVPR*, pages 14451–14460, 2022.

[Ke *et al.*, 2017] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *CVPR*, pages 3288–3297, 2017.

[Kim *et al.*, 2021] Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. Self-guided contrastive learning for bert sentence representations. In *ACL (Volume 1: Long Papers)*, pages 2528–2540, 2021.

[Li *et al.*, 2017] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Skeleton-based action recognition with convolutional neural networks. In *ICMEW*, pages 597–600. IEEE, 2017.

[Li *et al.*, 2019] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *CVPR*, pages 3595–3603, 2019.

[Li *et al.*, 2021a] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 3d human action representation learning via cross-view consistency pursuit. In *CVPR*, pages 4741–4750, 2021.

[Li *et al.*, 2021b] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *AAAI*, volume 35, pages 8547–8555, 2021.

[Liang *et al.*, 2019] Duohan Liang, Guoliang Fan, Guangfeng Lin, Wanjun Chen, Xiaorong Pan, and Hong Zhu. Three-stream convolutional neural network with multi-task and ensemble learning for 3d action recognition. In *CVPRW*, pages 0–0, 2019.

[Liu *et al.*, 2019] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE TPAMI*, 42(10):2684–2701, 2019.

[Liu *et al.*, 2020] Jiaying Liu, Sijie Song, Chunhui Liu, Yanghao Li, and Yueyu Hu. A benchmark dataset and comparison study for multi-modal human action analytics. *ACM TOMM*, 16(2):1–24, 2020.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NIPS*, 32, 2019.

[Pathak *et al.*, 2016] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016.

[Plizzari *et al.*, 2021] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial temporal transformer network for skeleton-based action recognition. *ICPR*, 2021.

[Richemond *et al.*, 2020] Pierre H Richemond, Jean-Bastien Grill, Florent Altché, Corentin Tallec, Florian Strub, Andrew Brock, Samuel Smith, Soham De, Razvan Pascanu, Bilal Piot, et al. Byol works even without batch statistics. *arXiv preprint arXiv:2010.10241*, 2020.

[Ruder, 2016] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[Shahroudy *et al.*, 2016] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, pages 1010–1019, 2016.

[Shi *et al.*, 2017] Yemin Shi, Yonghong Tian, Yaowei Wang, Wei Zeng, and Tiejun Huang. Learning long-term dependencies for action recognition with a biologically-inspired deep network. In *ICCV*, pages 716–725, 2017.

[Shi *et al.*, 2020] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *ACCV*, 2020.

[Su *et al.*, 2020] Kun Su, Xiulong Liu, and Eli Shlizerman. Predict & cluster: Unsupervised skeleton based action recognition. In *CVPR*, pages 9631–9640, 2020.

[Sun *et al.*, 2019] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 30, 2017.

[Wei *et al.*, 2019] Chen Wei, Lingxi Xie, Xutong Ren, Yingda Xia, Chi Su, Jiaying Liu, Qi Tian, and Alan L Yuille. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In *CVPR*, pages 1910–1919, 2019.

[Yan *et al.*, 2018] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

[Zhang *et al.*, 2020] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *CVPR*, pages 1112–1121, 2020.

[Zhang *et al.*, 2022] Jiahang Zhang, Lilang Lin, and Jiaying Liu. Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations. *arXiv preprint arXiv:2211.13466*, 2022.

[Zheng *et al.*, 2022] Ce Zheng, Matias Mendieta, Pu Wang, Aidong Lu, and Chen Chen. A lightweight graph transformer network for human mesh reconstruction from 2d human pose. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5496–5507, 2022.

[Zheng *et al.*, 2023] Ce Zheng, Matias Mendieta, Taojiannan Yang, Guo-Jun Qi, and Chen Chen. Feater: An efficient network for human reconstruction feature map-based transformer. In *CVPR*, 2023.