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

\{hy1997, shiqian.wu\}@wust.edu.cn, \{wwu25, aidong.lu\}@unc.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.

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 ac-
tion 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

SkeletonCLR [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...
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 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: $f_q = E_q(x_q; \theta_q)$ and $f_k = E_k(x_k; \theta_k)$, where $f_q, f_k \in \mathbb{R}^{n \times C}$. Among them, $E_k$ follows $E_q$ to up-

---

[Li et al., 2021b] rely 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.
date the parameters $\theta_k$ 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: $M = 1 - (1 - M_0) \cdot (\cos(\pi \cdot \text{iter}/\text{iter}_{\text{max}}) + 1)/2$.

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 $z_k$: $z_q = P_q(f_q)$ and $z_k = P_k(f_k)$, where $z_q, z_k \in \mathbb{R}^C$. $p_k$ 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, $z_k$ 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:

$$L_{\text{Info}} = -\log \left( \frac{\exp(z_q \cdot z_k / \tau)}{\exp(z_q \cdot z_k / \tau) + \exp(z_i \cdot z_k / \tau) \cdot N_\text{eg}(z_q)} \right) \quad (1)$$

$z_q, z_k$ and $m_i$ are all normalized. $\tau$ is a hyperparameter (set to 0.2 in experiments). $N_\text{eg}(z) = \sum_{i=1}^K \exp(z \cdot m_i / \tau)$ 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 $\gamma$ 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 $x_{\text{mix}}$ after data mixing augmentation through $\gamma$.

**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 $f_{\text{mix}} = E_q(x_{\text{mix}}; \theta_q)$, where $f_{\text{mix}} \in \mathbb{R}^{n \times C_f}$. The $Q, K$ and $V$ of the multi-head self-attention mechanism are calculated by the following formula:

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

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

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

where $\frac{1}{\sqrt{d_k}}$ is the normalization scaling factor. After calculation, $x_{\text{attn}}$ 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:

$$M_s = \text{Sigmoid}(\lambda \cdot \text{proj}(x_{\text{attn}})) \quad (4)$$

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

$$M_{ns} = I - M_s$$  \hspace{1cm} (5)

$$f_s = \left( \sum_{i} f_{mix} \cdot M_s \right) / n, f_{ks} = \left( \sum_{i} f_k \cdot M_s \right) / n$$  \hspace{1cm} (6)

$$f_{ns} = \left( \sum_{i} f_{mix} \cdot M_{ns} \right) / n, f_{kns} = \left( \sum_{i} f_k \cdot M_{ns} \right) / n$$  \hspace{1cm} (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 $q_s = P_q(f_s)$, $q_{ns} = P_q(f_{ns})$, $k_s = P_k(f_{ks})$ and $k_{ns} = P_k(f_{kns})$. 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:

$$L_s = -\log \left( \frac{\exp(q_s \cdot k_s / \tau)}{\exp(q_s \cdot k_s / \tau) + \exp(q_s \cdot q_{ns} / \tau) + \text{Neg}(q_s)} \right)$$  \hspace{1cm} (8)

$$L_{ns} = -\log \left( \frac{\exp(q_{ns} \cdot k_{ns} / \tau)}{\exp(q_{ns} \cdot k_{ns} / \tau) + \exp(q_{ns} \cdot q_s / \tau) + \text{Neg}(q_{ns})} \right)$$  \hspace{1cm} (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:

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

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

3.3 The Overall Objective of SkeAttnCLR

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:

$$L = L_{global} + L_{local} = L_{info} + L_{local}$$  \hspace{1cm} (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-GRB+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...
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Stream</th>
<th>NTU-60</th>
<th>NTU-120</th>
<th>PKU-MMD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Xsub</td>
<td>Xview</td>
<td>Xsub</td>
</tr>
<tr>
<td>Baseline</td>
<td>J</td>
<td>68.3</td>
<td>76.4</td>
<td>-</td>
</tr>
<tr>
<td>Baseline*</td>
<td>J</td>
<td>72.0+3.7</td>
<td>79.0+2.6</td>
<td>51.0</td>
</tr>
<tr>
<td>Ours</td>
<td>J</td>
<td>80.3+12.0</td>
<td>86.1+9.7</td>
<td>66.3+15.1</td>
</tr>
<tr>
<td>Baseline</td>
<td>M</td>
<td>53.3</td>
<td>50.8</td>
<td>-</td>
</tr>
<tr>
<td>Baseline*</td>
<td>M</td>
<td>56.5+3.2</td>
<td>57.2+6.4</td>
<td>46.1</td>
</tr>
<tr>
<td>Ours</td>
<td>M</td>
<td>63.9+10.6</td>
<td>58.7+7.9</td>
<td>49.9+3.8</td>
</tr>
<tr>
<td>Baseline</td>
<td>B</td>
<td>69.4</td>
<td>-</td>
<td>69.4</td>
</tr>
<tr>
<td>Baseline*</td>
<td>B</td>
<td>66.0+3.4</td>
<td>69.0+1.6</td>
<td>51.1</td>
</tr>
<tr>
<td>Ours</td>
<td>B</td>
<td>76.2+6.8</td>
<td>76.0+8.6</td>
<td>63.0+11.9</td>
</tr>
<tr>
<td>Baseline</td>
<td>3S</td>
<td>75.0</td>
<td>79.8</td>
<td>-</td>
</tr>
<tr>
<td>Baseline*</td>
<td>3S</td>
<td>75.9+0.9</td>
<td>79.8</td>
<td>65.0</td>
</tr>
<tr>
<td>Ours</td>
<td>3S</td>
<td>82.0+6.1</td>
<td>86.5+6.7</td>
<td>77.1+12.1</td>
</tr>
</tbody>
</table>

Table 2: Linear evaluation comparisons with different backbones 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.

**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
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.

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

**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.

**4.4 Ablation Study**

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


[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.


