Hierarchical Consistent Contrastive Learning for Skeleton-Based Action Recognition with Growing Augmentations

Jiahang Zhang, Lilang Lin, Jiaying Liu[*]

Wangxuan Institute of Computer Technology, Peking University, Beijing, China
{zjh2020, linlilang, liujiaying}@pku.edu.cn

Abstract

Contrastive learning has been proven beneficial for self-supervised skeleton-based action recognition. Most contrastive learning methods utilize carefully designed augmentations to generate different movement patterns of skeletons for the same semantics. However, it is still a pending issue to apply strong augmentations, which distort the images/skeletons' structures and cause semantic loss, due to their resulting unstable training. In this paper, we investigate the potential of adopting strong augmentations and propose a general hierarchical consistent contrastive learning framework (HiCLR) for skeleton-based action recognition. Specifically, we first design a gradual growing augmentation policy to generate multiple ordered positive pairs, which guide to achieve the consistency of the learned representation from different views. Then, an asymmetric loss is proposed to enforce the hierarchical consistency via a directional clustering operation in the feature space, pulling the representations from strongly augmented views closer to those from weakly augmented views for better generalizability. Meanwhile, we propose and evaluate three kinds of strong augmentations for 3D skeletons to demonstrate the effectiveness of our method. Extensive experiments show that HiCLR outperforms the state-of-the-art methods notably on three large-scale datasets, *i.e.,* NTU60, NTU120, and PKUMMD. Our project is publicly available at: https://jhang2020.github.io/Projects/HiCLR/HiCLR.html.

1 Introduction

Human action recognition is important for bridging artificial systems and humans in the real world. It has been widely used in video understanding, human-robot interaction, entertainment, *etc.* (Tang et al. 2020; Rodomagoulakis et al. 2016; Shotton et al. 2011). Owing to the advantages such as lightweight, robustness, and privacy protection, action recognition based on 3D skeleton data has attracted a lot of attention recently. There are many works targeted at skeleton-based action recognition (Shi et al. 2019; Cheng et al. 2020; Liu et al. 2020b; Chen et al. 2021), but most of them are designed in a fully-supervised manner and require a large amount of labeled data. Considering that annotation of large-scale datasets is expensive and time-consuming, recently more and more researchers pay attention to the study of representation learning from unlabeled data (Lin et al. 2020; Li et al. 2021; Kim et al. 2022).

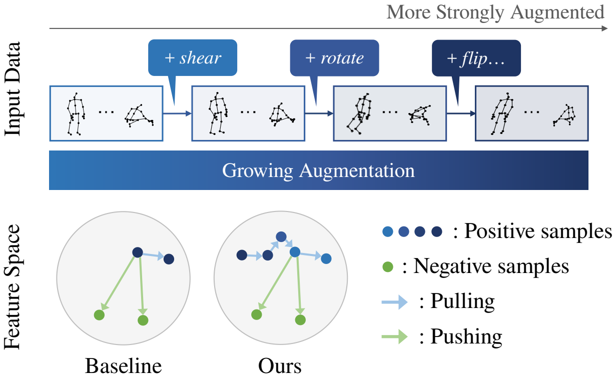

Figure 1: The proposed hierarchical consistent contrastive learning compared with the traditional contrastive learning pipelines. Instead of applying all augmentations directly, we utilize a growing augmentation to generate multiple ordered positive pairs that are augmented progressively. Then the model performs a directional feature clustering operation to constrain the consistency of adjacent positive samples.

Among the various self-supervised learning methods, contrastive learning is an effective one and has been shown successful for skeleton-based action recognition (Rao et al. 2021; Thoker, Doughty, and Snoek 2021; Li et al. 2021). For contrastive learning, augmentations have been proven to be very crucial, introducing various movement patterns for the same semantics and directly affecting the quality of feature representations learned by the model (Tian et al. 2020; Guo et al. 2022). However, it is still not fully investigated on what augmentation to use and how to use it for skeleton data.

Compared to the RGB representation of human action, 3D skeleton data is a more high-level modality representation, which intensifies the sensitivity of contrastive learning to the augmentations. This sensitivity leads to a cautious selection of augmentations, which becomes the bottleneck in designing more advanced contrastive learning methods. On the one hand, as shown on the right of Table 3, some augmentations like *Random Mask* cause the performance drop of the baseline algorithm. Following (Bai et al. 2022), these aug-

[*]Corresponding author. This work is supported by the National Natural Science Foundation of China under contract No.62172020.

mentations are called **strong augmentations** (also namely **heavy augmentations**), which distort the images/skeletons' structures and cause semantic loss, leading to unstable training. Some works have revealed the potential of using strong data augmentations (Cubuk et al. 2020; Wang and Qi 2021). However, it is still difficult to measure and constrain the consistency, which is the base of contrastive learning, directly and accurately from the strongly augmented views, as these augmentations can cause serious semantic information loss. On the other hand, most previous works based on contrastive learning simply treat all augmentations fairly, ignoring the differences in the importance of applied augmentations. Recent works (Tian et al. 2020; Zhang and Ma 2022) have shown that each augmentation has a different impact on the downstream tasks, and hence learning from the invariance after augmentation without distinction can inevitably cause non-optimal representations for the downstream task.

To address the aforementioned issues, we are inspired to explore a general contrastive framework that applies growing augmentations. In this paper, we propose a novel hierarchical consistent contrastive learning framework (HiCLR) that learns from the invariance of the hierarchical growing augmentation and treats different augmentations in a distinguished manner. Compared to previous works (Li et al. 2021; Guo et al. 2022), we focus on how to better utilize and benefit from multiple augmentations including strong augmentations. Specifically, a growing hierarchical augmentation policy is proposed to construct multiple correlated positive pairs. Each of the pairs is generated via more augmentations than the previous one and expands the feature distribution. Then, to better utilize the novel patterns brought by the strong augmentations, we propose an asymmetric hierarchical learning strategy. Instead of directly learning all augmentation invariances at once, our objective suggests a hierarchical consistent learning manner for different augmentations as shown in Figure 1. Meanwhile, this asymmetric design encourages the strongly augmented view to be similar to the weakly augmented view, which helps the model better generalize. Based on our new framework, we further analyze and evaluate different choices for strong augmentations. Extensive experiments on both Graph Convolutional Networks (GCNs) and transformers are conducted to verify the effectiveness of our method.

Our contributions can be summarized as follows:

- We propose a hierarchical consistent contrastive learning framework, HiCLR, which successfully introduces strong augmentations to the traditional contrastive learning pipelines for skeletons. The hierarchical design integrates different augmentations and alleviates the difficulty in learning consistency from strongly augmented views, which are accompanied by serious semantic information loss.
- We introduce the growing augmentation along with asymmetric hierarchical learning that constrains the representation consistency of the constructed positive pairs. By virtue of these, the model further improves the representation capacity by leveraging the rich information brought by the strong augmentations.
- With the proposed framework, we further design and analyze three strong augmentation strategies: Random Mask, Drop/Add Edges, and SkeleAdaIN. Despite the adverse effects observed when applying them directly, they become significantly effective with our HiCLR and exceed state-of-the-art performance.

2 Related Works

2.1 Skeleton-based Action Recognition

Skeleton-based action recognition aims to classify the action categories using 3D coordinates data of the human body. The current methods can be divided into recurrent neural network (RNN)-based, convolutional neural network (CNN)-based, GCN-based, and transformer-based styles. The work in (Du, Wang, and Wang 2015) directly uses RNN to tackle the skeleton sequence data. Song *et al.* (Song et al. 2017, 2018a,b) propose to utilize the attention mechanism and multi-modal information. Some works (Ke et al. 2017; Liu, Liu, and Chen 2017) transform each skeleton sequence into image-like representations and apply the CNN model to extract spatial-temporal information. Recently, inspired by the natural topology structure of the human body, GCN-based methods have attracted more attention. Spatial-temporal GCN (ST-GCN) (Yan, Xiong, and Lin 2018) first explores the potential of modeling the spatial-temporal relationship of skeleton data. Many works (Shi et al. 2019; Cheng et al. 2020) based on it have achieved success by virtue of the GCN's strong representation capacity. Meanwhile, transformer-based models (Shi et al. 2020; Plizzari, Cannici, and Matteucci 2021) also show remarkable results by utilizing the long-range temporal dependencies, owing to attention mechanism. We adopt the ST-GCN and DSTA-Net (Shi et al. 2020) as backbones to evaluate our method.

2.2 Contrastive Representation Learning

Contrastive learning (He et al. 2020; Chen et al. 2020a,b) is a popular and effective method for self-supervised learning. In many works (Tian et al. 2020; Zhang and Ma 2022), the design of augmentations has been found essential for the success of contrastive learning. Xiao *et al.* propose Leave-one-out Contrastive Learning (Xiao et al. 2020), which projects the input image into multiple embedding spaces corresponding to invariance learning of different augmentation combinations. The work (Zhang and Crandall 2022) proposes to decouple spatial-temporal contrastive learning by applying temporal and spatial augmentations separately.

Recent works have shown more and more interests in strong augmentations. Contrastive learning with stronger augmentations (CLSA) (Wang and Qi 2021) shows the improvements of strong augmentations in contrastive learning. The work in (Zhang and Ma 2022) proposes to apply more augmentations in different depths of the encoder to learn the augmentation invariances non-homogeneously. Bai *et al.* (Bai et al. 2022) explore a directional self-supervised objective for heavy image augmentations. These works provide an important basis for our research.

For contrastive learning in skeleton-based action recognition, AS-CAL (Rao et al. 2021) directly applies the current contrastive learning framework (He et al. 2020) for skeleton. Li *et al.* (Li et al. 2021) explores the cross-stream knowledge for contrastive learning. Recently, abundant information mining for self-supervised action representation (AimCLR) (Guo et al. 2022) migrates CLSA to skeleton data and uses more augmentations. However, these works still lack effective design for the use of strong augmentations, and leave the potential of strong augmentations underutilized. To this end, we propose a hierarchical consistent contrastive learning framework that can effectively borrow the knowledge of strong augmentations.

3 Proposed Method: HiCLR

3.1 Contrastive Learning for Skeleton

We first give a unified formulation of the contrastive learning (Bai et al. 2022) for skeleton following the recent works:

- **Data augmentation module** containing the augmentation strategy set $\mathcal{T}$ to generate the different views of the original data which are regarded as the positive pairs.
- **Query/key encoder** $f(\cdot)$ for mapping the input to the latent feature space.
- **Embedding projector** $h(\cdot)$ for mapping the latent feature into an embedding space where the self-supervised loss is applied.
- **Self-supervised loss** that performs the feature clustering operation in the embedding space.

SkeletonCLR (Li et al. 2021) follows the recent contrastive learning framework, MoCov2 (Chen et al. 2020b), and is used as the baseline algorithm of our method. Specifically, given a skeleton sequence s, the positive pair (x, x') is constructed via $\mathcal{T}$. Subsequently, we can obtain the corresponding feature representations (z, z') via the query/key encoder $f(\cdot)$ and embedding projector $h(\cdot)$, respectively. A memory queue $\mathbf{M}$ is maintained storing lots of negative samples for contrastive learning. The whole network is optimized by InfoNCE loss (Oord, Li, and Vinyals 2018):

$$\mathcal{L}_{Info} = -\log \frac{\exp(z \cdot z'/\tau)}{\exp(z \cdot z'/\tau) + \sum_{i=1}^{M} \exp(z \cdot m_i/\tau)}, \quad (1)$$

where m_i is the feature in $\mathbf{M}$ corresponding to the i-th negative sample, M is the number of negative features and τ is the temperature hyper-parameter. After each training step, all samples in a batch will be updated to $\mathbf{M}$ as negative samples in a first-in, first-out policy. The key encoder is a momentum-updated version of the query encoder that is updated via gradients. Concretely, denoting the parameters of query encoder and key encoder as θ_q and θ_k respectively, the key encoder is updated as: $\theta_k \leftarrow m\theta_k + (1-m)\theta_q$, where $m \in [0, 1)$ is a momentum coefficient.

3.2 Hierarchical Consistent Contrastive Learning

Traditional contrastive learning works directly apply the augmentation sets at once to generate positive pairs. When strong augmentations are applied, these positive samples heavily suffer from semantic information loss, sharing less correlation. However, it is quite difficult to learn useful information from the consistency constraint of these degraded pairs. To address this problem, we propose a hierarchical consistent contrastive learning framework. We generate a series of highly correlated positive pairs progressively via gradually growing augmentations. Therefore, these pairs provide hierarchical guidance of the feature similarity and benefit the model in learning the knowledge from strong augmentations with consistency of different views.

We first give an overview of our method. As shown in Figure 2, HiCLR has multiple branches to extract features and mainly comprises two components: (1) A gradual growing augmentation policy which constructs multiple positive pairs corresponding to the different augmentations. (2) Asymmetric hierarchical learning constraint of the representation consistency from strongly augmented views. Next, we will introduce each component in detail.

1) Gradual growing augmentation. To facilitate the learning process to achieve better augmentation invariance, a gradual growing augmentation policy is introduced. The augmentation policy consists of multiple augmentation sets, each of which is an extended version of the existing one. By virtue of this, multiple ordered positive pairs with increasing distortion are generated.

Here we give a formal description of our growing augmentation pipeline. The proposed growing hierarchical augmentation policy constructs the following augmentation sets: $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_{k-1}$, where k is the total number of different augmentation sets. $\mathcal{T}_0$ contains the basic augmentation strategy, and each set adopts more augmentation instances than the previous one. These sets can be formulated as:

$$\begin{aligned}\mathcal{T}_0 &= \{a_{0,0}\}, \\ \mathcal{T}_1 &= \{a_{0,1},\ a_{1,1}\}, \\ &\ldots \\ \mathcal{T}_{k-1} &= \{a_{0,k-1},\ a_{1,k-1},\ \ldots,\ a_{k-1,k-1}\},\end{aligned} \quad (2)$$

where $a_{i,j}$ represents the instances sampled from the i-th augmentation strategy belonging to the j-th augmentation set. Note that we re-sample the instances of each augmentation strategy in each augmentation set, which means that $a_{i,j} \neq a_{i,j'}$, $j \neq j'$. The re-sampling strategy further expands the feature distribution, enabling the model to learn a more distinguishable feature space for the downstream task.

Resorting to this module, we construct $k-1$ ordered positive pairs $(v_0, v_1), \ldots, (v_{k-2}, v_{k-1})$, where $v_i = \mathcal{T}_i(s)$. Different from the previous works, the augmentations applied to one positive pair are different, which allows the model to be directional in its feature clustering. Meanwhile, we can also obtain the basic positive pair as described in Section 3.1, (v_0, v_0') via $\mathcal{T}_0$ (for $query$ and key). The gradual growing augmentation policy enables the model to treat augmentations differently by adjusting their applied branches and decoupling the learning of different augmentations.

2) Asymmetric hierarchical learning. Previous works based on contrastive learning utilize InfoNCE loss in Equation (1) for the representation learning. However, it often leads to performance drop when applying strong augmentations, which is caused by the little mutual information among different augmented views. To this end, a hierarchi-

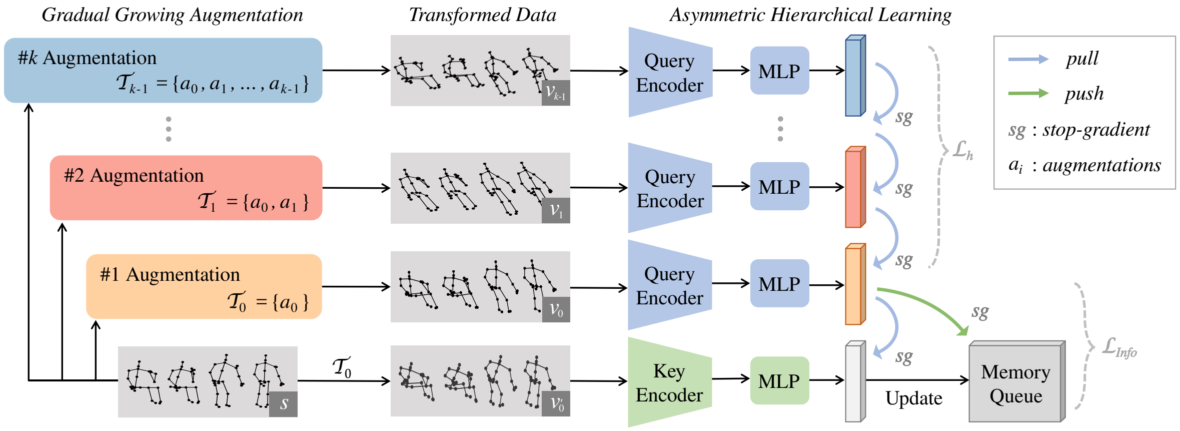

Figure 2: The overview architecture of the proposed HiCLR. There are k branches sharing the same query encoder weights corresponding to the hierarchical learning of different augmentations. The augmented view v_i is fed into the query encoder f_{θ_q} and the embedding projector h_{θ_q} to obtain z_i. Similarly, z_0' is obtained by the key encoder f_{θ_k} and the embedding projector h_{θ_k}. Meanwhile, a hierarchical self-supervised loss is proposed to align the feature distributions of adjacent branches, which is optimized jointly with the InfoNCE loss.

cal self-supervised learning objective is proposed to learn the representation consistency of multiple augmented views.

As shown in Figure 2, the positive pairs are first encoded to the feature embeddings. Formally, for a skeleton sequence s, we construct the positive pairs (v_0, v_1), ... , (v_{k-2}, v_{k-1}) and (v_0, v_0') as discussed above. Then, the query encoder f_{θ_q} and the MLP head h_{θ_q} are applied successively to extract the feature representations:

$$z_i = h_{\theta_q}\left(f_{\theta_q}\left(v_i\right)\right), i = 0, 1, \ldots, k-1. \tag{3}$$

Similarly, we can obtain the feature representation z_0' via the key encoder f_{θ_k} and the MLP h_{θ_k}:

$$z_0' = h_{\theta_k}\left(f_{\theta_k}\left(v_0'\right)\right). \tag{4}$$

The model optimizes the feature similarity of different augmented views (v_{i-1}, v_i) to learn the representation consistency among adjacent branches. Since these adjacent views share more augmentation strategies, it allows the target features to converge more smoothly to the center of the latent cluster. However, according to the previous work (Bai et al. 2022), it may lead to performance drop when the strongly augmented view is used as a mimic target due to the serious distortions. Therefore, we design an asymmetric loss to unilaterally pull the features closer. The hierarchical self-supervised learning objective is computed by the feature similarity of adjacent branches and can be formulated as:

$$\mathcal{L}_h = \sum_{i=1}^{k-1} sim\left(z_i, \text{stopgrad}\left(z_{i-1}\right)\right). \tag{5}$$

Here, we utilize the stop-gradient (stopgrad) operation to take a more confident target for similarity learning. The strongly augmented view z_i is constrained to reduce the feature distance from the weakly augmented view z_{i-1}, but not vice versa. $sim(\cdot)$ can be any function that measures the similarity between two feature embeddings, such as cosine similarity and Kullback–Leibler (KL) divergence (Kullback and Leibler 1951). This can be viewed as an asymmetric design of representation consistency learning, which is adopted in Simsiam (Chen and He 2021), BYOL (Grill et al. 2020), and CO2 (Wei et al. 2020). Through the asymmetric hierarchical learning from multiple positive pairs, the model exploits the rich information brought by the strong augmentations and further improves the generalization capacity to downstream tasks.

3) Instantiation. We next give an instantiation of our method. For the asymmetric hierarchical learning, we use KL divergence as the $sim(\cdot)$ function. One problem is that it is difficult to calculate an ideal accurate distribution of feature z_i. Inspired by (Wang and Qi 2021), we obtain the conditional distribution of feature z_i with the positive feature output by the key encoder and numerous negative features maintained in $\mathbf{M}$. Specifically, the conditional distribution for z_i is given as follows:

$$p\left(z|z_i\right) = \frac{\exp(z \cdot z_i/\tau)}{\exp(z_0' \cdot z_i/\tau) + \sum_{i=1}^{M} \exp(m_i \cdot z_i/\tau)}. \tag{6}$$

Equation (6) depicts the similarity distribution of feature z_i measured by positive features and negative features. According to Wang and Qi's discovery (Wang and Qi 2021), the distributions of $p\left(z|z_i\right)$ and $p\left(z|z_{i-1}\right)$ are similar via a randomly initialized network. It inspires us to optimize the distribution distances between $p\left(z|z_i\right)$ and $p\left(z|z_{i-1}\right)$, *i.e.*, $D_{KL}\left(\text{stopgrad}(p\left(z|z_{i-1}\right)), p\left(z|z_i\right)\right)$ as $sim(\cdot)$, to learn the consistency between different augmented views. Also, we apply the $\mathcal{L}_{Info}$ on the basic positive pairs (z_0, z_0') and jointly optimize the model. The overall loss is given by:

$$\mathcal{L} = \mathcal{L}_{Info} + \lambda_h \mathcal{L}_h, \tag{7}$$

where λ_h is the weight for hierarchical self-supervised loss.

For augmentations, the model is instantiated as k=3 with *Basic Augmentation Set* (for $a_{0,*}$), *Normal Augmentation Set* (for $a_{1,*}$) and *Random Mask* (for $a_{2,*}$). We will discuss more about strong augmentations including *Random Mask* in the next Section.

3.3 Strong Augmentation for Skeleton

We first introduce the *Basic Augmentation Set* and *Normal Augmentation Set* following the previous works (Li et al. 2021; Guo et al. 2022; Rao et al. 2021):

- *Basic Augmentation Set* (*BA*) contains a spatial transformation *Shear* and a temporal transformation *Crop*.
- *Normal Augmentation Set* (*NA*) contains the following augments: *Spatial Flip*, *Rotation*, *Gaussian Noise*, *Gaussian Blur*, and *Channel Mask*.

In addition to the augmentations above, we consider the strong augmentations targeted at skeleton data to introduce more novel patterns for the representation learning. The augmentations for the skeleton are divided into three categories:

- **Semantic-Dependent Augmentation.** Since human skeleton sequences have natural semantic information, we can perform linear transformations (*e.g.*, rotation, scaling) or nonlinear transformations (*e.g.*, joint replacement, flip) on 3D skeleton data to keep the essential semantic unchanged and computationally available.
- **Feature-Wise Augmentation.** We can apply the disturbance to the features of graph nodes which is the human joints for skeleton data. It enables the model to obtain more robust representations for the noise in collecting data, such as the error caused by the camera view.
- **Structure-Wise Augmentation.** Considering the topology of the human body, we hope that the model can output consistent semantic information under a slight perturbation of the joint adjacency graph. It is because human action is often global and slight structural perturbations can be compensated by the information aggregation at other joints.

Based on this categorization, we propose the following three strong augmentation strategies as our *Strong Augmentation Set* for skeleton:

- *Random Mask.* A random mask for the spatial-temporal 3D coordinate data of the joints. It can be viewed as a random perturbation of the joint coordinates.
- *Drop/Add Edges (DAE).* We randomly drop/add connections between different joints in each information aggregation layer. The target to be augmented is the predefined or learnable adjacency matrix for the graph convolution layer and the attention map for the transformer block.
- *SkeleAdaIN.* Inspired by the practice of style transfer (Huang and Belongie 2017; Karras, Laine, and Aila 2019), we exchange statistics of two skeleton samples on the spatial-temporal dimension, *i.e.*, the mean and the variance of the style sample are transferred to the content sample, to generate the augmented views. Since this transformation does not change the relative order of joint coordinates, we maintain the semantics of skeleton sequences unchanged.

More details can be found in the Appendix. These augmentations cause varying degrees of performance degradation when applied directly as shown on the right of Table 3. Therefore, we regard these augmentations as the examples of **strong augmentations** for the skeleton.

4 Experiment Results

4.1 Dataset

1) NTU RGB+D Dataset 60 (NTU60) (Shahroudy et al. 2016) is a large-scale dataset that contains 56,578 samples with 60 action categories and 25 joints. We follow the two recommended protocols: a) Cross-Subject (xsub): the data for training and testing are collected from different subjects. b) Cross-View (xview): the data for training and testing are collected from different camera views.

2) NTU RGB+D Dataset 120 (NTU120) (Liu et al. 2019) is an extension to NTU60. 114,480 videos are collected with 120 action categories. Two recommended protocols are adopted: a) Cross-Subject (xsub): the data for training and testing are collected from 106 different subjects. b) Cross-Setup (xset): the data for training and testing are collected from 32 different setups.

3) PKU Multi-Modality Dataset (PKUMMD) (Liu et al. 2020a) is a large-scale dataset covering a multi-modality 3D understanding of human actions with almost 20,000 instances and 51 action labels. Two subsets are divided: Part I is an easier version; Part II provides more challenging data caused by large view variation and the cross-subject protocol is adopted.

4.2 Implementation Details and Evaluation

We evaluate the performance of our method on both ST-GCN (Yan, Xiong, and Lin 2018) and DSTA-Net (Shi et al. 2020) as backbones. The experimental settings of pre-training are following the previous works (Li et al. 2021; Guo et al. 2022) for a fair comparison. All skeleton data are pre-processed into 50 frames. We reduce the number of channels in each graph convolution layer to 1/4 of the original setting for ST-GCN and 1/2 for DSTA-Net, respectively. The dimension of the final output feature is 128 and the size of the memory bank $\mathbf{M}$ is set to 32,768. The model is trained for 300 epochs with a batch-size of 128 using the SGD optimizer. λ_h is set to 0.5. A multi-stream fusion strategy is adopted following the previous works, *i.e.*, a weighted fusion of joint, bone, and motion streams. We adopt the following protocols to give a comprehensive evaluation:

1) KNN Evaluation. We apply a K-Nearest Neighbor (KNN) classifier which is a non-parametric supervised learning method. It directly reflects the quality of the feature space learned by the encoder.

2) Linear Evaluation. A linear classifier is applied to the fixed encoder for linear evaluation. The classifier is trained to predict the corresponding label of the input sequences.

3) Semi-supervised Evaluation. In semi-supervised evaluation, we pre-train the encoder with all unlabeled data, and then train the whole model with randomly sampled 1%, 10% of the training data.

4) Supervised Evaluation. We fine-tune the whole model after pre-training the encoder. Both the encoder $f(\cdot)$ and classifier are trained for the downstream task.

4.3 Ablation Study

We first conduct ablation studies to give a more detailed analysis of our method. All results reported in this section

Augmentation	Stream	xsub (%)	xview (%)
Baseline	Joint	68.3	76.4
Random Mask		**77.6**	82.0
DAE		77.2	81.7
SkeleAdaIN		77.3	**82.4**
Baseline	Bone	69.4	67.4
Random Mask		73.9	78.0
DAE		**76.9**	**80.5**
SkeleAdaIN		75.3	79.2
Baseline	Motion	53.3	50.8
Random Mask		69.1	**74.3**
DAE		68.0	72.2
SkeleAdaIN		**69.5**	71.8
Baseline	Ensemble	75.0	79.8
Random Mask		**80.4**	**85.5**
DAE		79.8	84.9
SkeleAdaIN		**80.4**	84.4

Table 1: Ablation studies on the strong augmentations. Ensemble represents the fusion of joint-bone-motion streams.

are under linear evaluation on NTU60 dataset.

1) Strong Augmentation Analysis. We set *BA, NA* as the hierarchical augmentation set of the first and the second branch, and give an analysis when introducing the different strong augmentations as the extra augmentation for the third branch. The linear evaluation results are shown in Table 1. Compared with the baseline, the model performance is significantly improved by applying the proposed strong augmentations with our HiCLR. Meanwhile, we also find some interesting results:

(a) Different streams correspond to the different optimal augmentation methods. For example, *DAE* performs significantly better than the other two augmentations on the bone stream. This may be relative to the association between invariances and streams, *i.e.*, the bone view of the skeleton data implies the topological information of the human body structure, which can be more robust to *DAE* augmentation.

(b) The performance of the same augmentation strategy can have a marked difference under different protocols. As shown in Table 1, *SkeleAdaIN* gives better results under cross-subject protocol than those under cross-view protocol. This is because *SkeleAdaIN* can be regarded as a linear transformation of the action sequences under the same view, and the statistics which usually contain information about the performer's body shapes and range of motions are exchanged. Therefore, better robustness can be obtained under a cross-subject evaluation protocol.

These results indicate that as a high-level representation, skeleton data faces more challenges in contrastive learning research. More efforts are needed in the design of augmentations for the skeleton data. To make our method more general, we finally adopt the *Random Mask* augmentation in the third branch of our implementation.

2) Data Augmentation Arrangement. Table 2 shows the results of different augmentation arrangements, where BA, NA, and Mask represent the *Basic*, *Normal Augmentation Set*, and *Random Mask* augmentation, respectively. As we can see, different arrangements can have a marked influence on the results which demonstrates the necessity of making a discriminate treatment for different augmentations. It is found that the optimal method approximates a kind of arrangement from weak augmentations to strong augmentations. This also proves that strong augmentation is not suitable as the basic augmentation strategy, confirming our hierarchical learning ideas from easy to difficult.

Arrangement	k	xsub (%)	xview (%)
[BA, NA, Mask]	3	77.6	**82.0**
[NA, BA, Mask]	3	**77.8**	80.3
[Mask, BA, NA]	3	74.7	79.7
[BA+NA, Mask]	2	74.0	79.0
[BA, NA+Mask]	2	76.7	79.4

Table 2: Ablation studies on the data augmentation arrangement of the single joint stream.

$sim(\cdot)$	Acc.
Cosine	75.8%
L1	73.6%
KL div.	**77.6%**

Augmentation	Baseline	Ours
[BA]	68.3%	-
[BA, NA]	72.9%	**76.8%**
[BA, NA, Mask]	56.7%	**77.6%**
[BA, NA, DAE]	65.5%	**77.2%**
[BA, NA, AdaIN]	13.2%	**77.3%**

Table 3: The accuracy is reported under cross-subject protocol. Left: The effect of different similarity functions in hierarchical self-supervised loss. Right: Ablation studies of the hierarchical design when applying different augmentations.

3) Hierarchical Consistent Learning. As shown on the left of Table 3, KL divergence gives the best results as $sim(\cdot)$ function, indicating that the distribution of $p\left(z|z_i\right)$ and $p\left(z|z_{i-1}\right)$ should be similar for a well-pre-trained model (Wang and Qi 2021). It can be regarded as a soft version of InfoNCE loss, which introduces more samples to measure and constrain the consistency of different augmented views. Meanwhile, the results when more and strong augmentations are applied are shown on the right of Table 3. As we can see, HiCLR can bring a consistent improvement even though some augmentations such as *Random Mask*, *DAE*, and *SkeleAdaIN* show adverse effects on the baseline algorithm, verifying the effectiveness of HiCLR.

4.4 Comparison with State-of-the-art Methods

We compare our method with the state-of-the-art methods for self-supervised skeleton-based action recognition under different evaluation protocols.

1) Linear Evaluation Results. We use both GCNs and transformers as our backbone to comprehensively demonstrate the effectiveness of our approach. First, compared with other GCN-based methods (Zheng et al. 2018; Lin et al. 2020; Su, Liu, and Shlizerman 2020; Rao et al. 2021; Thoker, Doughty, and Snoek 2021; Nie, Liu, and Liu 2020; Li et al. 2021; Guo et al. 2022; Yang et al. 2021), HiCLR has achieved the best performance on NTU datasets as shown in Table 4. By virtue of the hierarchical design, our method benefits better from strong augmentations and significantly outperforms the results of other methods. Compared with AimCLR (Guo et al. 2022), which also considers the strong augmented views, we obtain a notable improvement on both

Method	Backbone	Params	NTU60		NTU120	
			xsub (%)	xview (%)	xsub (%)	xset (%)
LongT GAN (AAAI 18)	GRU	40.2M	39.1	48.1	-	-
MS^{2}L (ACM MM 20)	GRU	2.28M	52.6	-	-	-
P&C (CVPR 20)	GRU	-	50.7	76.3	42.7	41.7
AS-CAL (Information Sciences 21)	LSTM	0.43M	58.5	64.8	48.6	49.2
ISC (ACM MM 21)	GRU+GCN+CNN	10.0M	76.3	85.2	67.1	67.9
SeBiReNet (ECCV 20)	GRU	0.27M	-	79.7	-	-
3s-CrosSCLR (CVPR 21)	ST-GCN	0.85M	77.8	83.4	67.9	66.7
3s-AimCLR (AAAI 22)	ST-GCN	0.85M	78.9	83.8	68.2	68.8
Ours	ST-GCN	**0.85M**	**80.4**	**85.5**	**70.0**	**70.4**
H-Transformer (ICME 21)	Transformer	>100M	69.3	72.8	-	-
GL-Transformer (ECCV 22)	Transformer	214M	76.3	**83.8**	66.0	68.7
Ours	Transformer	**1.56M**	**78.8**	83.1	**67.3**	**69.9**

Table 4: Linear evaluation results on NTU60 and NTU120 datasets.

Method	Stream	NTU60 (%)		PKUMMD Part I (%)
		xsub	xview	
SkeletonCLR	Joint	56.1	61.7	68.9
AimCLR		62.0	71.5	72.0
Ours		**67.3**	**75.3**	**73.8**
SkeletonCLR	Motion	37.4	41.6	51.0
AimCLR		50.8	56.9	60.6
Ours		**55.3**	**60.7**	**63.8**

Table 5: KNN evaluation results of different streams.

Method	*1% data*		*10% data*	
	xsub	xview	xsub	xview
ASSL (20)	-	-	64.3	69.8
MCC (21)	-	-	55.6	59.9
3s-CrosSCLR (21)	51.1	50.0	74.4	77.8
3s-Colorization (21)	48.3	52.5	71.7	78.9
3s-AimCLR (22)	54.8	54.3	78.2	81.6
Ours (GCN)	**58.5**	**58.3**	**79.6**	**84.0**
3s-Hi-TRS (22)	49.3	51.5	77.7	81.1
Ours (Transformer)	**54.7**	**53.7**	**82.1**	**84.8**

Table 6: Semi-supervised results on NTU60 dataset.

Method	Params	Protocol	
NTU60 Dataset		xsub	xview
3s-CrosSCLR (CVPR 21)	0.85M	86.2	92.5
3s-AimCLR (AAAI 22)	0.85M	86.9	92.8
Ours (GCN)	0.85M	**88.3**	**93.2**
3s-Hi-TRS (ECCV 22)	7.05M	90.0	**95.7**
Ours (Transformer)	**1.56M**	**90.4**	**95.7**
NTU120 Dataset		xsub	xset
3s-CrosSCLR (CVPR 21)	0.85M	80.5	80.4
3s-AimCLR (AAAI 22)	0.85M	80.1	80.9
Ours (GCN)	0.85M	**82.1**	**83.7**
3s-Hi-TRS (ECCV 22)	7.05M	85.3	87.4
Ours (Transformer)	**1.56M**	**85.6**	**87.5**

Table 7: Supervised results on NTU dataset.

the single joint stream (**77.6%** vs. 74.3% on xsub and **82.0%** vs. 79.7% on xview) and the fusion results.

We also compare the latest works using transformers (Cheng et al. 2021; Kim et al. 2022) as shown in Table 4. HiCLR uses only one percent of the model parameters to achieve comparable or better performance than others, indicating the efficiency and effectiveness of our method.

2) KNN Evaluation Results. The KNN evaluation is a direct reflection of the quality of the feature space (Wu et al. 2018). In Table 5, we can see that our method outperforms the SkeletonCLR and AimCLR by a large margin on both the joint and motion streams. It indicates that a higher quality feature space is learned by the model owing to the introduction of more strong augmentations.

3) Semi-supervised Evaluation Results. The semi-supervised results are presented in Table 6. Our method can significantly improve the performance in semi-supervised learning compared with GCN-based methods (Si et al. 2020; Su, Lin, and Wu 2021), especially when there is little training data available. Meanwhile, a remarkable gain is observed when using transformers as the backbone. Compared with Hi-TRS (Chen et al. 2022), HiCLR improves the semi-supervised results by a large margin, verifying the strong representation ability of our method.

4) Supervised Evaluation Results. We conduct supervised evaluation experiments and the results are shown in Table 7. Compared with other GCN-based methods, HiCLR consistently outperforms other methods, especially on NTU120 dataset. Moreover, for the latest transformer-based method, our method can exceed Hi-TRS (Chen et al. 2022) with fewer parameters and renews the state-of-the-art score.

5 Conclusion

In this paper, we propose a new hierarchical contrastive learning framework, HiCLR, to fully take advantage of the strong augmentations. Instead of learning all augmentations without distinction, HiCLR learns from hierarchical consistency with growing augmentations, alleviating the difficulty in learning consistency from the strongly augmented views. An asymmetric loss is applied to align the feature extracted from the strongly augmented view to the one from the weakly augmented view. Extensive experiments verify the effectiveness of HiCLR for GCNs and transformers as backbones. HiCLR can generate a more distinguishable feature space and outperforms the state-of-the-art methods under various protocols.

References

Bai, Y.; Yang, Y.; Zhang, W.; and Mei, T. 2022. Directional Self-Supervised Learning for Heavy Image Augmentations. In *IEEE CVPR*, 16692–16701.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved baselines with momentum contrastive learning. arXiv:2003.04297.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *IEEE CVPR*, 15750–15758.

Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; and Hu, W. 2021. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *IEEE ICCV*, 13359–13368.

Chen, Y.; Zhao, L.; Yuan, J.; Tian, Y.; Xia, Z.; Geng, S.; Han, L.; and Metaxas, D. N. 2022. Hierarchically Self-Supervised Transformer for Human Skeleton Representation Learning. arXiv:2207.09644.

Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; and Lu, H. 2020. Skeleton-based action recognition with shift graph convolutional network. In *IEEE CVPR*, 183–192.

Cheng, Y.-B.; Chen, X.; Chen, J.; Wei, P.; Zhang, D.; and Lin, L. 2021. Hierarchical transformer: Unsupervised representation learning for skeleton-based human action recognition. In *IEEE ICME*, 1–6.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *IEEE CVPR Workshops*, 702–703.

Du, Y.; Wang, W.; and Wang, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *IEEE CVPR*, 1110–1118.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 33: 21271–21284.

Guo, T.; Liu, H.; Chen, Z.; Liu, M.; Wang, T.; and Ding, R. 2022. Contrastive Learning from Extremely Augmented Skeleton Sequences for Self-supervised Action Recognition. In *AAAI*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *IEEE CVPR*, 9729–9738.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE ICCV*, 1501–1510.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *IEEE CVPR*, 4401–4410.

Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; and Boussaid, F. 2017. A new representation of skeleton sequences for 3d action recognition. In *IEEE CVPR*, 3288–3297.

Kim, B.; Chang, H. J.; Kim, J.; and Choi, J. Y. 2022. Global-local Motion Transformer for Unsupervised Skeleton-based Action Learning. arXiv:2207.06101.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1): 79–86.

Li, L.; Wang, M.; Ni, B.; Wang, H.; Yang, J.; and Zhang, W. 2021. 3d human action representation learning via cross-view consistency pursuit. In *IEEE CVPR*, 4741–4750.

Lin, L.; Song, S.; Yang, W.; and Liu, J. 2020. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *ACM MM*, 2490–2498.

Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.-Y.; and Kot, A. C. 2019. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE TPAMI*, 42(10): 2684–2701.

Liu, J.; Song, S.; Liu, C.; Li, Y.; and Hu, Y. 2020a. A benchmark dataset and comparison study for multi-modal human action analytics. *ACM TOMM*, 16(2): 1–24.

Liu, M.; Liu, H.; and Chen, C. 2017. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68: 346–362.

Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; and Ouyang, W. 2020b. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *IEEE CVPR*, 143–152.

Nie, Q.; Liu, Z.; and Liu, Y. 2020. Unsupervised 3d human pose representation with viewpoint and pose disentanglement. In *ECCV*, 102–118.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. arXiv:1807.03748.

Plizzari, C.; Cannici, M.; and Matteucci, M. 2021. Skeleton-based action recognition via spatial and temporal transformer networks. *CVIU*, 208: 103219.

Rao, H.; Xu, S.; Hu, X.; Cheng, J.; and Hu, B. 2021. Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. *Information Sciences*, 569: 90–109.

Rodomagoulakis, I.; Kardaris, N.; Pitsikalis, V.; Mavroudi, E.; Katsamanis, A.; Tsiami, A.; and Maragos, P. 2016. Multimodal human action recognition in assistive human-robot interaction. In *IEEE ICASSP*, 2702–2706.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *IEEE CVPR*, 1010–1019.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *IEEE CVPR*, 12026–12035.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2020. Decoupled spatial-temporal attention network for skeleton-based action recognition. arXiv:1709.04875.

Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; and Blake, A. 2011. Real-time human pose recognition in parts from single depth images. In *IEEE CVPR*, 1297–1304.

Si, C.; Nie, X.; Wang, W.; Wang, L.; Tan, T.; and Feng, J. 2020. Adversarial self-supervised learning for semi-supervised 3d action recognition. In *ECCV*, 35–51.

Song, S.; Lan, C.; Xing, J.; Zeng, W.; and Liu, J. 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*.

Song, S.; Lan, C.; Xing, J.; Zeng, W.; and Liu, J. 2018a. Skeleton-indexed deep multi-modal feature learning for high performance human action recognition. In *IEEE ICME*, 1–6.

Song, S.; Lan, C.; Xing, J.; Zeng, W.; and Liu, J. 2018b. Spatio-temporal attention-based LSTM networks for 3D action recognition and detection. *IEEE TIP*, 27(7): 3459–3471.

Su, K.; Liu, X.; and Shlizerman, E. 2020. Predict & cluster: Unsupervised skeleton based action recognition. In *IEEE CVPR*, 9631–9640.

Su, Y.; Lin, G.; and Wu, Q. 2021. Self-supervised 3d skeleton action representation learning with motion consistency and continuity. In *IEEE ICCV*, 13328–13338.

Tang, Y.; Ni, Z.; Zhou, J.; Zhang, D.; Lu, J.; Wu, Y.; and Zhou, J. 2020. Uncertainty-aware score distribution learning for action quality assessment. In *IEEE CVPR*, 9839–9848.

Thoker, F. M.; Doughty, H.; and Snoek, C. G. 2021. Skeleton-contrastive 3D action representation learning. In *ACM MM*, 1655–1663.

Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; and Isola, P. 2020. What makes for good views for contrastive learning? *NeurIPS*, 33: 6827–6839.

Wang, X.; and Qi, G.-J. 2021. Contrastive learning with stronger augmentations. arXiv:2104.07713.

Wei, C.; Wang, H.; Shen, W.; and Yuille, A. 2020. Co2: Consistent contrast for unsupervised visual representation learning. arXiv:2010.02217.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE CVPR*, 3733–3742.

Xiao, T.; Wang, X.; Efros, A. A.; and Darrell, T. 2020. What should not be contrastive in contrastive learning. arXiv:2008.05659.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 7444–7452.

Yang, S.; Liu, J.; Lu, S.; Er, M. H.; and Kot, A. C. 2021. Skeleton cloud colorization for unsupervised 3d action representation learning. In *IEEE ICCV*, 13423–13433.

Zhang, J.; and Ma, K. 2022. Rethinking the Augmentation Module in Contrastive Learning: Learning Hierarchical Augmentation Invariance with Expanded Views. In *IEEE CVPR*, 16650–16659.

Zhang, Z.; and Crandall, D. 2022. Hierarchically decoupled spatial-temporal contrast for self-supervised video representation learning. In *IEEE WACV*, 3235–3245.

Zheng, N.; Wen, J.; Liu, R.; Long, L.; Dai, J.; and Gong, Z. 2018. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *AAAI*, 2644–2651.