Hierarchical Contrast for Unsupervised Skeleton-Based Action Representation Learning

Jianfeng Dong[1,2], Shengkai Sun[1], Zhonglin Liu[1], Shujie Chen[1,2*], Baolong Liu[1,2], Xun Wang[1,2]

[1]College of Computer Science and Technology, Zhejiang Gongshang University, China
[2]Zhejiang Key Lab of E-Commerce, China

Abstract

This paper targets unsupervised skeleton-based action representation learning and proposes a new Hierarchical Contrast (HiCo) framework. Different from the existing contrastive-based solutions that typically represent an input skeleton sequence into instance-level features and perform contrast holistically, our proposed HiCo represents the input into multiple-level features and performs contrast in a hierarchical manner. Specifically, given a human skeleton sequence, we represent it into multiple feature vectors of different granularities from both temporal and spatial domains via sequence-to-sequence (S2S) encoders and unified downsampling modules. Besides, the hierarchical contrast is conducted in terms of four levels: instance level, domain level, clip level, and part level. Moreover, HiCo is orthogonal to the S2S encoder, which allows us to flexibly embrace state-of-the-art S2S encoders. Extensive experiments on four datasets, *i.e.,* NTU-60, NTU-120, PKU-I and PKU-II, show that HiCo achieves a new state-of-the-art for unsupervised skeleton-based action representation learning in two downstream tasks including action recognition and retrieval, and its learned action representation is of good transferability. Besides, we also show that our framework is effective for semi-supervised skeleton-based action recognition. Our code is available at https://github.com/HuiGuanLab/HiCo.

1 Introduction

Human action recognition has a broad application in fields of human-computer interaction, intelligent surveillance, video content analysis, game control, etc (Gao, Zhang, and Xu 2019; Li et al. 2019; Dong et al. 2022; Zheng et al. 2023). In recent years, 3D skeleton-based action recognition has made remarkable progress with deep learning networks (Zhang 2012; Liu et al. 2019; Wang, Ni, and Yang 2020). However, how to learn a more discriminative representation of skeletons is still an open question for skeleton-based action recognition. To conquer this question, the majority of early works (Ni, Wang, and Moulin 2011; Wang et al. 2012; Vemulapalli and Chellapa 2016; Zhang et al. 2017; Liu et al. 2019; Cheng et al. 2020) trained a network in a fully-supervised manner, which required massive annotated 3D skeleton data and hence is expensive and time-consuming. Recently, we

(a) A typical contrastive learning framework
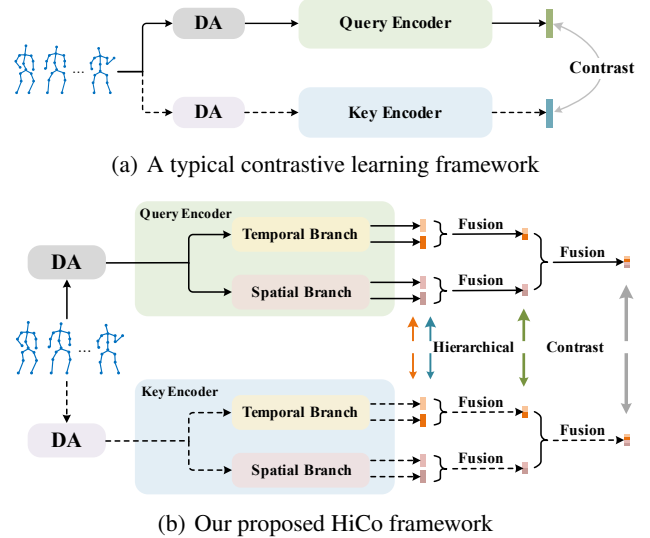

(b) Our proposed HiCo framework

Figure 1: Frameworks for unsupervised SARL. (a) A typical contrastive learning framework that represents skeleton instances at instance-level features, and performs contrast holistically. (b) Our proposed HiCo represents them at multi-level features, and performs contrast hierarchically. DA indicates data augmentation.

observe a new trend of proposing unsupervised Skeleton-based Action Representation Learning (SARL) method to liberate the workload of annotation (Zheng et al. 2018; Su, Liu, and Shlizerman 2020; Lin et al. 2020). In this paper, we also focus on unsupervised SARL.

Existing efforts on unsupervised SARL can be roughly categorized into three groups: encoder-decoder methods (Zheng et al. 2018; Nie, Liu, and Liu 2020), contrastive-learning methods (Rao et al. 2021; Thoker, Doughty, and Snoek 2021; Guo et al. 2022), and hybrid methods (Su, Lin, and Wu 2021; Chen et al. 2022). The encoder-decoder methods first encode the input skeleton sequence into latent features, then decode the latent features under the guidance of various hand-crafted pretext tasks, such as skeleton reconstruction (Zheng et al. 2018), skeleton colorization prediction (Yang et al. 2021), and skeleton displacement pre-

---

diction (Kim et al. 2022). The contrastive-learning methods usually augment an input skeleton sequence into two augmented instances, and train an encoder to make the instances of the same skeleton have more similar representations than that of different skeletons. The hybrid methods integrate the ideas of both encoder-decoder and contrastive-learning. Among them, the contrastive-learning methods are dominated in recent years due to their simple mechanism and superior performance (Li et al. 2021; Guo et al. 2022).

The first contrastive-learning method for unsupervised SARL was proposed by (Rao et al. 2021), which adapted MoCo (He et al. 2020) that was originally designed for unsupervised image representation learning to unsupervised skeleton representation learning. After that, a number of improved contrastive-learning works were proposed by modeling the uncertainty of skeletons (Su et al. 2021), exploring more positive pairs (Li et al. 2021), and utilizing superior skeleton-specific augmentations (Thoker, Doughty, and Snoek 2021). As illustrated in Figure 1(a), these contrastive-learning methods usually first represent skeleton sequences into instance-level features, and then perform instance-level contrast *holistically*. Although the contrastive-learning methods have shown much better performance, such holistic instance-level contrast may be suboptimal considering human skeletons naturally have *hierarchical* structures. A skeleton sequence can be commonly regarded as a list of whole skeletons (frames) temporally, or a list of skeleton joints spatially. Moreover, either frames or joints are the fundamental elements in the temporal or spatial domain, which can be built as larger-granularity elements such as frame clips or body parts.

Inspired by the hierarchical nature of human skeletons, we propose a new framework Hierarchical Contrast (HiCo) for unsupervised SARL. HiCo has a hierarchical encoder network to encode skeletons into part-level, clip-level, domain-level, and instance-level representations, and perform multiple-level contrast hierarchically, as shown in Figure 1(b). Specifically, the hierarchical encoder network has two branches corresponding to the temporal domain and the spatial domain respectively. For the temporal branch, it encodes a skeleton sequence into multiple clip-level features of different temporal granularities by extracting features from clips of various lengths. Similarly, the spatial branch extracts features from human parts of various sizes to obtain part-level features. Furthermore, the clip-level features and part-level features derived from temporal and spatial domains are progressively used to compose the domain-level representation and the instance-level representation. Additionally, given the multiple-level representations, a hierarchical contrast is conducted in terms of four levels. Our hypothesis is that such hierarchical contrast is consistent with the hierarchical nature of human skeletons and provides more supervision which is crucial for unsupervised learning. In summary, this paper makes the following contributions.

- We propose a hierarchical encoder network that represents skeleton sequences into multiple feature vectors of different granularities from both temporal and spatial domains via sequence-to-sequence (S2S) encoders and unified downsampling modules. Moreover, our framework is orthogonal to the S2S encoder, which allows us to flexibly embrace state-of-the-art S2S encoders.

- Based on the multi-level representations of human skeletons, we propose a new hierarchical contrastive loss for unsupervised learning, which is more effective than the plain contrastive loss over instance-level representation in the context of unsupervised SARL.

- Extensive experiments on four datasets show that our HiCo achieves a new state-of-the-art for unsupervised SARL in two downstream tasks, and its learned representation is of good transferability. Besides, we also show that our method is also effective for semi-supervised skeleton-based action recognition.

2 Related Work

2.1 Unsupervised SARL

Unsupervised methods have achieved increasing attention in various tasks (Yang et al. 2018; Deng et al. 2019; Gao et al. 2020; Liu et al. 2022) due to its good characteristic of training without human annotations. Existing works on unsupervised SARL can be roughly categorized into three groups: encoder-decoder methods, contrastive-learning methods, and hybrid methods.

Encoder-decoder methods. The early works mainly adopt the encoder-decoder paradigm, which typically encodes the input skeleton sequence into latent features, then decodes the latent features under the guidance of various hand-crafted pretext tasks. The key to this paradigm is how to design a pretext task for unsupervised SARL. Skeleton reconstruction is popular as the pretext task. For instance, (Zheng et al. 2018) learned to reconstruct the whole skeleton from the corrupted skeleton, (Kundu et al. 2019; Su, Liu, and Shlizerman 2020) reconstructed from the whole skeleton via auto-encoders. (Nie, Liu, and Liu 2020) first disentangled the skeleton into the pose-dependent and view-dependent features, then reconstructed skeletons from the disentangled features. Recently, a number of novel pretext tasks have been proposed. (Su, Lin, and Wu 2021) colorized each joint of human skeletons according to its temporal and spatial orders, and a skeleton colorization prediction pretext task was employed. (Cheng et al. 2021) devised a motion prediction pretext task that predicted the motion between adjacent frames. In (Kim et al. 2022), a multi-interval pose displacement prediction pretext task was proposed, which aimed to estimate the whole-body and joint motions at various time scales.

Contrastive-learning methods. (Rao et al. 2021) was the first to introduce contrastive learning for unsupervised SARL. In (Rao et al. 2021), a skeleton sequence was first transformed into two instances by random augmentation strategies, which were further fed into a query encoder and a key encoder respectively to obtain instance-level features. The contrast was conducted between the outputs of both the key and query encoders, encouraging augmentations of the same skeleton (positive pairs) to have more similar representations than that of different skeletons (negative pairs). Besides the positive pairs generated by random augmentation,
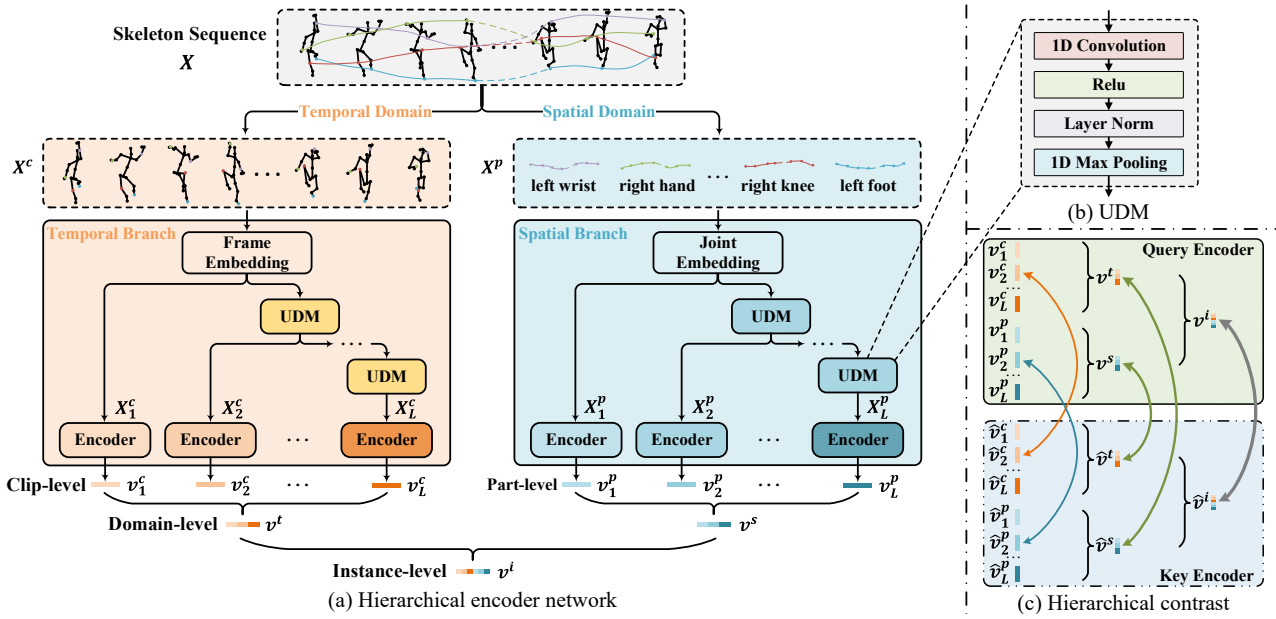
Figure 2: The illustration of our HiCo framework. It consists of (a) a hierarchical encoder network using S2S encoders with our designed (b) unified downsampling modules (UDM) that encodes a skeleton sequence at multiple granularities of both temporal and spatial domains, (c) a hierarchical contrast that conducts contrastive learning in terms of four levels jointly. Note that encoders are shared in the same branch.

(Li et al. 2021) exploited the multi-view knowledge of skeleton data for mining more positive pairs. (Thoker, Doughty, and Snoek 2021) proposed skeleton-specific spatial and temporal augmentations to enhance the spatio-temporal invariances of the learned representation. Moreover, three encoders of different architectures were utilized to conduct cross-architecture contrasts. In (Su et al. 2021), they represented skeleton sequences into a probabilistic embedding space, and the similarity was measured as the distance between probabilistic distributions.

Hybrid methods. Hybrid methods combine the idea of the encoder-decoder and contrastive-learning together (Su, Lin, and Wu 2021; Xu et al. 2021). A typical method was proposed by (Lin et al. 2020), where besides the contrastive learning between skeletons, a motion prediction and a jigsaw puzzle recognition pretext tasks were jointly employed. Different from (Lin et al. 2020) that added pretext tasks on the final learned features, (Chen et al. 2022) employed pretext tasks on the outputs of middle layers.

2.2 Hierarchical Modeling on Skeletons

There are also some previous works that modeled human skeletons hierarchically. For instance, (Du, Wang, and Wang 2015; Cheng et al. 2021; Wang, Ni, and Yang 2020) modeled the spatial hierarchical structure with hand-crafted designs, but they did not consider the temporal hierarchical structure. (Chen et al. 2022) exploited the temporal hierarchical clues from the frame, clip, and video levels. By contrast, we jointly model the temporal and spatial hierarchical structure, which can be readily implemented using our proposed more unified framework. Besides, different from the

previous models that present skeletons at the instance level, we present skeletons at the instance level, domain level, clip level, and part level.

3 Method

In this paper, we propose a hierarchical contrast framework for unsupervised SARL. As shown in Figure 2, our framework contains two main components: 1) a hierarchical encoder network that encodes a skeleton sequence at multiple granularities of both temporal and spatial domains, deriving the hierarchical multi-level representations. 2) a hierarchical contrast for unsupervised learning, which conducts contrastive learning in terms of four levels jointly. In what follows, we first depict the hierarchical encoder network, followed by the description of the hierarchical contrast.

3.1 Hierarchical Encoder Network

The hierarchical encoder network consists of a temporal branch and a spatial branch, which encodes the input skeleton sequence from the temporal domain and the spatial domain, respectively. The temporal branch encodes a skeleton sequence at multiple *temporal* granularities, resulting in a clip-level representation. Similarly, the spatial branch encodes it at multiple *spatial* granularities, resulting in a part-level representation. Both the clip-level representation and part-level representation are progressively used to assemble the domain-level and instance-level representations.

Clip-level Representation. Given a skeleton sequence $X \in \mathbb{R}^{T \times J \times 3}$ that has T frames with J joints, and each joint consists of three spatial coordinates, we first reshape it in the

time-majored domain as a list of frames $X^c = \{x_i^c\}_{i=1}^T \in \mathbb{R}^{T \times 3J}$, where x_i^c denotes the i-th frame that is essentially a whole human skeleton. Besides, a frame embedding is employed to project the input into a C-dimensional dense feature space by two fully connected layers. Formally, a projected feature from x_i^c is obtained as:

$$\widetilde{x_i^c} = W_2(\sigma(W_1 x_i^c + b_1)) + b_2, \qquad (1)$$

where $W_1 \in \mathbb{R}^{C \times 3J}$ and $W_2 \in \mathbb{R}^{C \times C}$ are transformation matrix, $b_1 \in \mathbb{R}^C$ and $b_2 \in \mathbb{R}^C$ indicate the bias vectors, and σ is the ReLU activation function. All projected features are jointly denoted as $X_1^c = \{\widetilde{x_i^c}\}_{i=1}^T \in \mathbb{R}^{T \times C}$. For ease of reference, we name these features as the initial frame representation. As the initial frame features are represented independently, they naturally lack temporal dependencies that are crucial for action representation. Therefore, we propose to model their temporal dependency with multiple temporal granularities. We first generate clips of different temporal granularities, and then utilize S2S encoders to model their temporal dependencies.

Clip Construction. In order to generate clips of different temporal granularities, we propose a Unified Donwsampling Module (UDM), which sequentially merges multiple consecutive frame/clip features to obtain more coarse-grained clips. The structure of UDM is illustrated in Figure 2(b), which can be formally defined as:

$$\text{UDM}(\cdot) = \text{MaxPool1D}(\text{LN}(\sigma(\text{Conv1D}(\cdot)))), \quad (2)$$

where Conv1D indicates a 1D convolution layer with a kernel size of 5 and stride size of 1, LN denotes the Layer Norm, and MaxPool1D is the 1D max pooling with kernel size 2. Here the 1D convolution layer is used to collect nearby contextual information and the max pooling is used for information aggregation. Note that UDM can be stacked in series, which leads to more coarse-grained clips. Additionally, we regard frames as single-frame clips at the granularity of 1 for ease of description. With the help of UDM, given clips at the granularity of n, clips of larger granularity can be readily obtained as:

$$X_{n+1}^c = \text{UDM}(X_n^c). \qquad (3)$$

By jointly employing UDM $L-1$ times in series, we are able to obtain clips of different temporal granularities $\{X_1^c, X_2^c, \ldots, X_L^c\}$.

Temporal Dependency Modeling. For each granularity of clips, we further model the temporal dependency by an S2S encoder. Concretely, given clips X_n^c, we feed it into an S2S encoder, such as GRU, followed by a temporal max pooling layer (TMP) for feature aggregation, and the clip-level feature v_n^c at the granularity of n is obtained as:

$$v_n^c = \text{TMP}(\text{S2S}(X_n^c)). \qquad (4)$$

Here, the outputs from all time steps of the S2S encoder are used. Accordingly, given $\{X_1^c, X_2^c, \ldots, X_L^c\}$ of L granularities, we are able to obtain the final clip-level representation V^c containing L feature vectors as:

$$V^c = \{v_1^c, v_2^c, \ldots, v_L^c\}. \qquad (5)$$

Note that our network is orthogonal to the S2S encoder, and any S2S encoders can be employed theoretically. In our implementation, we try GRU, LSTM, and Transformer.

Part-level Representation. Obtaining the part-level representation is similar to obtaining the clip-level representation, so we mainly specify choices that are unique at the part-level. Given a skeleton sequence $X \in \mathbb{R}^{T \times J \times 3}$, here we reshape it in the space-majored domain as a list of joints $X^p = \{x_i^p\}_{i \in J} \in \mathbb{R}^{J \times 3T}$, where x_i^p denotes the i-th joint that is essentially a joint trajectory along the time. Similarly, a joint embedding is employed to obtain the initial joint representation $X_1^p \in \mathbb{R}^{J \times C}$.

Additionally, we generate human parts of different spatial granularities by UDM. The UDM here is employed to merge nearby joints/parts to obtain parts of larger sizes. Similarly, by jointly employing UDM $L-1$ times in series, we are able to obtain parts of different spatial granularities $\{X_1^p, X_2^p, \ldots, X_L^p\}$. Furthermore, for each granularity, an S2S encoder with a max pooling layer is used to capture the spatial dependency. Consequently, we are able to obtain the final part-level representation containing clues of multiple spatial granularities as $V^p = \{v_1^p, v_2^p, \ldots, v_L^p\}$.

Domain-level and Instance-level Representation. Thus far we have a clip-level representation of multiple temporal granularities V^t and a part-level representation of multiple spatial granularities V^s. We further combine them progressively to get the domain-level representation and the instance-level representation.

For the domain-level representation, we fuse all L feature vectors of clip-level representation to derive the temporal-domain representation v^t, and fuse all the corresponding feature vectors of part-level representation to derive the spatial-domain representation v^s:

$$\begin{aligned} v^t &= F(V^t) = F(v_1^c, v_2^c, \ldots, v_L^c), \\ v^s &= F(V^p) = F(v_1^p, v_2^p, \ldots, v_L^p), \end{aligned} \qquad (6)$$

where $F(\cdot)$ denotes the fusion operator over multiple feature vectors. The concatenation is used in our implementation.

Moreover, the temporal-domain and spatial-domain representations are jointly combined into the instance-level representation v^i:

$$v^i = F(v^t, v^s). \qquad (7)$$

It is worth pointing out that though multiple-level representations have been derived, only the instance-level representation is used for the final skeleton sequence representation. Other representations are used for hierarchical contrast, which will be described in the following section.

3.2 Hierarchical Contrast

Different from the previous contrastive-learning methods that conduct contrast only at instance-level features, we propose a hierarchical contrast that conducts contrast at instance-level, domain-level, clip-level, and part-level features. For the contrastive way, we adopt the MoCo manner (He et al. 2020) that a query encoder and a key encoder are jointly used with the dynamic dictionary queue and moving-averaged update mechanism. Besides, following (Chen et al. 2020), we utilize a two-layer multilayer perceptron for feature projection before contrast. For concise descriptions, we use the same symbol of the projected features as the original

ones. We utilize ˆ to indicate the features derived from the key encoder, and m to denote the feature from the queue.

Instance-level Contrast. Following (Rao et al. 2021), we use the noise contrastive estimation loss InfoNCE (Van den Oord et al. 2018) for contrast. At the instance level, the contrastive loss is computed as:

$$\mathcal{L}_{Instance} = -\log \frac{\exp(v^i \cdot \hat{v^i}/\tau)}{\exp(v^i \cdot \hat{v^i}/\tau) + \sum_{m^i_j \in M^i} \exp(v^i \cdot m^i_j/\tau)},$$
(8)

where τ is the temperature hyper-parameter, m^i_j denotes j-th negative samples from the first-in-first-out queue M^i of previous projected instance-level features.

Domain-level Contrast. At the domain level, we expect a skeleton sequence have similar temporal-domain and spatial-domain representations. Our intuition is that the temporal-domain and spatial-domain representations are descriptions from different perspectives of the same skeleton sequence, so they essentially should be similar in high-level semantics. Therefore, we add a contrastive loss over two cross-domain features v^t and v^s, which encourages the network to pull the temporal-domain and spatial-domain features of the same skeleton sequence while pushing away that of the different skeleton sequences:

$$\mathcal{L}_{domain} = -\log \frac{\exp(v^t \cdot \hat{v^s}/\tau)}{\exp(v^t \cdot \hat{v^s}/\tau) + \sum_{m^s_j \in M^s} \exp(v^t \cdot m^s_j/\tau)}$$
$$-\log \frac{\exp(v^s \cdot \hat{v^t}/\tau)}{\exp(v^s \cdot \hat{v^t}/\tau) + \sum_{m^t_j \in M^t} \exp(v^s \cdot m^t_j/\tau)}.$$
(9)

Clip-level Contrast. Recall that clip-level representation has multiple-feature vectors of different temporal granularities. Although they are at different granularities, they also should be similar in high-level semantics. Therefore, we consider the clip-level features of different granularities of the same input instance positive. For ease of implementation, the clip-level feature at the granularity of 1 is regarded as the anchor sample, which is positive with other granularities of the same input. The contrastive loss at the clip-level is obtained by InfoNCE with more positive pairs:

$$\mathcal{L}_{clip} = -\log \frac{\sum_{l=1}^{L} \exp(v^c_1 \cdot \hat{v^c_l}/\tau)}{\sum_{l=1}^{L} \exp(v^c_1 \cdot \hat{v^c_l}/\tau) + \sum_{m^c_j \in M^c} \exp(v^c_1 \cdot m^c_j/\tau)},$$
(10)

where M^c denotes the queue containing previous projected clip-level features of all granularities.

Part-level Contrast. Similar to the clip-level contrast, the contrastive loss at the part-level is computed on part-level representation, which is defined as:

$$\mathcal{L}_{part} = -\log \frac{\sum_{l=1}^{L} \exp(v^p_1 \cdot \hat{v^p_l}/\tau)}{\sum_{l=1}^{L} \exp(v^p_1 \cdot \hat{v^p_l}/\tau) + \sum_{m^p_j \in M^p} \exp(v^p_1 \cdot m^p_j/\tau)}.$$
(11)

Finally, we train our network by minimizing the sum of the above losses. The total loss is computed as:

$$\mathcal{L}_{total} = \mathcal{L}_{instance} + \mathcal{L}_{domain} + \mathcal{L}_{clip} + \mathcal{L}_{part}.$$

4 Experiments

4.1 Experimental Setup

Datasets Experiments are conducted on four popular skeleton-based action datasets, *i.e.,* NTU-60 (Shahroudy et al. 2016), NTU-120 (Liu et al. 2019), PKU-I, and PKU-II (Liu et al. 2020).

Evaluation Metrics On NTU-60 and NTU-120, we follow two standard evaluation protocols: cross-subject (x-sub), and cross-view (x-view). Following (Lin et al. 2020), we report the x-sub results on PKU I and II. The top-1 accuracy is used for performance evaluation.

4.2 Comparison to the State-of-the-art

In this section, we compare our HiCo to the recently proposed unsupervised state-of-the-art methods, including:
(1) Encoder-decoder methods: LongT GAN (Zheng et al. 2018), EnGAN-PoseRNN (Kundu et al. 2019), P&C (Su, Liu, and Shlizerman 2020), SeBiReNet (Nie, Liu, and Liu 2020), H-Transformer (Cheng et al. 2021), Colorization (Yang et al. 2021), GL-Transformer (Kim et al. 2022). (2) Contrastive-learning methods: AS-CAL (Rao et al. 2021), CrosSCLR (Li et al. 2021), ISC (Thoker, Doughty, and Snoek 2021), AimCLR (Guo et al. 2022). (3) Hybrid methods: MS^{2}L (Lin et al. 2020), PCRP (Xu et al. 2021).

The evaluations are conducted in the context of two downstream tasks, *i.e.,* skeleton-based action recognition, and skeleton-based action retrieval, which are commonly adopted in previous works (Lin et al. 2020; Li et al. 2021; Thoker, Doughty, and Snoek 2021). For both two downstream tasks, each model should be first trained in an unsupervised manner without using any labeled data, and then used for further evaluation.

Skeleton-based Action Recognition. In this task, an extra linear classifier (a fully connected layer) is further added on the skeleton sequence representation obtained by the corresponding pre-training model, which is trained on the target dataset. Following the previous works (Li et al. 2021; Lin et al. 2020; Guo et al. 2022), the pre-training model is frozen, and only the linear classifier will be trained.

As shown in Table 1, our proposed HiCo models consistently outperform previous methods of all types with a large margin. Among the four datasets, all models on PKU-II perform the worst, which is due to that this dataset is more challenging with more noise caused by the view variation (Guo et al. 2022). Even on this challenging dataset, our best model surpasses the state-of-the-art, *i.e.,* AimCLR, by 16%. The results verify the effectiveness of our proposed framework for unsupervised SARL. In more detail, for our three HiCo variants with different S2S encoders, their performance is comparable, showing a good generality of our framework for S2S encoders. Additionally, for the compared contrastive-learning methods, they represent a skeleton sequence into instance-level features and perform contrast holistically. By contrast, we represent it into multiple-level features and perform contrast hierarchically, which enables the network to perceive more fine-grained supervision. Hence, it gains large

Method	Type	Encoder	NTU-60		NTU-120		PKU-I	PKU-II
			x-sub	x-view	x-sub	x-setup	x-sub	x-sub
LongT GAN (AAAI'18)	encoder-decoder	GRU	52.1	56.4	-	-	67.7	26.5
MS^{2}L (ACM MM'20)	hybrid	GRU	52.6	-	-	-	64.9	27.6
PCRP (TMM'21)	hybrid	GRU	54.9	63.4	43.0	44.6	-	-
AS-CAL (INS'21)	contrastive-learning	LSTM	58.5	64.8	48.6	49.2	-	-
EnGAN-PoseRNN (WACV'19)	encoder-decoder	LSTM	68.6	77.8	-	-	-	-
H-Transformer (ICME'21)	encoder-decoder	Transformer	69.3	72.8	-	-	-	-
SeBiReNet (ECCV'20)	encoder-decoder	GRU	-	79.7	-	-	-	-
CrosSCLR (CVPR'21)	contrastive-learning	GCN	72.9	79.9	-	-	84.9	21.2
AimCLR (AAAI'22)	contrastive-learning	GCN	74.3	79.7	63.4	63.4	87.8	38.5
Colorization (ICCV'21)	encoder-decoder	GCN	75.2	83.1	-	-	-	-
GL-Transformer (ECCV'22)	encoder-decoder	Transformer	76.3	83.8	66.0	68.7	-	-
ISC (ACM MM'21)	contrastive-learning	GRU&CNN&GCN	76.3	85.2	67.1	67.9	80.9	36.0
HiCo-GRU (ours)	contrastive-learning	GRU	**80.6**	**88.6**	**72.5**	**73.8**	**88.9**	**52.2**
HiCo-LSTM (ours)	contrastive-learning	LSTM	**81.4**	**88.8**	**73.7**	**74.5**	**89.4**	**54.7**
HiCo-Transformer (ours)	contrastive-learning	Transformer	**81.1**	**88.6**	**72.8**	**74.1**	**89.3**	**49.4**

Table 1: Comparisons to the state-of-the-art methods for skeleton-based action recognition downstream task.
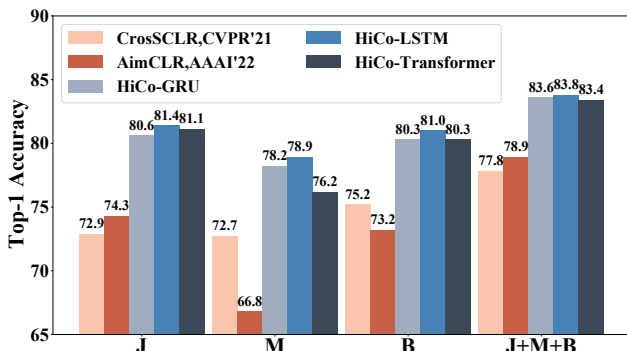

Figure 3: Performance comparison using different views (J:joint, M:motion, B:bone) of skeleton sequences in terms of x-sub on NTU-60.

improvements compared to the existing contrastive-learning methods.

Until now we only report our performance using the joint view of skeletons as input. For a more comprehensive comparison, we also try other views of skeleton sequences and three-streams fusion version of joints, motions, and bones as previous works (Li et al. 2021; Guo et al. 2022). For fusion version, we follow the setting of (Li et al. 2021). The results are summarized in Figure 3. Our models consistently outperform the counterparts with a clear margin, which further verifies the effectiveness of our framework. Besides, jointly using three views as input achieves a performance gain.

Skeleton-based Action Retrieval. In this experiment, we follow the setting in (Thoker, Doughty, and Snoek 2021; Guo et al. 2022). Specifically, given an action query, the most similar action is retrieved from training samples using

Method	NTU-60		NTU-120	
	x-sub	x-view	x-sub	x-setup
LongT GAN (AAAI'18)	39.1	48.1	31.5	35.5
P&C (CVPR'20)	50.7	76.3	39.5	41.8
AimCLR (AAAI'22)	62.0	-	-	-
ISC (ACM MM'21)	62.5	82.6	50.6	52.3
HiCo-GRU (ours)	**67.9**	**84.4**	**55.9**	**58.7**
HiCo-LSTM (ours)	**66.9**	**84.3**	**56.3**	**59.1**
HiCo-Transformer (ours)	**68.3**	**84.8**	**56.6**	**59.1**

Table 2: Comparisons to the state-of-the-art methods for skeleton-based action retrieval on NTU-60 and NTU-120.

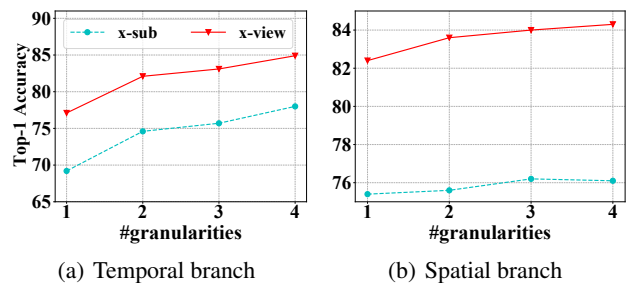(a) Temporal branch   (b) Spatial branch

Figure 4: The effectiveness of representation from multiple temporal and spatial granularities.

cosine similarity. Table 2 shows the results on NTU-60 and NTU-120. Again, our HiCo models with various encoders consistently perform better than the existing methods. The result further demonstrates that the action representation obtained by our HiCo is more discriminative.

Temporal branch	Spatial branch	x-sub	x-view
✓	-	78.0	84.9
-	✓	76.1	84.3
✓	✓	**79.8**	**87.0**

Table 3: The effectiveness of two-branch structure that jointly models temporal and spatial dependencies.

Instance	Domain	Clip&Part	x-sub	x-view
✓	-	-	79.8	87.0
✓	✓	-	80.7	88.4
✓	✓	✓	**81.1**	**88.6**

Table 4: The effectiveness of hierarchical contrast on multi-level features. All variants are based on the same representation of multiple granularities, but utilizing different losses.

4.3 Ablation Studies

To verify the effectiveness of each component in our proposed framework, we conduct ablation studies on NTU-60. All the experiments are conducted in the context of the skeleton-based action recognition downstream task, and Transformer is used as the default encoder of our model.

The effectiveness of representation on multiple granularities. Figure 4 illustrates the performance of our models using the different numbers of granularities. For both branches, the performance is gradually increased with more granularities being added. The results show the effectiveness of modeling multiple temporal and spatial granularities.

The effectiveness of two-branch structure. Table 3 shows the results of our model using one branch or two branches. The two-branch consistently performs better than the one-branch counterparts. It not only demonstrates the effectiveness of the two-branch structure, but also shows the complementary of the temporal and the spatial branches.

The effectiveness of hierarchical contrast. Table 4 summarizes the performance of conducting contrast over features of different levels, and all models use the hierarchical encoder network. It is worth noting that the model only with the instance-level contrast utilizes the same contrastive learning way of AS-CAL (Rao et al. 2021), CrosSCLR (Li et al. 2021). Progressively adding domain-level contrast, and clip&part-level contrasts gain clear improvement, which demonstrates the effectiveness of hierarchical contrast.

4.4 Transfer Learning for Action Recognition

To explore the transferability of our learned representation, we evaluate whether it learned from a source dataset generalizes well to another target dataset. Specifically, a model is first pre-trained on a source dataset in an unsupervised manner, then finetuned on a target dataset. For a cross-paper comparison, following (Lin et al. 2020), we choose NTU-60 and PKU-I as the source datasets, and PKU-II as the target dataset. All the results are evaluated under the x-sub pro-

Method	Transfer to PKU-II	
	PKU-I	**NTU-60**
LongT GAN (Zheng et al. 2018)	43.6	44.8
M^{2}L (Lin et al. 2020)	44.1	45.8
ISC (Thoker, Doughty, and Snoek 2021)	45.1	45.9
HiCo-Transformer (ours)	**53.4**	**56.3**

Table 5: Comparisons on transfer learning.

Method	x-sub		x-view	
	1%	10%	1%	10%
LongGAN (Zheng et al. 2018)	35.2	62.0	-	-
M^{2}L (Lin et al. 2020)	33.1	65.2	-	-
ASSL (Si et al. 2020)	-	64.3	-	69.8
MCC (Su et al. 2021)	-	60.8	-	65.8
Colorization (Yang et al. 2021)	48.3	71.7	52.5	**78.9**
CrosSCLR (Li et al. 2021)	-	67.6	-	73.5
Hi-TRS (Chen et al. 2022)	-	70.7	-	74.8
GL-Transformer (Kim et al. 2022)	-	68.6	-	74.9
HiCo-Transformer (Ours)	**54.4**	**73.0**	**54.8**	78.3

Table 6: Comparisons to the state-of-the-art methods with semi-supervised learning on NTU-60 dataset.

tocol. As shown in Table 5, our proposed HiCo outperforms its competitors with a large margin, which shows good transferability of our proposed action representation learning for skeleton-based action recognition.

4.5 The Potential for Semi-Supervised Learning

For semi-supervised learning, we follow the setting of (Lin et al. 2020; Chen et al. 2022), and report the results of using 1% and 10% randomly sampled training data with labels for finetuning. Table 6 summarizes the results on NTU-60. Our HiCo with Transformer as encoders performs the best on the setting of 1% data, and is slightly worse than Colorization (Yang et al. 2021) on the setting of 10% data. The results allow us to conclude that our HiCo is more effective when fewer labeled training data are available. We attribute it to our hierarchical contrast on multiple-level representation, which enables the network to acquire more supervision thus be more effective in the scenario of fewer labeled data.

5 Conclusion

This paper proposes a new contrastive-learning framework HiCo for unsupervised SARL. By jointly exploiting multiple-level representation and hierarchical contrast, HiCo is capable of learning more discriminative representation under unsupervised scenarios and has good transferability. Extensive experiments show the effectiveness of HiCo and its new state-of-the-art performance. Given HiCo is simple and effective, we believe it can also be used as a new strong baseline for unsupervised SARL.

References

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607.

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. In *European Conference on Computer Vision*, 185–202.

Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; and Lu, H. 2020. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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 *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.

Deng, C.; Yang, E.; Liu, T.; Li, J.; Liu, W.; and Tao, D. 2019. Unsupervised semantic-preserving adversarial hashing for image search. *IEEE Transactions on Image Processing*, 28(8): 4032–4044.

Dong, J.; Li, X.; Xu, C.; Yang, X.; Yang, G.; Wang, X.; and Wang, M. 2022. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4065–4080.

Du, Y.; Wang, W.; and Wang, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1110–1118.

Gao, J.; Yang, X.; Zhang, Y.; and Xu, C. 2020. Unsupervised video summarization via relation-aware assignment learning. *IEEE Transactions on Multimedia*, 23: 3203–3214.

Gao, J.; Zhang, T.; and Xu, C. 2019. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8303–8311.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 762–770.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

Kim, B.; Chang, H. J.; Kim, J.; and Choi, J. Y. 2022. Global-local motion transformer for unsupervised skeleton-based action learning. In *European Conference on Computer Vision*, 209–225.

Kundu, J. N.; Gor, M.; Uppala, P. K.; and Radhakrishnan, V. B. 2019. Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1459–1467.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4741–4750.

Li, X.; Xu, C.; Yang, G.; Chen, Z.; and Dong, J. 2019. W2VV++ fully deep learning for ad-hoc video search. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1786–1794.

Lin, L.; Song, S.; Yang, W.; and Liu, J. 2020. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2490–2498.

Liu, D.; Qu, X.; Wang, Y.; Di, X.; Zou, K.; Cheng, Y.; Xu, Z.; and Zhou, P. 2022. Unsupervised temporal video grounding with deep semantic clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1683–1691.

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 Transactions on Pattern Analysis and Machine Intelligence*, 42(10): 2684–2701.

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

Ni, B.; Wang, G.; and Moulin, P. 2011. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 1147–1153.

Nie, Q.; Liu, Z.; and Liu, Y. 2020. Unsupervised 3D human pose representation with viewpoint and pose disentanglement. In *European Conference on Computer Vision*, 102–118.

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.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1010–1019.

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 *European Conference on Computer Vision*, 35–51.

Su, K.; Liu, X.; and Shlizerman, E. 2020. Predict & cluster: Unsupervised skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9631–9640.

Su, Y.; Lin, G.; Sun, R.; Hao, Y.; and Wu, Q. 2021. Modeling the uncertainty for self-supervised 3D skeleton action representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, 769–778.

Su, Y.; Lin, G.; and Wu, Q. 2021. Self-supervised 3D skeleton action representation learning with motion consistency and continuity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13328–13338.

Thoker, F. M.; Doughty, H.; and Snoek, C. G. 2021. Skeleton-contrastive 3D action representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1655–1663.

Van den Oord, A.; Li, Y.; Vinyals, O.; et al. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Vemulapalli, R.; and Chellapa, R. 2016. Rolling rotations for recognizing human actions from 3D skeletal data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4471–4479.

Wang, J.; Liu, Z.; Wu, Y.; and Yuan, J. 2012. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 1290–1297.

Wang, M.; Ni, B.; and Yang, X. 2020. Learning multi-view interactional skeleton graph for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xu, S.; Rao, H.; Hu, X.; Cheng, J.; and Hu, B. 2021. Prototypical contrast and reverse prediction: Unsupervised skeleton based action recognition. *IEEE Transactions on Multimedia*, 1–1.

Yang, E.; Deng, C.; Liu, T.; Liu, W.; and Tao, D. 2018. Semantic structure-based unsupervised deep hashing. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 1064–1070.

Yang, S.; Liu, J.; Lu, S.; Er, M. H.; and Kot, A. C. 2021. Skeleton cloud colorization for unsupervised 3D action representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13423–13433.

Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; and Zheng, N. 2017. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE International Conference on Computer Vision*, 2117–2126.

Zhang, Z. 2012. Microsoft kinect sensor and its effect. *IEEE Multimedia*, 19(2): 4–10.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2644–2651.

Zheng, Q.; Dong, J.; Qu, X.; Yang, X.; Wang, Y.; Zhou, P.; Liu, B.; and Wang, X. 2023. Progressive localization networks for language-based moment localization. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2): 1–21.