Machine Learning (2025) 114:64
https://doi.org/10.1007/s10994-024-06667-z

Progressive semantic learning for unsupervised skeleton-based action recognition

Hao Qin[1] · Luyuan Chen[2] · Ming Kong[1] · Zhuoran Zhao[1] · Xianzhou Zeng[1] · Mengxu Lu[1] · Qiang Zhu[1]

Received: 7 May 2024 / Revised: 10 August 2024 / Accepted: 13 December 2024 /
Published online: 6 February 2025

Abstract
Traditional contrastive learning frameworks for skeleton-based action recognition use data augmentation and memory bank techniques to obtain positive/negative samples required for training, but this instance-level pseudo-label generation mechanism does not take full advantage of the rich cluster-level semantic information contained in human skeleton sequences. In this paper, we propose a Progressive Semantic Learning method (ProSL), which gradually optimizes the pseudo-label generation mechanism in self-supervised contrastive learning through an iterative framework, so that representation learning can effectively capture action semantic information. Specifically, the existing contrastive learning methods can output an initial skeleton encoder. Then, on the basis of this encoder, clustering methods can be applied to generate a Codebook containing the semantic information of human actions, which is further used to improve the pseudo-label generation mechanism. Finally, based on the above two-step iterations, we achieve progressive semantic learning and obtain a more reasonable skeleton encoder. Extensive experiments on four datasets demonstrate that our proposed method achieves SOTA on multiple downstream tasks.

Keywords Action recognition · Unsupervised learning · Semantic learning · Progressive optimization

1 Introduction

Action recognition has broad application prospects in the fields of human-computer interaction, video understanding, intelligent monitoring (Peng et al., 2020), etc. Considering that high-quality 3D skeleton annotation requires a lot of manpower, self-supervised contrastive learning has dominated the research on skeleton-based action recognition (Thoker et al., 2021; Rao et al., 2021; Guo et al., 2022; Shah et al., 2023). The core mechanism of contrastive learning is pseudo-label generation. Usually, we regard a sample and its augmented version as the positive pair, while considering it with other

Editors: Kee-Eung Kim, Shou-De Lin.

Extended author information available on the last page of the article

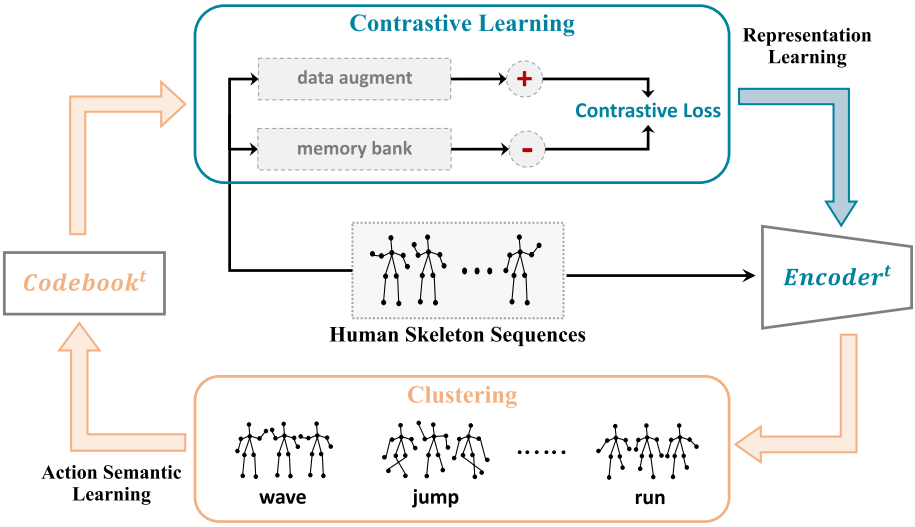


Fig. 1 Schematic diagram of the ProSL framework. The blue part in the diagram represents the contrastive representation learning module, while the orange part represents the semantic learning module we introduced. In this framework, the *Encoder* and *Codebook* iteratively progress and continuously optimize each other

samples as negative pairs. However, this sample generation mechanism tends to uniformly distribute the features of all samples on the hypersphere (Wang and Isola, 2020), without fully utilizing the inherent distribution and semantic information of the data itself (Wang and Liu, 2021).

Recently, many researchers have tried to improve the representation learning of human skeleton by improving the generation mechanism of positive and negative sample pairs in contrastive learning, such as mining positive samples from negative samples (Guo et al., 2022; Li et al., 2021; Zhang et al., 2022) or generating multi-scale positive and negative sample pairs (Dong et al., 2023; Lin et al., 2023; Hua et al., 2023). These pseudo-label optimization techniques only refine the generation of positive and negative samples at the instance-level, and do not take into account the semantic information contained in action categories, such as wave, run, jump and etc. The cluster-level semantic information contains truly valuable information that is crucial for action recognition and can be used to optimize representation learning for skeleton sequences. Therefore, how to optimize the pseudo-label generation mechanism from action semantic learning, beyond the instance-level definition of positive and negative samples, is the starting point of our work.

We propose a progressive semantic learning framework (ProSL), as shown in Fig. 1. Under unsupervised conditions, *action semantic* is an unknown variable that cannot be directly solved through mathematical methods. Therefore, we design an iterative training process to capture more accurate semantic information, primarily involving two stages: (1) **Contrastive Learning**: the input is a skeleton sequence, positive samples are generated through data augment, and negative samples come from stored samples in the memory bank. The output is the *Encoder*^{t} at the current moment; (2) **Semantic Clustering**: all skeleton sequences are encoded by the *Encoder*^{t}, and then are clustered to mine action semantic information. The output is the *Codebook*^{t} at the moment, which can be used to

optimize the next round of contrastive learning. This process is repeated continuously to form the entire progressive optimization mechanism.

Based on this iterative training process, we design novel optimization strategies for self-supervised skeleton sequence representation learning. Specifically, the regulation of the contrastive loss by the Codebook can be divided into two parts: (1) To enhance the rationality of the features extracted by the encoder, the features of each anchor not only need to be "close to positive samples and far from negative samples," but also need to be close to the semantic centroid to which they belong; (2) To avoid the convergence oscillation caused by semantically ambiguous samples, negative samples that are semantically similar to the anchor will be excluded. In addition, we introduce a new adaptive semantic branch into the contrastive framework. By regulating the encoder's input with semantic coefficients, the encoder is able to extract richer semantic information.

We conduct experiments on four widely used skeleton-based action recognition datasets, and the experimental results demonstrate that our approach achieves state-of-the-art performance across various evaluation metrics. Ablation experiments and analyses confirm the effectiveness of each sub-module within the progressive semantic learning framework. Additionally, we further substantiate the rationality of the motion features extracted by our framework through feature visualization techniques.

The main contributions can be summarized as follows:

- We propose ProSL, an iterative learning framework suitable for the skeleton-based action recognition task, to optimize the skeleton sequence representation learning process by the semantic signal obtained by clustering. During the iterative process, ProSL can progressively extract more and more precise semantic information.
- We design a novel mechanism based on the Semantic Codebook to refine the pseudo-label generation required for contrastive learning, and build more reliable positive and negative sample pairs by introducing semantic centers, evaluating semantic similarity, and calculating semantic coefficients.
- ProSL achieves the SOTAs in multiple downstream tasks on the four datasets of NTU-60, NTU-120, PKUMMD-I, and PKUMMD-II. And the comprehensive experimental and visualization results prove the effectiveness of our method.

2 Related work

2.1 Self-supervised representation learning

Self-supervised representation learning refers to training models without using actual labels, often applied in scenarios where annotating is costly or labels are scarce. Common approaches to self-supervised representation learning include pretext tasks (Noroozi et al., 2018), contrastive learning (Wu et al., 2018), prediction/reconstruction-based strategies (He et al., 2022), among others (Caron et al., 2020, 2021; Bardes et al., 2021). Contrastive learning plays a central role in self-supervised tasks. For example, SimCLR (Chen et al., 2020) trains models by leveraging data augmentation to create positive samples and treating other samples as negative, achieving strong training performance with the large batch size. MoCo (He et al., 2020) uses a memory bank to store negative samples, reducing the usage of GPU memory. To overcome the limitations of negative samples, Grill et al. (2020) devise the asymmetric BYOL model, which surpasses previous methods without

the need to calculate the loss of negative samples. However, these techniques often capture low-level features from samples, missing out on higher-level semantic information. Li et al. (2020) introduce semantic information of images into the embedding space by combining clustering with contrastive learning. Building on this, Guo et al. (2022) enable the models to learn fine-grained hierarchical image semantic information through cyclic clustering. Skeleton sequences also contain rich semantic information. Inspired by this, we try to use clustering to generate a Semantic Codebook to optimize the positive and negative sample pairs in skeleton representation learning so that the skeleton encoder can learn more reasonable embeddings.

2.2 Self-supervised skeleton representation

With the advancement of self-supervised representation learning, numerous self-supervised skeleton representation methods have emerged (Yang et al., 2021; Su et al., 2021; Kim et al., 2022; Mao et al., 2022; Zhou et al., 2023). LongT GAN (Zheng et al., 2018) pretrains a skeleton encoder through an encoder-decoder structure and employs a discriminator to provide supervisory signals. Building on this, P&C (Su et al., 2020) introduces additional prediction and clustering tasks to guide the encoder in learning more robust feature representations. Li et al. (2021) leverage contrastive learning principles to develop SkeletonCLR, also proposing cross-view positive sample mining techniques. MS^{2}L (Lin et al., 2020) suggests jointly training skeleton encoders through various unsupervised tasks, including motion prediction, jigsaw puzzle recognition, and contrastive learning. Recognizing the limited diversity of skeleton data augmentation in traditional methods, Guo et al. present AimCLR (Guo et al., 2022), a model capable of accommodating extremely enhanced augmentations, and design a more concise positive sample mining approach. HiCLR (Zhang et al., 2023) generates a richer set of positive samples through hierarchical data augmentation. ActCLR (Lin et al., 2023) extracts motion joints from human bodies and processes motion and non-motion joints differently to achieve adaptive action modeling for different body parts. Considering that local body parts are crucial for action recognition, SkeAttnCLR (Hua et al., 2023) adds a number of local feature contrastive losses to the global feature contrastive loss. HiCo (Dong et al., 2023) employs a multi-level structure to extract features at different spatial and temporal scales of skeleton sequences, performing contrastive learning at various scales to enhance the model's accuracy across multiple downstream tasks. However, most of these methods primarily handle individual samples, leading to the loss of cluster-level semantic information. Therefore, we attempt to guide the model's self-supervised training process by constructing a Semantic Codebook through clustering, aiming to enable the model to capture more accurate skeleton semantic information.

3 Method

3.1 Framework overview

We propose a progressive semantic learning framework (ProSL), as shown in Fig. 2. In traditional unsupervised skeleton representation learning methods, a dual-branch (query branch and key branch) structure is usually used for instance-level contrastive learning ($\mathcal{L}_{CL}$). However, this approach overlooks the cluster-level semantic information in skeleton sequences.

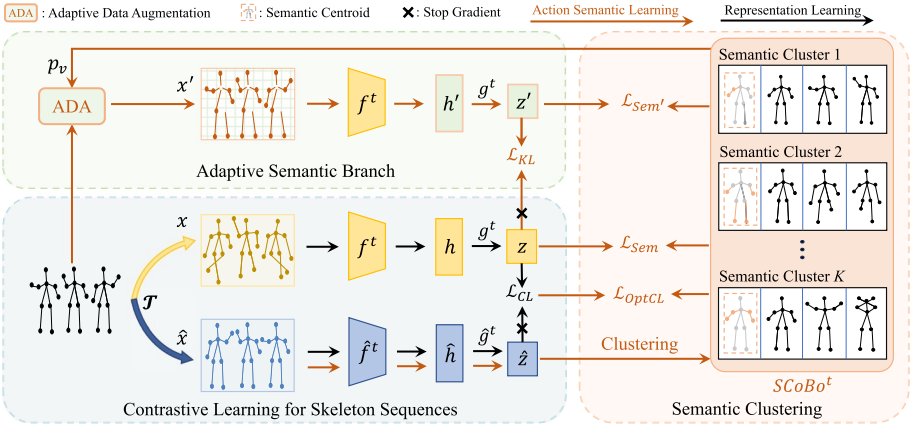


Fig. 2 The overview architecture of the proposed ProSL. ProSL adopts a three-branch structure, with the three branches from top to bottom being: *semantic branch*, *query branch*, and *key branch*. The Semantic Codebook (*SCoBo*) is obtained by clustering the output of the key branch. It can be utilized to guide the outputs of the semantic and query branches, thereby optimizing the contrastive learning process. Data augmentation for the semantic branch is controlled by the semantic coefficient, and 'ADA' stands for Adaptive Data Augmentation

Therefore, we introduce the concept of clustering to handle the features of skeleton sequences. ProSL adopts a three-branch structure, where the query branch corresponds to the gradient update branch in traditional contrastive representation learning, the key branch corresponds to the momentum update branch, and the semantic branch receives more semantically rich samples compared to the other branches. To introduce more accurate semantic information, we cluster the output $\hat{z}$ of the key branch to obtain a Semantic Codebook (*SCoBo*). *SCoBo* stores the semantic information of each sample along with k semantic centroids. It can be utilized to enhance the traditional negative sample selection process in contrastive learning ($\mathcal{L}_{OptCL}$) and compel the outputs of the query branch ($\mathcal{L}_{Sem}$) and semantic branch ($\mathcal{L}_{Sem'}$) to possess high-level semantic information. Simultaneously, the semantic coefficients computed in the semantic space can be applied to the input of the semantic branch, allowing the encoder to receive semantically richer skeleton sequences and learn more meaningful semantic information. To improve the stability of the training process, we calculate the similarity between the outputs of the semantic branch and the query branch as an auxiliary loss ($\mathcal{L}_{KL}$). In the following sections, we will provide a detailed explanation of each component in ProSL.

3.2 Contrastive learning for skeleton sequences

Contrastive Learning for Skeleton Sequences part is a classic framework for applying contrastive learning to skeleton-based action recognition, which are optimized by InfoNCE loss Oord et al. (2018):

$$\mathcal{L}_{CL} = -\log \frac{exp(z \cdot \hat{z}/\tau)}{exp(z \cdot \hat{z}/\tau) + \sum_{m_i \in M} exp(z \cdot m_i/\tau)} \tag{1}$$

where $z = f^t(g^t(x))$ and $\hat{z} = \hat{f}^t(\hat{g}^t(\hat{x}))$ represent the output of the query branch and the key branch, respectively. m_i is the feature in memory bank corresponding to the i-th negative

sample, and τ is a hyper-parameter. The encoder f and projector g of the query branch are updated by gradient descent, and the parameters of $\hat{f}$ and $\hat{g}$ depend on f and g for momentum update, respectively.

3.3 Semantic clustering

In previous self-supervised skeleton representation learning, the positive and negative sample pairs used in contrastive learning are instance-level and fixed, leading to the loss of rich cluster-level semantic information within skeleton sequences. We regulate and optimize this process through semantic clustering.

First, to obtain cluster-level semantic information from the skeleton sequences, we perform k-means clustering on the features $\hat{z}$ output by the key branch to obtain a Semantic Codebook μ *SCoBo*. As shown in Fig. 2, *SCoBo* stores k semantic (cluster) centroids along with the 'categories' to which each sample belongs. These 'categories' correspond solely to semantic information (i.e., general abstract features) and are independent of actual labels.

Next, we utilize *SCoBo* to optimize the output of the query branch. Contrastive learning is an instance-wise representation method that overlooks the underlying connections between different samples. In recognition tasks, we desire a small intra-cluster distance among sample features compared to the inter-cluster distance. Hence, we employ semantic loss to compel the output feature z of the query branch to be close to its semantic centroid. This enables the encoder to capture cluster-level semantic information:

$$\mathcal{L}_{Sem} = -\sum_{z_b \in B} Onehot(SCoBo\{z_b\}) \log y_b \tag{2}$$

where y_b represents the cosine similarity distribution between z_b and all semantic centroids, and *Onehot*() represents the one-hot encoding operation. B denotes the sample set of z. In order to make the model's clustering process more stable, we exclude some outliers in the process of calculating $\mathcal{L}_{Sem}$, so the size of B is in the range of [0, *batchsize*]. The judgment principle of outliers is as follows:

$$B = [z_b | \max\{y_b\} = SCoBo\{z_b\} \; and \; \max(y_b) > \xi] \tag{3}$$

where $\max\{y_b\}$ represents the index of the maximum value in y_b, $\max(y_b)$ represents the maximum value in y_b, and ξ is a hyper-parameter.

Finally, we utilize the semantic information within *SCoBo* to optimize the contrastive loss. The memory bank stores features of various samples, among which some features exhibit semantic similarity to query features. Treating them as negative samples could lead to the model learning vague semantics. To prevent potential positive samples within the memory bank from affecting the model's convergence, we remove samples in the memory bank that are semantically similar to z and no longer regard them as negative samples. The updated memory bank can be represented as:

$$\widetilde{M} = [m_j | SCoBo\{m_j\} \neq SCoBo\{z\}] \tag{4}$$

where $SCoBo\{m_j\}$ and $SCoBo\{z\}$ respectively represent the index of the semantic centroid to which m_j and z belong. Then the contrastive loss in Eq. 1 can be optimized with this memory bank, expressed as:

$$\mathcal{L}_{OptCL} = -\log \frac{exp(z \cdot \hat{z}/\tau)}{exp(z \cdot \hat{z}/\tau) + \sum_{m_j \in \widetilde{M}} exp(z \cdot m_j/\tau)} \tag{5}$$

3.4 Adaptive semantic branch

The quality of samples is dependent on the rationality of data augmentation in contrastive learning. Existing methods (Guo et al., 2022; Zhang et al., 2023) have attempted to add new branches in the contrastive framework to enhance the diversity of samples; however, the rationality of samples has not been explicitly constrained. Therefore, we design an adaptive semantic branch with semantically regulated data augmentation to make the encoder receive more reasonable inputs.

Each joint in a skeleton sequence contributes differently to the overall semantics. We aim to retain joints with rich semantic information as much as possible during the masking process and avoid the model from learning incorrect positive samples. To achieve this, we introduce a semantic-driven adaptive masking step. In order to explore the impact of each joint on the overall semantics, we calculate the difference in similarity between the $\hat{z}$ obtained when a certain joint is present or absent and its corresponding semantic centroid:

$$p_v = sim(\hat{z}, SCoBo(\hat{z})) - sim(\hat{z}_v, SCoBo(\hat{z})) \tag{6}$$

where v represents skeleton joint, and $\hat{z}_v$ represents the features obtained after removing the joint v. $SCoBo(\hat{z})$ represents the semantic centroid to which $\hat{z}$ belongs. $sim(,)$ is the similarity measure function, which is used to calculate the semantic ambiguity of a feature, and p_v is the semantic coefficient, which represents the influence of joint v on semantics.

If the loss of a certain joint leads to a significant increase in semantic ambiguity, we consider that this joint contains rich semantic information. Conversely, if the loss of a certain joint results in only a minor change in semantic ambiguity, we infer that this joint contributes minimally to the overall semantics. In practical applications, to simplify the computation process, we make the following approximations:

$$p_v \approx sim(\hat{z}, \hat{z}_v) \tag{7}$$

where $sim(,)$ employs KL divergence (Hershey and Olsen, 2007). We aim for joints with higher semantic coefficients to be more likely preserved during the masking process. As a result, the actual masking probability for joint v is:

$$p_v^{mask} = \varphi * \frac{1}{p_v} \tag{8}$$

where φ is the normalization coefficient.

Similar to the query branch, we employ semantic loss to compel z' to be close to its semantic centroid:

$$\mathcal{L}_{Sem'} = -\sum_{z'_b \in B'} Onehot(SCoBo\{z'_b\}) \log y'_b \tag{9}$$

where y'_b represents the cosine similarity distribution between z'_b and all semantic centroids. B' denotes the sample set of z', and uses the same outlier removal strategy as B.

In order to stabilize the training process, we introduce a KL loss to supervise z' in addition to the semantic loss:

$$\mathcal{L}_{KL} = KL(z' \cdot M || z \cdot M) = \sum_{m_i \in M} z' \cdot m_i \log \frac{z' \cdot m_i}{z \cdot m_i} \tag{10}$$

3.5 Training process

Figure 2 illustrates all the loss functions utilized in ProSL, most of which are closely related to *SCoBo*. In the early stages of training, we initially train the encoder using classical contrastive learning without generating *SCoBo*. At this point, the loss functions can be represented as:

$$\mathcal{L} = \mathcal{L}_{CL} + \mathcal{L}_{KL} \tag{11}$$

As the features output by the encoder stabilized over time, we introduce *SCoBo* to regulate the training process to obtain a better encoder, and the optimized encoder can then generate more reasonable *SCoBo*. This process is iterated continuously. At this stage, the loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{OptCL} + \mathcal{L}_{KL} + \mathcal{L}_{Sem} + \mathcal{L}_{Sem'} \tag{12}$$

For the simplicity of the model, here we directly add all loss functions with equal weights, and a more complex fusion of various loss functions may further improve the performance of the model. More analysis of the rationality of this training process can be found in Appendix A.

4 Experiments

4.1 Datasets

NTU RGB+D Dataset 60 (Shahroudy et al., 2016): NTU-60 is a comprehensive skeleton-based action recognition dataset collected using Microsoft Kinect v2 sensors. It encompasses 60 distinct actions performed by 40 individual subjects, resulting in a dataset with over 56,000 skeleton sequences. Each skeleton sequence comprises 25 joints, and each sequence features only one subject. NTU-60 offers two different dataset splits based on varying factors: Cross-Subject (x-sub) and Cross-View (x-view).

NTU RGB+D Dataset 120 (Liu et al., 2019): NTU-120 is an expanded version of the NTU-60 dataset, comprising 113,945 sequences across 120 distinct action categories, performed by 106 individual subjects. This dataset offers two evaluation protocols: the Cross-Subject (x-sub) protocol and the Cross-Setup (x-set) protocol.

PKU Multi-Modality Dataset (Liu et al., 2020): PKU-MMD is a substantial benchmark designed for continuous multi-modality 3D action understanding. The dataset encompasses nearly 20,000 action instances spanning 51 categories and involving 66 distinct subjects. PKU-MMD is segmented into two parts: Part I comprises 21,539 sequences, while Part II contains 6,904 sequences.

4.2 Implementation details

Settings: Our approach can be combined with skeleton representation models based on contrastive learning. In this paper, we use HiCo (Dong et al., 2023) as the backbone. For a fair comparison, we employ the exact same hyperparameters as HiCo-LSTM. We use two stacked LSTM in a bidirectional manner as the encoder. The output size of the encoder is set to 512, and the output dimension of projection MLP for feature projection is 128. During the training, we set the size of M and the value of τ to 2048 and 0.2, respectively. For the optimizer, we adopt the SGD method with a momentum of 0.9 and weight decay of 0.0001. The model is trained for 450 epochs with the initial learning rate of 0.01 and the learning rate is reduced to 0.001 at epoch 350, and the batch size is set to 64. For the query and key branches, we directly use the data augmentation strategies used in Thoker et al. (2021), i.e., shearing, joint jittering, and temporal cropping. For the semantic branch, in addition to the adaptive mask, we use the normal augmentation set used in Zhang et al. (2023), i.e., spatial flip, rotation, gaussian noise, gausssian blur, and channel mask. Regarding the *SCoBo*, for the NTU-60, NTU-120, and PKU-MMD datasets, we set k as 65, 130, and 55 respectively, and the value of the hyperparameter ξ is set to 0.027. We initiate *SCoBo* generation and optimize the original contrastive learning process from 400 epochs, with the *SCoBo* updated at the end of each epoch.

4.2.1 Evaluation metrics

To make a comprehensive evaluation, we evaluate the effect of our method on multiple downstream tasks: (1) Linear Evaluation, (2) Semi-Supervised Evaluation, (3) KNN Evaluation, (4) Transfer Learning.

4.3 Comparison with state-of-the-art methods

In this section, we compare our ProSL to recently proposed unsupervised state-of-the-art methods, including: LongT GAN (Zheng et al., 2018), MS^{2}L (Lin et al., 2020), CrosSCLR (Li et al., 2021), ISC (Thoker et al., 2021), AimCLR Guo et al. (2022), GL-Transformer (Kim et al., 2022), Hi-TRS (Chen et al., 2022), HaLP (Shah et al., 2023), HiCLR (Zhang et al., 2023), ActCLR (Lin et al., 2023), SkeAttnCLR (Hua et al., 2023) and HiCo (Dong et al., 2023). Asterisks (*) in tables indicate the results obtained by reproducing public code and weights. We repeat all experiments five times and take the average for presentation.

4.3.1 Linear evaluation

In the linear evaluation mechanism, we freeze the weights of the pre-trained encoder, and then add a trainable fully connected layer as the classification head. Table 1 shows the linear evaluation accuracy of our method and other classical models on NTU-60, NTU-120, PKU-MMD I and II datasets. It can be seen that under all datasets and settings, our method achieves the optimum, and for the more challenging datasets, the performance improvement of our method is greater. These results show that the features learnt by ProSL have better linear separability compared to other methods. To make the

Table 1 Comparisons to the state-of-the-art methods for skeleton-based action recognition task on NTU-60, NTU-120, PKU-MMD I and II. (3s) means three views (joint, bone, and motion) fusion, and other models use only the joint view of skeletons as the input

Method	Encoder	NTU-60		NTU-120		PKU-MMD I	PKU-MMD II
		x-sub	x-view	x-sub	x-set	x-sub	x-sub
LongT GAN (Zheng et al. 2018)	BiGRU	52.1	56.4	-	-	67.7	26.5
MS^{2}L (Lin et al., 2020)	BiGRU	52.6	-	-	-	64.9	27.6
CrosSCLR (Li et al., 2021)	GCN	72.9	79.9	-	-	-	-
ISC (Thoker et al., 2021)	BiGRU	76.3	85.2	67.1	67.9	80.9	36.0
AimCLR (Guo et al., 2022)	GCN	74.3	79.7	63.4	63.4	83.4	-
GL-Transformer (Kim et al., 2022)	Transformer	76.3	83.8	66.0	68.7	-	-
HiCLR (Zhang et al., 2023)	Transformer	78.8	83.1	67.3	69.9	-	-
ActCLR (Lin et al., 2023)	GCN	80.9	86.7	69.0	70.5	-	-
SkeAttnCLR (Hua et al., 2023)	GCN	80.3	86.1	66.3	74.5	87.3	52.9
HiCo-LSTM* (Dong et al., 2023)	BiLSTM	81.6	89.0	73.1	74.4	89.3	52.8
HiCo-LSTM-3s (Dong et al., 2023)	BiLSTM	83.8	90.4	75.8	77.1	89.3	-
ProSL (Ours)	BiLSTM	**82.5**	**89.3**	**74.0**	**74.9**	**89.8**	**54.4**
ProSL-3s (Ours)	BiLSTM	**85.3**	**91.8**	**77.1**	**78.4**	**90.8**	**59.0**

comparisons with other methods more comprehensive, Fig. 3 shows the accuracy of our method when using information from different views (joint, bone, and motion). It can be seen that ProSL can obtain stable accuracy improvement regardless of the view, and our method still achieves optimality when using information from all views. The greater performance improvement with multi-view fusion further underscores the validity of our proposed semantic clustering and the regulatory approach to contrastive learning.

4.3.2 Semi-supervised evaluation

To investigate the performance of our method under low-data conditions, we conduct a semi-supervised evaluation, where only 1% or 10% of the data are used for fine-tuning during the linear evaluation. We perform experiments on NTU-60, and the results are shown in Table 2. Compared to the baseline, with only 1% of the available training data, our method achieves performance gains of 1.6% and 2.4% under the x-sub and x-view protocols, respectively. The experimental results demonstrate the strong robustness of our method and the importance of semantic information in scenarios with limited data.

4.3.3 KNN evaluation

The k-nearest neighbor evaluation is an evaluation protocol without training parameters, which directly extracts sample features using pre-trained encoder weights and applies a k-nearest neighbor classifier (k=1) to all features for classification. It can also reflect

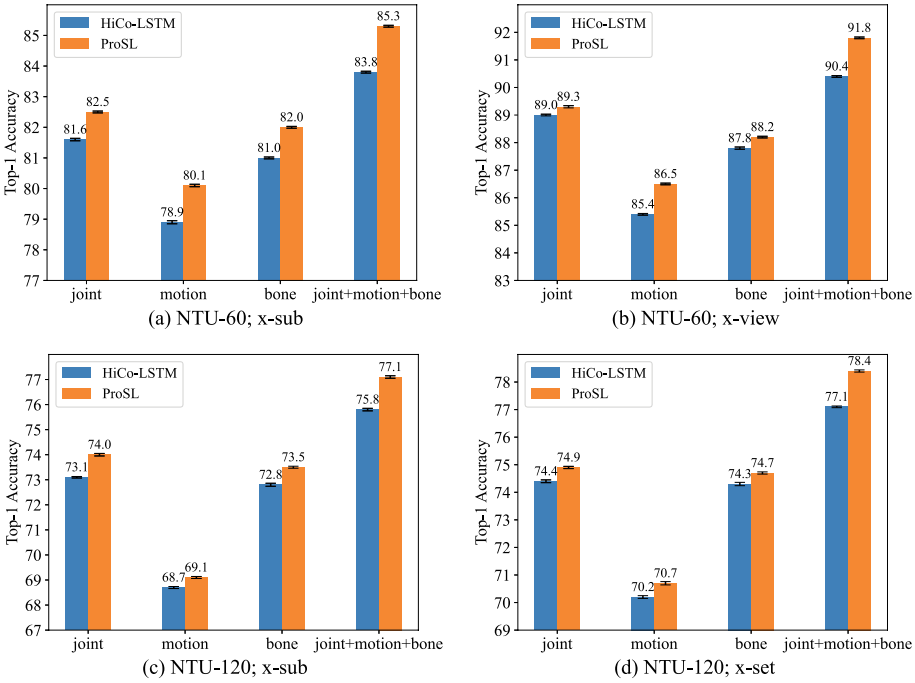


Fig. 3 Performance comparison using different views of skeleton sequences. The greater performance improvement with multi-view fusion underscores the validity of our proposed semantic clustering and the regulatory approach to contrastive learning

Table 2 Comparisons to the state-of-the-art methods with semi-supervised learning on NTU-60 dataset

Method	x-sub		x-view	
	1%	10%	1%	10%
LongT GAN (Zheng et al., 2018)	35.2	62.0	–	–
MS^{2}L (Lin et al., 2020)	33.1	65.2	–	–
ISC (Thoker et al., 2021)	35.7	65.9	38.1	72.5
CrosSCLR (Li et al., 2021)	–	67.6	–	73.5
Hi-TRS (Chen et al., 2022)	–	70.7	–	74.8
HaLP (Shah et al., 2023)	46.6	72.6	48.7	77.1
HiCo-LSTM* (Dong et al., 2023)	53.4	73.2	53.4	77.4
ProSL (Ours)	**55.1**	**74.7**	**55.7**	**78.6**

the quality of the features learned by the encoder. According to the traditional settings, we conduct action retrieval experiments on NTU-60 and NTU-120, and the results are shown in Table 3. We find that our method improves the performance of the pre-trained encoder without any fine-tuning. Compared with linear evaluation, KNN evaluation can reflect the clustering quality of the learned features. And the experiment results show that progressive semantic learning effectively improves the reasonableness of the features.

Table 3 Comparisons to the state-of-the-art methods for skeleton-based action retrieval on NTU-60 and NTU-120

Method	NTU-60		NTU-120	
	x-sub	x-view	x-sub	x-set
LongT GAN (Zheng et al., 2018)	39.1	48.1	31.5	35.5
ISC (Thoker et al., 2021)	62.5	82.6	50.6	52.3
AimCLR (Guo et al., 2022)	62.0	–	–	–
ActCLR (Lin et al., 2023)	–	78.0	–	–
SkeAttnCLR (Hua et al., 2023)	**69.4**	76.8	46.7	58.0
HiCLR (Zhang et al., 2023)	67.3	75.3	–	–
HiCo-LSTM* (Dong et al., 2023)	67.1	83.9	56.1	58.9
ProSL (Ours)	68.8	**84.2**	**56.8**	**59.5**

4.3.4 Transfer learning

To explore the cross-domain capabilities of the encoder trained by our method, we conduct unsupervised pretraining on the source dataset and perform linear evaluation on the target dataset. Following (Dong et al., 2023), we use NTU-60 and PKU-MMD I as the source datasets, and PKU-MMD II as the target dataset. All experiments are conducted under the x-sub protocol, and the results are presented in Table 4. It can be observed that our approach achieves optimal performance even in cross-dataset testing, which proves the robustness and generalization of the encoder obtained through progressive semantic learning.

4.4 Ablation study

4.4.1 Effectiveness of each sub-module

ProSL utilizes *SCoBo* to regulate contrastive learning and employs semantic coefficients for adaptive data augmentation. Here, we conduct ablation experiments to investigate whether these two sub-modules function effectively as intended, as shown in Table 5. For 'Semantic Clustering,' we train the encoder on the basis of the original contrastive learning framework using $\mathcal{L}_{OptCL}$ and $\mathcal{L}_{Sem}$. For the 'Normal Branch,' we introduce a new branch and utilize the same normal augmentation set as in Zhang et al. (2023), incorporating additional $\mathcal{L}_{KL}$ and $\mathcal{L}_{Sem'}$ for training. In the 'Semantic Branch', we replace the masking step from the 'Normal

Table 4 Comparisons on transfer learning

Method	Transfer to PKU-MMD II	
	PKU-MMD I	NTU-60
LongT GAN (Zheng et al., 2018)	43.6	44.8
MS^{2}L (Lin et al., (2020)	44.1	45.8
ISC (Thoker et al., 2021)	45.1	45.9
HiCo-LSTM* (Dong et al., 2023)	55.7	53.3
ProSL (Ours)	**56.3**	**54.1**

Table 5 Ablation experiments of sub-modules under the x-sub protocol on NTU-60. ✓ indicates that this sub-module is used

Semantic Clustering	Normal Branch	Semantic Branch	Linear	KNN
			81.6	67.1
✓			82.2	68.0
✓	✓		82.1	68.1
✓		✓	**82.5**	**68.8**

Branch' with adaptive data augmentation. It can be observed that 'Semantic Clustering' effectively enhances the performance of the encoder over traditional contrastive learning, indicating the importance of semantic information. Adding a normal branch has almost no impact on the model's accuracy, while replacing the masking approach with adaptive data augmentation further enhances the model's performance. This suggests that the adaptive semantic branch contributes to the enhancement of the encoder not through the introduction of the new branch, but rather through semantic-driven data augmentation.

4.4.2 Generalizability to the other backbone

Our starting point is to optimize the skeleton representation learning process by exploring semantic information and designing novel optimization strategies. To investigate the generalization of our method, we conduct experiments using SkeletonCLR (Li et al., 2021) as the backbone, and the experimental results are shown in Table 6. SkeletonCLR (Li et al., 2021) is a fundamental skeleton-based contrastive learning framework with significant representativeness. It can be seen that our method brings 4.8% and 2% improvement in accuracy for linear evaluation under the x-sub protocol and x-view protocol on NTU-60, respectively. This demonstrates that our method can provide a stable improvement in accuracy for the skeleton-based contrastive learning framework.

4.4.3 Impact of the number of clustering centroids

In the process of generating *SCoBo*, the number of clustering centroids, denoted as k, represents the number of action semantic categories. At different classification scales, k exhibits a certain degree of ambiguity. In practical applications, the number of possible action semantic categories may be unknown in different scenarios. Therefore, we expect the algorithm to have a certain robustness to changes in k. We conduct ablation experiments on the value of k, as shown in Fig. 4a. It can be observed that within a certain range, the accuracy of our method has been improved compared to the baseline, indicating the robustness of ProSL to variations in k. We believe this is because the semantics of actions are abstract concepts, and different numbers of semantic centroids can adaptively adjust to focus on motion categories at different scales. Within a reasonable range, the setting of k does not compromise the effectiveness of

Table 6 The linear accuracy of our method on NTU-60 when take SkeletonCLR as the backbone

Method	x-sub	x-view
SkeletonCLR (Li et al. 2021)	68.3	76.4
ProSL-SkeletonCLR (Ours)	**73.1**	**78.4**

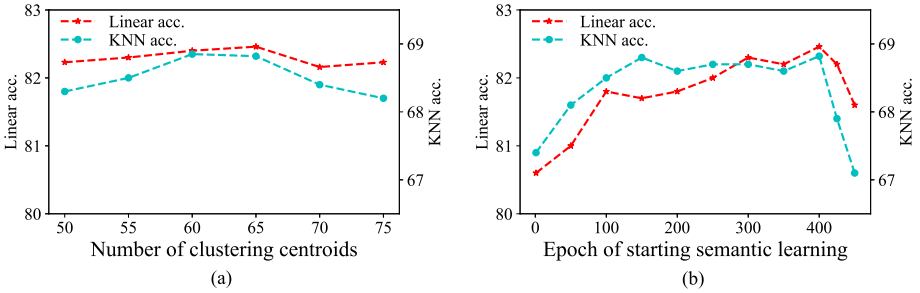


Fig. 4 Ablation experiments of hyperparameters under the x-sub protocol on NTU-60. **a** As the number of cluster centroids changes, the accuracy of linear evaluation and KNN evaluation changes; **b** As the epoch of starting semantic learning changes, the accuracy of linear evaluation and KNN evaluation changes

SCoBo, thereby ensuring that the encoder's ability to extract category-related abstract features does not decrease.

4.4.4 Effect of the epoch at which clustering begins

The performance of iterative algorithms is often significantly influenced by initialization (Dempster et al., 1977). In the training of ProSL, we do not introduce *SCoBo* until the training process gradually stabilized. We undertake a series of ablation experiments to pinpoint the optimal epoch for commencing semantic information acquisition, spanning a training duration of 450 epochs, and the results are illustrated in Fig. 4b. Notably, the results suggest that initiating semantic learning during the later stages of training is more advantageous. Commencing semantic learning prematurely, when the encoder lacks the foundational capacity for effective feature extraction, results in undue noise injection into *SCoBo*. Conversely, if the initiation of semantic learning is deferred excessively, the encoder will not be able to undergo effective adjustments.

4.5 Visualization analysis

In order to more intuitively observe the advantages of ProSL over the baseline, we perform a t-SNE (Van der Maaten and Hinton, 2008) visualization analysis of the learned skeleton features. As shown in Fig. 5, both methods are trained for 450 epochs, and our approach demonstrates a superior clustering of features of the same class. A horizontal observation reveals that in unsupervised skeleton sequence representation learning, contrastive learning enables the model to quickly converge and learn some semantic information during the initial training stages. However, due to its essence as an instance-level feature learning strategy, the model cannot further optimize the feature distribution as the number of training epochs increases. Before introducing progressive semantic learning, ProSL and HiCo-LSTM learn similar feature distributions, suggesting that the added branch not subject to semantic control has a minimal impact on the model. After incorporating progressive semantic learning, only minimal additional training is required to further enhance the model's feature extraction capability.

We also visualize the semantic coefficients from the adaptive data augmentation to analyze which joints have a greater impact on action recognition, as shown in Fig. 6. It can

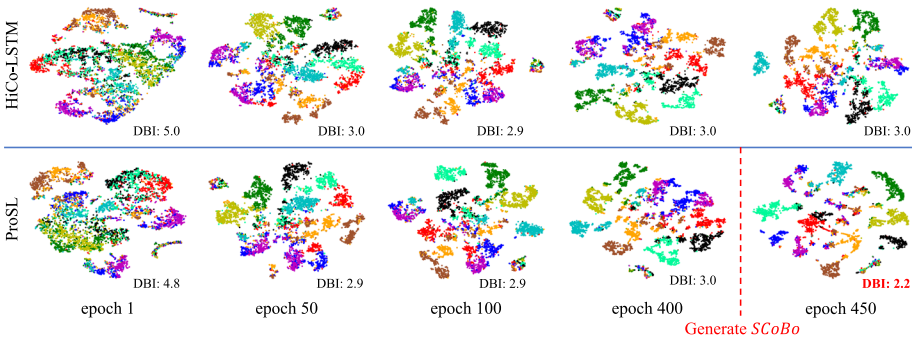
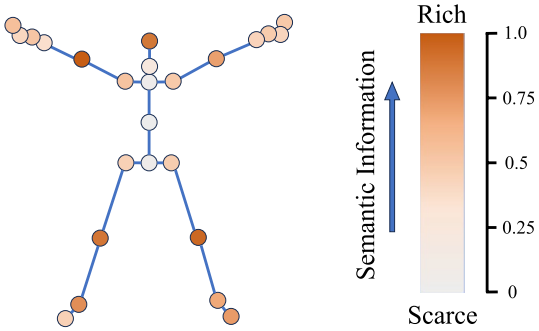


Fig. 5 t-SNE visualization of the features obtained by ProSL and HiCo-LSTM on NTU-60. We present the visualization results of both models at training epochs 1, 50, 100, 400, and 450, sequentially. 'Generate *SCoBo*' indicates the initiation of progressive semantic learning at the *400*-th epoch. DBI denotes Davies-Bouldin index, which is negatively correlated with the quality of clustering. It can be observed that our semantic learning phase effectively enhances the rationality of the feature distribution

Fig. 6 Visualization of semantic coefficients p_v. Deeper colors of the joints indicate richer semantic information



be observed that the elbow and knee joints contain richer semantic information, while the torso joint hardly reflects any semantic information.

4.6 Discussion and prospect

4.6.1 Training time

Taking the experiment under the x-sub protocol on NTU-60 as an example, compared to HiCo (~8 h, on the RTX 3090 GPU), our method adds 51min of time overhead, which accounts for less than 10% of the total training time, and this is acceptable. Among this, extracting $\hat{z}$ takes 21min, clustering takes 18min (utilizing *sklearn.cluster*), and calculating p_v takes 12min. Other operations related to clustering, such as outlier removal and clustering loss calculation, consume very little time (less than 1min).

4.6.2 Limitation

When the semantic information in the sequence is clear and fixed, our method can usually get significant gains. However, since a semantic centroid in *SCoBo* corresponds to a single and fixed semantic, our method may fail to learn effective semantic information in

scenarios where a skeleton sequence contains multiple actions or the types of semantics frequently change.

4.6.3 Outlook

We hope to avoid the aforementioned limitation by conducting a detailed semantic analysis and designing the extraction process with greater precision. Specifically, we believe that research on the interpretability of semantic information is crucial. Moreover, using large language models to achieve more refined control over the semantic extraction process may be effective.

5 Conclusion

In this paper, we propose a progressive semantic learning framework that efficiently accomplishes the self-supervised skeleton sequence representation task by iteratively learning cluster-level semantic information within the skeleton sequences. We introduce a Semantic Codebook to store the semantic information and guide the computation of the loss function, enhancing the rationality of the features extracted by the encoder. Additionally, we regulate the input using semantic coefficients to facilitate the encoder in capturing more semantic information. Extensive experiments and visualization results validate the effectiveness of our approach.

6 Supplementary information

The source code is provided in the Supplementary Materials.

Appendix A Rationality of the training process

Iterative algorithms have proven their effectiveness in many fields (Dempster et al., 1977; MacQueen et al., 1967; Wu et al., 2008). Here, we analyze ProSL based on the Expectation-Maximization algorithm. Suppose in the t-th iteration, the parameters of the network are denoted as θ^t, and the parameters of *SCoBo* are denoted as φ^t. For n samples, the objective of self-supervised skeleton sequence representation learning is to maximize the log-likelihood function:

$$\theta^{t+1} = argmax \sum_{i=1}^{n} logp(x_i;\theta^t) \tag{13}$$

After introducing *ScoBo*, we maximize the log-likelihood function of the model distribution as follows:

$$\theta^{t+1} = argmax \sum_{i=1}^{n} log \sum_{\varphi_i^t} p(x_i, \varphi_i^t;\theta^t) \tag{14}$$

Equation 14 is derived based on the marginal probability of x_i, and we cannot directly calculate θ^{t+1}. To address this, introduce an unknown new distribution $Q_i(\varphi_i^t)$ and scale Eq. 14 using Jensen's inequality as follows:

$$\sum_{i=1}^{n} \log \sum_{\varphi_i^t} p\left(x_i, \varphi_i^t; \theta^t\right) \geq \sum_{i=1}^{n} \sum_{\varphi_i^t} Q_i\left(\varphi_i^t\right) \log \frac{p\left(x_i, \varphi_i^t; \theta^t\right)}{Q_i\left(\varphi_i^t\right)} \tag{15}$$

If we want to satisfy the equality of Jensen's inequality, then:

$$\frac{p(x_i, \varphi_i^t; \theta^t)}{Q(\varphi_i^t)} = c = \sum_{\varphi_i^t} p(x_i, \varphi_i^t; \theta^t) \tag{16}$$

$$Q_i(\varphi_i^t) = \frac{p(x_i, \varphi_i^t; \theta^t)}{\sum_{\varphi_i^t} p(x_i, \varphi_i^t; \theta^t)} = \frac{p(x_i, \varphi_i^t; \theta^t)}{\sum_{\varphi_i^t} p(x_i; \theta^t)} = p(\varphi_i^t | x_i; \theta^t) \tag{17}$$

Equation 15 provides a lower bound for the log-likelihood function. Therefore, in the ***Maximization*** step, i.e., the representation learning process, our objective is:

$$\theta^{t+1} = argmax \sum_{i=1}^{n} \sum_{\varphi_i} Q_i\left(\varphi_i^t\right) \log \frac{p\left(x_i, \varphi_i^t; \theta^t\right)}{Q_i\left(\varphi_i^t\right)} \tag{18}$$

For the ***Expectation*** step, which is the action semantic learning process in ProSL, we need to estimate $p(\varphi_i^{t+1}; x_i, \theta^{t+1})$ to obtain $SCoBo^{t+1}$.

The iterative process has been proven to converge (Dempster et al., 1977), however, the final iteration results are highly influenced by the initialization, and model parameters may get stuck in local optimal points. In ProSL, to ensure the quality of φ_i^0, we choose to first perform several rounds of representation learning before starting the iterative training. Figure 4b in the main text illustrates the necessity of this strategy.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10994-024-06667-z.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant 42394060 and 42394064.

Author Contributions Hao Qin and Luyuan Chen conceived the ideas in this paper, conducted the experiments, and wrote the manuscript. Ming Kong and Zhuoran Zhao assisted with the theoretical analysis. Xianzhou Zeng and Mengxu Lu revised the manuscript. Qiang Zhu was responsible for the overall direction and planning. All authors discussed the results and contributed to the final manuscript.

Funding This work was supported by the National Natural Science Foundation of China under Grant 42394060 and 42394064.

Data availability The data used in this work are all public.

Code availability The source code is provided in the Supplementary Materials and it will be released after publishing.

Declarations

Conflict of interest None.

Ethics approval Not applicable.

Consent for publication Not applicable.

References

Bardes, A., Ponce, J., LeCun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906 (2021)

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660 (2021)

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems, 33*, 9912–9924.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607 (2020). PMLR

Chen, Y., Zhao, L., Yuan, J., Tian, Y., Xia, Z., Geng, S., Han, L., Metaxas, D.N.: Hierarchically self-supervised transformer for human skeleton representation learning. In: European Conference on Computer Vision, pp. 185–202 (2022). Springer

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological), 39*(1), 1–22.

Dong, J., Sun, S., Liu, Z., Chen, S., Liu, B., Wang, X.: Hierarchical contrast for unsupervised skeleton-based action representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 525–533 (2023)

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. *Advances in neural information processing systems, 33*, 21271–21284.

Guo, T., Liu, H., Chen, Z., Liu, M., Wang, T., Ding, R.: Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 762–770 (2022)

Guo, Y., Xu, M., Li, J., Ni, B., Zhu, X., Sun, Z., Xu, Y.: Hcsc: Hierarchical contrastive selective coding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9706–9715 (2022)

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)

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

Hershey, J.R., Olsen, P.A.: Approximating the kullback leibler divergence between gaussian mixture models. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, vol. 4, p. 317 (2007). IEEE

Hua, Y., Wu, W., Zheng, C., Lu, A., Liu, M., Chen, C., Wu, S.: Part aware contrastive learning for self-supervised action recognition. arXiv preprint arXiv:2305.00666 (2023)

Kim, B., Chang, H.J., Kim, J., Choi, J.Y.: Global-local motion transformer for unsupervised skeleton-based action learning. In: European Conference on Computer Vision, pp. 209–225 (2022). Springer

Li, L., Wang, M., Ni, B., Wang, H., Yang, J., Zhang, W.: 3d human action representation learning via cross-view consistency pursuit. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4741–4750 (2021)

Li, J., Zhou, P., Xiong, C., Hoi, S.C.: Prototypical contrastive learning of unsupervised representations. arXiv preprint arXiv:2005.04966 (2020)

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

Lin, L., Zhang, J., Liu, J.: Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2363–2372 (2023)

Liu, J., Song, S., Liu, C., Li, Y., Hu, Y.: 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 (2020)

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., & 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.

Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)

MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol.1, pp. 281–297 (1967). Oakland, CA, USA

Mao, Y., Zhou, W., Lu, Z., Deng, J., Li, H.: Cmd: Self-supervised 3d action representation learning with cross-modal mutual distillation. In: European Conference on Computer Vision, pp. 734–752 (2022). Springer

Noroozi, M., Vinjimoor, A., Favaro, P., Pirsiavash, H.: Boosting self-supervised learning via knowledge transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9359–9367 (2018)

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

Peng, W., Hong, X., Chen, H., Zhao, G.: Learning graph convolutional network for skeleton-based human action recognition by neural searching. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 2669–2676 (2020)

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

Shah, A., Roy, A., Shah, K., Mishra, S., Jacobs, D., Cherian, A., Chellappa, R.: Halp: Hallucinating latent positives for skeleton-based self-supervised learning of actions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18846–18856 (2023)

Shahroudy, A., Liu, J., Ng, T.-T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019 (2016)

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

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

Thoker, F.M., Doughty, H., Snoek, C.G.: Skeleton-contrastive 3d action representation learning. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1655–1663 (2021)

Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning, pp. 9929–9939 (2020). PMLR

Wang, F., Liu, H.: Understanding the behaviour of contrastive loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2495–2504 (2021)

Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742 (2018)

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems, 14*, 1–37.

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

Zhang, H., Hou, Y., Zhang, W., Li, W.: Contrastive positive mining for unsupervised 3d action representation learning. In: European Conference on Computer Vision, pp. 36–51 (2022). Springer

Zhang, J., Lin, L., Liu, J.: Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 3427–3435 (2023)

Zheng, N., Wen, J., Liu, R., Long, L., Dai, J., Gong, Z.: Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

Zhou, Y., Duan, H., Rao, A., Su, B., Wang, J.: Self-supervised action representation learning from partial spatio-temporal skeleton sequences. arXiv preprint arXiv:2302.09018 (2023)

Authors and Affiliations

Hao Qin[1] · Luyuan Chen[2] · Ming Kong[1] · Zhuoran Zhao[1] · Xianzhou Zeng[1] · Mengxu Lu[1] · Qiang Zhu[1]

✉ Qiang Zhu
zhuq@zju.edu.cn

Hao Qin
haoqin@zju.edu.cn

Luyuan Chen
chenly@bistu.edu.cn

Ming Kong
zjukongming@zju.edu.cn

Zhuoran Zhao
zhuoranzhao@zju.edu.cn

Xianzhou Zeng
xzhouzeng@zju.edu.cn

Mengxu Lu
lumengxu@zju.edu.cn

[1] School of Computer Science and Technology, Zhejiang University, HangZhou 310013, China

[2] Computer School, Beijing Information Science and Technology University, Beijing 100101, China