Contrastive Positive Mining for Unsupervised 3D Action Representation Learning

Haoyuan Zhang (Corresponding author)[1], Yonghong Hou[1], Wenjing Zhang[1], and Wanqing Li[2]

[1] Tianjin University, School of Electrical and Information Engineering, Tianjin, China
{zhy0860,houroy,zwj759}@tju.edu.cn
[2] Advanced Multimedia Research Lab, University of Wollongong, Wollongong, Australia.
wanqing@uow.edu.au

Abstract. Recent contrastive based 3D action representation learning has made great progress. However, the strict positive/negative constraint is yet to be relaxed and the use of non-self positive is yet to be explored. In this paper, a Contrastive Positive Mining (CPM) framework is proposed for unsupervised skeleton 3D action representation learning. The CPM identifies non-self positives in a contextual queue to boost learning. Specifically, the siamese encoders are adopted and trained to match the similarity distributions of the augmented instances in reference to all instances in the contextual queue. By identifying the non-self positive instances in the queue, a positive-enhanced learning strategy is proposed to leverage the knowledge of mined positives to boost the robustness of the learned latent space against intra-class and inter-class diversity. Experimental results have shown that the proposed CPM is effective and outperforms the existing state-of-the-art unsupervised methods on the challenging NTU and PKU-MMD datasets.

Keywords: Unsupervised learning, 3D action representation, Skeleton, Positive mining.

1 Introduction

Human action recognition is an active research in recent years. Due to being lightweight, privacy-preserving and robust against complex conditions [26,27,2,28], 3D skeleton is becoming a popular modality for capturing human action dynamics [10,39,31,43]. Majority of previous skeleton-based action recognition approaches [18,35,38,42] are developed with a fully-supervised manner. However, in order to learn a good action representation, supervised methods require a huge number of labeled skeleton samples which is expensive and difficult to obtain. It impels the exploration of learning skeleton-based action representation in an unsupervised manner [15,24,30,14]. Often unsupervised methods use pretext tasks to generate the supervision signals, such as reconstruction [7,44], autoregression [12,30] and jigsaw puzzles [22,36]. Consequently, the learning highly

relies on the quality of the designed pretext tasks, and those tasks are hard to be generalized for different downstream tasks. Recent unsupervised methods employ advanced contrastive learning [15,24,14] for instance discrimination in a latent space and have achieved promising results.

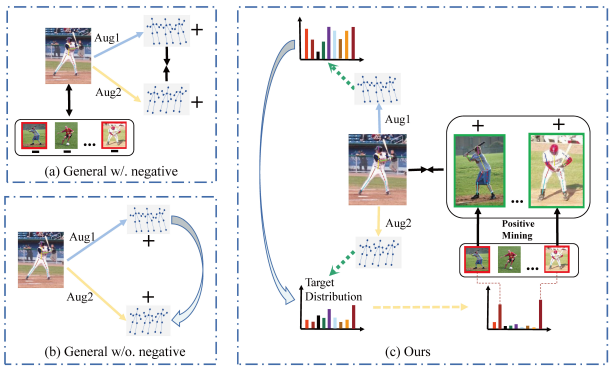

Fig. 1. Illustrations about the proposed CPM and previous contrastive methods. (a) contrastive learning methods with negative [15,24]. (b) contrastive learning methods without negative [4,6]. (c) the proposed Contrastive Positive Mining (CPM) method.

Although contrastive methods can improve the learning of skeleton representation, there are several issues, as illustrated in Fig. 1, in the current methods. Fig. 1(a) shows that the conventional contrastive learning methods require negatives [15,24]. They only regard different augmentations of the same instance as positives to be drawn close during the learning, while other instances in the queue, usually formed by training samples in the previous round of epochs or batches, are all regarded as negatives and pushed apart from the current instance. Although these methods consider the correlation of current instance with others, there are inevitably instances in queue that belong to the same category as the current instance (marked with red rectangular box) and these instances are mistaken as negatives, which could degrade the learned representation. To address this issue, as shown in Fig. 1(c), this paper proposes to search for the instances in queue that are likely to be the same class of the current instance, then to consider those instances as non-self positives (marked with green rectangular box) and draw them close to current instance so as to improve the learning.

Fig. 1(b) shows the conventional contrastive learning methods without negatives [4,6]. The positive setting is similar to the previous methods illustrated in Fig. 1(a). Only different augmentations of individual instance are used as positive, consistency among current instance and the non-self instances with the same class are ignored during learning, limiting the representation ability for intra-class diversity. Besides, although non-negative manner avoids the in-

stances of same class being pushed apart, the correlation of different instances are not considered.

Notice that contrastive objective of both methods (i.e. with or without negatives) is on individual instances, which challenges learning a feature space for all instances. To overcome the above shortcomings, as illustrated in Fig. 1(c), the proposed method extends the contrastive objective from individual instances by keeping a queue of instances and mining the non-self positives in the queue to boost learning. Specifically, a novel Contrastive Positive Mining (CPM) framework is proposed for unsupervised skeleton 3D action recognition. The proposed CPM is a siamese structure with a student and a target branch, which follows the SimSiam [4]. The student network is trained to match the target network in terms of the similarity distribution of the augmented instance in reference to all instances in a contextual queue, so that the non-self positive instances with high similarity can be identified in the queue. Then a positive-enhanced learning strategy is proposed to leverage the mined non-self positives to guide the learning of the student network. This strategy boosts the robustness of the learned latent space against intra-class and inter-class diversity. Experimental results on NTU-60 [25], NTU-120 [17] and PKU-MMD [16] datasets have validated the effectiveness of the proposed strategy.

To summarize, the key contributions include:

- A novel Contrastive Positive Mining (CPM) framework for unsupervised learning of skeleton representation for 3D action recognition.
- A simple but effective non-self positive mining scheme to identify the positives in a contextual queue.
- A novel positive-enhanced leaning strategy to guide the learning of the student network via the target network.
- Extensive evaluation of the CPM on the widely used datasets, NTU and PKU-MMD, with state-of-the-art results obtained.

2 Related Works

2.1 Unsupervised Contrastive Learning

Contrastive learning is derived from noise-contrastive estimation [8], which contrasts different type of noises to estimate the latent distribution. It has been extended in different ways for unsupervised learning. Contrastive Prediction Coding (CPC) [23] develops the info-NCE to learn image representation, with an auto-regressive model used to predict future in latent space. Contrastive Multiview Coding (CMC) [34] leverages multi-view as positive samples, so that the information shared between multiple views can be captured by the learned representation. However, there often lacks of negative instances for the above methods. To solve this issue, a scheme called memory-bank [37] is developed in which the previous random representations are stored as negative instances, and each of them are regarded as an independent class. Recently, MoCo [9] utilizes a dynamic dictionary to improve the memory-bank, and introduces the momentum

updated encoder to boost the representation learning. Another way to enrich the negative instances is to use large batch-size such as in SimCLR [3]. Particularly, SimCLR samples negatives from a large batch and shows that different augmentation, large batch size, and nonlinear projection head are all important for effective contrastive learning. However, these methods all regard different augmentations of the same instance as the only positives, while other instances in the queue including the ones with same category are all considered as negatives which cannot fully leverage capability of contrastive learning due to highly likely mixture of positives in the negatives.

To deal with this issue, some negative-sample-free approaches are recently developed. SimSiam [4] shows that simple siamese twin networks with a stop-gradient operation to prevent collapsing can learn a meaningful representation. Barlow Twins [41] proposes an unsupervised objective function by measuring the cross-correlation matrix between the outputs of two identical networks. BYOL [6] learns a potentially enhanced representation from an online network by predicting the representation from a given representation learned from a target network with slow updating. However, these methods do not consider consistency learning among current instances and the non-self instances with the same class.

2.2 Unsupervised 3D Action Recognition

Unsupervised methods [29,20,13] for video based action recognition are well developed, while few works are specifically for skeletons. LongT GAN (Generative adversarial network) [44] is an auto-encoder-based GAN for skeleton sequence reconstruction. P&C [30] employs an encoder-decoder learning structure, the encoder is weakened compared with decoder to learn more representative features. ASCAL [24] is a momentum LSTM with a dynamic updated memory-bank, augmented instances of the input skeleton sequence are contrasted to learn representation. MS^{2}L [15] is a multi-task learning framework, with both pretext tasks and contrastive learning. CrosSCLR [14] adopts a cross-view contrastive learning scheme and leverages multi-view complementary supervision signal. However, these methods either require pretext tasks or a large amount of negative samples, or rely on the reconstruction.

3 Proposed Method

3.1 Overview

Fig. 2 shows the basic framework of CPM. CPM adopts siamese twin networks as inspired by SimSiam [4]. 3D skeleton sequences are randomly augmented. Assume that a skeleton sequence s has T frames, V joints, and C coordinate channels, which can be represented as $s \in \mathrm{R}^{C \times T \times V}$. To augment s into different versions x and x', a skeleton-specific augmentation strategy is needed. Different from the augmentations implemented for images, augmentation of skeleton sequences needs to be effective for learning spatial-temporal dynamics. In this

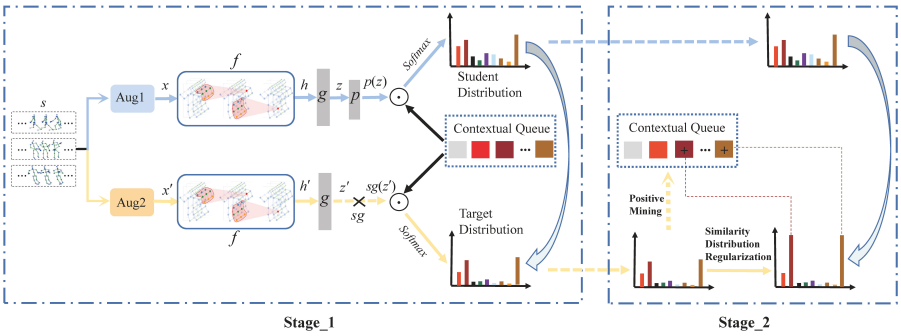

Fig. 2. Overview of the CPM framework. CPM includes two stages. In the first stage, the student branch is trained to predict the inter-skeleton similarity distribution inferred by the target network, so as to excavate non-self positive. Then in the second stage the information of mined positives is injected into the target branch through similarity distribution regularization to guide the learning of student, which achieves positive-enhanced learning (different colors in distribution and contextual queue represent the embeddings of different instances, '+' means mined positive).

paper, shear and crop in the spatial and temporal domain are to augment samples. Specifically, shear is applied as a spatial augmentation and is implemented as a linear transformation that displaces the skeleton joint in a fixed direction. Skeleton sequences are multiplied by a transformation matrix on the channel dimension, so as to slant the shape of 3D coordinates from body joints at a random angle. Crop is to pad a number of frames to a sequence symmetrically, then the sequence is randomly cropped into a fixed length [24,14].

The siamese encoders with identical network structure are to encode the augmented skeleton sequences, as shown in Fig. 2, in a latent feature space. One branch is referred to as the student and the other serves as the target [32]. ST-GCN [39] is adopted as the encoder networks. The siamese encoders consist of several GCN layers and embed the two augmented skeleton sequences x and x' into a latent space. In each layer, human pose in spatial-dimension and joint's motion in temporal-dimension are alternatively encoded, i.e. a spatial graph convolution is followed by a temporal convolution.

After the siamese encoders, a projection MLP g is attached to project the vector h and h' in the encoding space: $z = g(h)$, $z' = g(h')$, where z and z' are assumed to be mean-centered along the batch dimension so that each unit has 0 mean output over the batch. The projection MLP consists of two layers, the first one is followed by a batch normalization layer and rectified linear units. After the projection MLP, a prediction MLP p with same architecture of g is attached to the student branch to produce the prediction $p(z)$, while the stop-gradient operation is used in the target branch with the output $sg(z')$. In addition, a "first-in first-out" [9] contextual queue $Q = [a_1, ..., a_N]$ is used to measure how

well the encoded augmented instance by student network matches that by the target network with respect to the instances in the queue.

The key idea of the proposed method is to use the output of the student network to predict the output of the target network. More specifically, our objective is to train the siamese encoders such that the student network matches the target network in terms of the similarity distribution of the augmented instance in reference to all instances in the queue.

3.2 Similarity Distribution and Positive Mining

Similarity between the encoded feature of the augmentations and the instances in queue is first calculated and similarity distribution for the student network and the target network are calculated through softmax. The learning process is to train the network so that the similarity distribution of x with respect to the instances in the queue can predict the distribution of x'. Compared with the previous methods, this strategy has the following advantages. No strict definition of positives/negatives is required and the match on similarity distribution over the instances in the queue is more reliable than that over individual instances. Since the similarity between the augmented instances and instances of the same class in the queue is expected to be high, resulting in an implicit mining of non-self positive instances in the queue.

Let $Q = [a_1, ..., a_N]$ be the queue of N instances, where a_i is the embedding of the i-th instance. The contextual queue comes from the preceding several batches of target network, which is updated in "first-in first-out" [9] strategy. Similarity distributions, d_i and d'_i, between $\bar{p}(z)$ and a_i and between $\bar{z}'$ and a_i are computed as follows, respectively,

$$d_i = \frac{e^{\bar{p}(z)\cdot a_i/\tau}}{\sum_{j=1}^{N} e^{\bar{p}(z)\cdot a_j/\tau}} \tag{1}$$

$$d'_i = \frac{e^{\bar{z}'\cdot a_i/\tau'}}{\sum_{j=1}^{N} e^{\bar{z}'\cdot a_j/\tau'}} \tag{2}$$

where $\bar{p}(z)$ and $\bar{z}'$ are l_2 normalization of $p(z)$ and z'. The overall similarity distributions, D and D' of the two arguments in the latent space with respect to the instances in the queue are,

$$D = \{d_i\}, D' = \{d'_i\}, i \in N \tag{3}$$

The idea is to training siamese encoders to match D with D'. In this paper, we adopt to minimize the Kullback-Leibler divergence between D and D', i.e.,

$$L = D_{KL}(D'||D) = H(D', D) - H(D') \tag{4}$$

By minimizing L, prediction $p(z)$ can be aligned with z'. Meanwhile, instances that belong to the same class could be pushed close in the latent space, while those from different classes are pushed apart.

The similarity measures provide information for mining the positive instance in the queue. Specifically, given one instance's embedding z and the corresponding queue Q, instances in queue with top-k high similarity are considered as positives, i.e.,

$$\Gamma\left(Q\right) = \mathrm{Topk}\left(Q\right) \tag{5}$$

which generates the index set of positive instances,

$$D'_{+} = \left\{d'_i\right\}, i \in N_{+} \tag{6}$$

where N_{+} is the index set of non-self positive instances. These positives can be used to facilitate a positive-enhanced learning as described below.

3.3 Positive-enhanced Learning

The non-self positives can be used to boost the representation learning. Intuitively, it is reasonable to inject the information of mined positives into the target branch to guide the learning of the student encoder. To do this, it is proposed to regularize the similarity distribution of the target branch D' in each batch, so as to make use of the non-self positives iteratively. Specifically, we set the similarities of the K mined positive instances in target branch to 1, which means those instances are considered the same action category with current instance. This strategy is referred to as "positive-enhanced leaning". The positive-enhanced similarity distribution can be expressed as,

$$d_i^e = \begin{cases} \frac{e^{1/\tau'}}{\sum_{j=1}^{N} e^{\tilde{z}'\cdot a_j/\tau'}}, i \in N_{+} \\ d'_i, \text{otherwise} \end{cases} \tag{7}$$

Then we train distribution of student to continue predicting the regularized target distribution, so that the student is guided to learn more informative intra-class diversity brought by the non-self skeleton positives knowledge we inject,

$$L' = H(D'_{NP}, D) - H(D'_{NP}) \tag{8}$$

where $D'_{NP} = \{d_i^e\}$ is the non-self positive-enhanced target distribution. Compared to Eq.(4), Eq.(8) intends to pull positive instances closer.

3.4 Learning of CPM

In the early training stage, the model is likely not stable and capable enough of providing reasonable measures of the similarity distribution to identify the

positives in the queue. Therefore, a two-stage training strategy is adopted for CPM: the student branch is first trained to predict the similarity distribution inferred by the target network without enhanced positives in Eq.(4). When it is stable, the model is trained using the positive-enhanced learning strategy in Eq.(8).

4 Experiments

4.1 Datasets

NTU RGB+D 60 (NTU-60) Dataset [25]: NTU-60 is one of the widely used indoor-captured datasets for human action recognition. 56880 action clips in total are performed by 40 different actors in 60 action categories. The clips are captured by three cameras simultaneously at different horizontal angles and heights in a lab environment. Experiments are conducted on the Cross-Subject (X-Sub) and Cross-View (X-View) benchmarks.

NTU RGB+D 120 (NTU-120) Dataset [17]: NTU-120 is an extended version of NTU-60. There are totally 114480 action clips in 120 action categories. Most settings of NTU-120 follow the NTU-60. Experiments are conducted on the Cross-Subject (X-Sub) and Cross-Setup (X-Set) benchmarks.

PKU-MMD Dataset [16]: There are nearly 20,000 action clips in 51 action categories. Two subsets PKU-MMD I and PKU-MMD II are used in the experiments. PKU-MMD II is more challenging than PKU-MMD I as it has higher level of noise. Experiments are conducted on the Cross-Subject (X-Sub) benchmark for both subsets.

4.2 Implementation

Architecture: The 9-layer ST-GCN [39] network is chosen as the encoders. In each layer, the spatial graph convolution is followed by a temporal convolution, the temporal convolutional kernel size is 9. A projector of 2-layer MLP is attached to the output of both networks. The first layer is followed by a batch normalization layer and rectified linear units, with output size of 512, while the output dimension of the second layer is 128. A predictor with the same architecture is used in the student branch, while the stop-gradient operation is applied in target branch. The contextual queue size N is set to 65536, 32768 and 16384 for NTU-60/120, PKU-MMD I and PKU-MMD II datasets, respectively.

Unsupervised Pre-training: LARS [40] is utilized as optimizer and trained for 400 epochs with batch size 512, note that the positive-enhanced learning is conducted after 300 epochs. The learning rate starts at 0 and is linearly increased to 0.5 in the first 10 epochs of training and then decreased to 0.0005 by a cosine decay schedule [19]. All experiments are conducted on one Nvidia RTX3090 GPU using PyTorch.

Linear Evaluation Protocol: The pre-trained models are verified by linear evaluation. Specifically, a linear classifier (a fully-connected layer followed by a softmax layer) is trained supervisedly for 100 epochs while the pre-trained model is fixed.

4.3 Results and Comparison

Unsupervised Results: The performance of the proposed CPM is compared with the state-of-the-art supervised and unsupervised methods on the NTU and PKU-MMD datasets and results are shown in Table 1. Following the standard practice in literature the recognition performance in terms of top-1 classification accuracy is reported. Note that, if not specified, the experiments including ablation study are conducted on the joint data. 3S means the ensemble results of joint, bone and motion data. The obvious performance improvement compared with the recent advanced unsupervised counterparts [14,33] has been obtained and demonstrates the effectiveness of CPM. In addition, CPM (3S) outperforms the supervised ST-GCN [39] on both NTU and PKU-MMD datasets.

Table 1. Performance and comparison with the state-of-the-art methods on the NTU and PKU-MMD datasets.

Architectures	NTU-60 (%)		NTU-120 (%)		PKU-MMD (%)	
	X-Sub	X-View	X-Sub	X-Set	Part I	Part II
Supervised						
C-CNN + MTLN [11]	79.6	84.8	-	-	-	-
TSRJI [1]	73.3	80.3	67.9	62.8	-	-
ST-GCN [39]	81.5	88.3	70.7	73.2	84.1	48.2
Unsupervised						
LongT GAN [44]	39.1	48.1	-	-	67.7	27.0
ASCAL [24]	58.5	64.8	48.6	49.2	-	-
MS^2L [15]	52.6	-	-	-	64.9	27.6
P&C [30]	50.7	76.3	-	-	-	-
ISC [33]	76.3	85.2	67.9	67.1	80.9	36.0
CrosSCLR (joint) [14]	72.9	79.9	-	-	-	-
CrosSCLR (3S) [14]	77.8	83.4	67.9	66.7	84.9	-
CPM (joint)	78.7	84.9	68.7	69.6	88.8	48.3
CPM (3S)	**83.2**	**87.0**	**73.0**	**74.0**	**90.7**	**51.5**

Semi-supervised Results: The CPM is first pre-trained on all training data in an unsupervised way, then the classifier is fine-tuned with 1% and 10% annotated data respectively. Table 2 shows the semi-supervised results on the NTU-60 dataset. The results have shown the proposed CPM performs significantly better than the compared methods. Compared with MS^2L [15] and ISC [33], CPM improves the performance by a large margin and shows its robustness when fewer labels are available for fine-tuning.

Fully Fine-tuned Results: The model is first unsupervisedly pre-trained, then a linear classifier is appended to the learnable encoder. Both the pre-trained model and the classifier undergo a supervised training using all training data [41], results are shown in Table 3. On both NTU-60 and NTU-120 datasets the fully

Table 2. Semi-supervised performance and comparison with the state-of-the-art methods on the NTU-60 dataset.

Architectures	Label fraction (%)	X-Sub (%)	X-View (%)
LongT GAN [44]	1	35.2	-
MS^{2}L [15]	1	33.1	-
ISC [33]	1	35.7	38.1
CPM	1	**56.7**	**57.5**
LongT GAN [44]	10	62.0	-
MS^{2}L [15]	10	65.2	-
ISC [33]	10	65.9	72.5
CPM	10	**73.0**	**77.1**

fine-tuned CPM outperforms the supervised ST-GCN [39], demonstrating the effectiveness of the unsupervised pretraining.

Table 3. Fully fine-tuned performance and comparison on the NTU-60 and NTU-120 datasets.

Architectures	NTU-60 (%)		NTU-120 (%)	
	X-Sub	X-View	X-Sub	X-Set
C-CNN + MTLN [11]	79.6	84.8	-	-
TSRJI [1]	73.3	80.3	67.9	62.8
ST-GCN [39]	81.5	88.3	70.7	73.2
CPM	**84.8**	**91.1**	**78.4**	**78.9**

4.4 Ablation Study

On positive mining: To verify the effectiveness of positive-enhanced learning, we pre-train the CPM (w/o. PM) without identifying the positive instances and, hence, positive-enhanced learning, other settings are kept the same. Performance of CPM and CPM (w/o. PM) is shown in Table 4. On the NTU-60 X-Sub and X-View tasks, CPM improves the recognition accuracy by 3.1 percentage points and 3.2 percentage points, respectively. On the NTU-120 X-Sub and X-Set tasks, 3.9 percentage points and 4.9 percentage points improvements are obtained by CPM. This demonstrates that identification of positive instances and the positive-enhanced learning strategy do improve the representation learning.

To further verify how well the non-self positives in the queue can be identified for the positive-enhanced learning, Fig. 3 shows the precision of the positives selected by CPM and CPM (w/o. PM) in one epoch in the top-100 identified positive instances. The results show that even CPM (w/o. PM) is capable of identifying many true positives. This is in significant contrast to the methods in [15,24] where all instances in the queue would be considered as negatives. When positive-enhanced learning is applied, the precision has been significantly

Table 4. Benefit of positive mining.

Datasets	CPM (w/o. PM) (%)	CPM (%)
X-Sub (NTU-60)	75.6	78.7
X-View (NTU-60)	81.7	84.9
X-Sub (NTU-120)	64.8	68.7
X-Set (NTU-120)	64.7	69.6

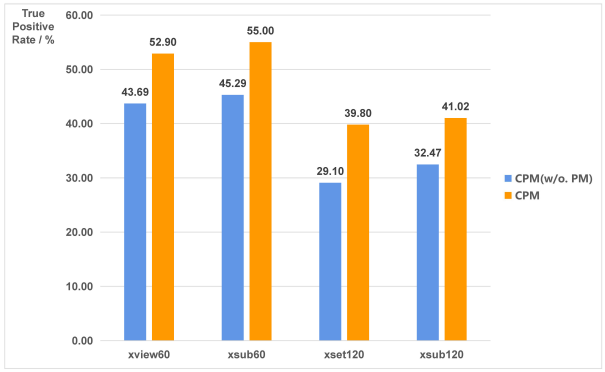

Fig. 3. Precision of positive instances identified by CPM and CPM (w/o. PM) on different datasets.

increased and so that the learned representation is more robust against the intra-class diversity.

On the value of K: Hyper-parameter K refers to the number of positives identified in the queue. This study shows how K affects the performance. Experiments have shown that when K is 100, best results are obtained on the NTU-60 and NTU-120 datasets. Results on NTU-60 X-View are shown in Fig. 4. It is found that too large or too small K both decreases the performance. A large value of K could include unexpected false positives with low similarity that misleads the learning. A small value of K might ignore too many true positives that would potentially decrease representation ability to accommodate intra-class diversity. Good performance was observed when K is 50 and 25 for PKU-MMD I and PKU-MMD II datasets, respectively. It is conjectured that choice of K may depend on the scale of the dataset.

On the value of τ': Fig. 5 shows the performance of CPM (w/o. PM) using different τ' with τ fixed to 0.1 [5], the optimal performance is obtained when τ' is 0.05. A large value of τ' could lead to a flatter target distribution so that the learned representation becomes less discriminative. A small value of τ' would suppress the difference in similarities between the positive and the negative, leading to many false positives included in the positive-enhanced learning. If τ' is too small, less positive instances could be identified and this would again adversely affect the effectiveness of learning.

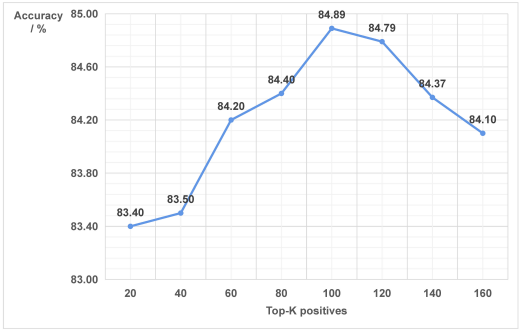

Fig. 4. Effect of the number K of top positives on the proposed CPM in the NTU-60 X-View task.

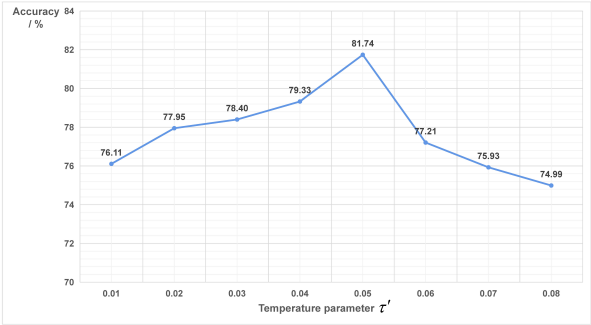

Fig. 5. Effect of different temperature τ' on the performance in the NTU-60 X-View task.

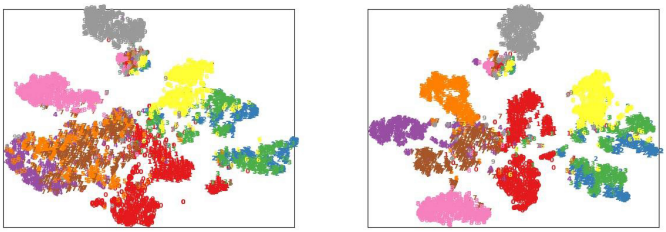

Fig. 6. t-SNE visualization of embedding for (a) CPM (w/o. PM) and (b) CPM on the NTU-60 X-View task (best view in color).

Embedding Visualization: t-SNE [21] is used to visualize the embedding clustering produced by CPM (w/o. PM) and CPM as shown in Fig. 6. Note that embedding of 10 different action categories are sampled and visualized with different colors. The visual results show how well the embedding of the same type of actions form clusters while different types of actions are separated. By comparing the t-SNE of CPM and CPM (w/o. PM), CPM has clearly improved the clustering of actions, which indicates that the learned latent space is more discriminative than the space learned without using positive-enhanced learning strategy.

5 Conclusion

In this paper, a novel unsupervised learning framework called Contrastive Positive Mining (CPM) is developed for learning 3D skeleton action representation. The proposed CPM follows the SimSiam [4] structure, consisting of siamese encoders, student and target. By constructing a contextual queue and identifying non-self positive instances in the queue, the student encoder is able to learn a discriminative latent space by matching the similarity distributions of individual instance's two augments with respect to the instances in the queue. In addition, by identifying positive instances in the queue, a positive-enhanced learning strategy is developed to boost the robustness of the learned latent space against intra-class and inter-class diversity. Experiments on the NTU and PKU-MMD datasets have shown that the proposed CPM obtains the state-of-the-art results.

References

1. Caetano, C., Brémond, F., Schwartz, W.R.: Skeleton image representation for 3d action recognition based on tree structure and reference joints. In: 2019 32nd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI). pp. 16–23. IEEE (2019)
2. Chen, J., Samuel, R.D.J., Poovendran, P.: Lstm with bio inspired algorithm for action recognition in sports videos. Image and Vision Computing **112**, 104214 (2021)
3. 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. PMLR (2020)
4. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021)
5. Fang, Z., Wang, J., Wang, L., Zhang, L., Yang, Y., Liu, Z.: Seed: Self-supervised distillation for visual representation. arXiv preprint arXiv:2101.04731 (2021)
6. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al.: Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733 (2020)
7. Gui, L.Y., Wang, Y.X., Liang, X., Moura, J.M.: Adversarial geometry-aware human motion prediction. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 786–803 (2018)
8. Gutmann, M.U., Hyvärinen, A.: Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. Journal of Machine Learning Research **13**(2) (2012)
9. 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)
10. Hou, Y., Li, Z., Wang, P., Li, W.: Skeleton optical spectra-based action recognition using convolutional neural networks. IEEE Transactions on Circuits and Systems for Video Technology **28**(3), 807–811 (2018)
11. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3d action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3288–3297 (2017)
12. Kundu, J.N., Gor, M., Uppala, P.K., Radhakrishnan, V.B.: Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In: 2019 IEEE winter conference on applications of computer vision (WACV). pp. 1459–1467. IEEE (2019)
13. Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.S.: Unsupervised learning of view-invariant action representations. arXiv preprint arXiv:1809.01844 (2018)
14. 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)
15. 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)
16. 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)

17. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: 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 (2019)
18. Liu, M., Liu, H., Chen, C.: 3d action recognition using multiscale energy-based global ternary image. IEEE Transactions on Circuits and Systems for Video Technology **28**(8), 1824–1838 (2017)
19. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
20. Luo, Z., Peng, B., Huang, D.A., Alahi, A., Fei-Fei, L.: Unsupervised learning of long-term motion dynamics for videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2203–2212 (2017)
21. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
22. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision. pp. 69–84. Springer (2016)
23. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
24. Rao, H., Xu, S., Hu, X., Cheng, J., Hu, B.: Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. Information Sciences **569**, 90–109 (2021)
25. 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)
26. Shi, Z., Kim, T.K.: Learning and refining of privileged information-based rnns for action recognition from depth sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3461–3470 (2017)
27. Si, C., Chen, W., Wang, W., Wang, L., Tan, T.: An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1227–1236 (2019)
28. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: Spatio-temporal attention-based lstm networks for 3d action recognition and detection. IEEE Transactions on image processing **27**(7), 3459–3471 (2018)
29. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using lstms. In: International conference on machine learning. pp. 843–852. PMLR (2015)
30. 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)
31. Sun, N., Leng, L., Liu, J., Han, G.: Multi-stream slowfast graph convolutional networks for skeleton-based action recognition. Image and Vision Computing **109**, 104141 (2021)
32. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780 (2017)
33. 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)

34. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 776–794. Springer (2020)
35. Wang, P., Li, W., Gao, Z., Zhang, Y., Tang, C., Ogunbona, P.: Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 595–604 (2017)
36. Wei, C., Xie, L., Ren, X., Xia, Y., Su, C., Liu, J., Tian, Q., Yuille, A.L.: Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1910–1919 (2019)
37. 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)
38. Xiao, Y., Chen, J., Wang, Y., Cao, Z., Zhou, J.T., Bai, X.: Action recognition for depth video using multi-view dynamic images. Information Sciences **480**, 287–304 (2019)
39. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
40. You, Y., Gitman, I., Ginsburg, B.: Large batch training of convolutional networks. arXiv preprint arXiv:1708.03888 (2017)
41. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. arXiv preprint arXiv:2103.03230 (2021)
42. Zhang, H., Hou, Y., Wang, P., Guo, Z., Li, W.: Sar-nas: Skeleton-based action recognition via neural architecture searching. Journal of Visual Communication and Image Representation **73**, 102942 (2020)
43. Zhang, X., Xu, C., Tao, D.: Context aware graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14333–14342 (2020)
44. 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)