

# Cross-Stream Contrastive Learning for Self-Supervised Skeleton-Based Action Recognition

Ding Li<sup>1,3</sup>, Yongqiang Tang<sup>3</sup>, Zhizhong Zhang<sup>2</sup>, Wensheng Zhang<sup>3</sup>

<sup>1</sup> School of Artificial Intelligence, UCAS    <sup>2</sup> East China Normal University

<sup>3</sup> State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, CAS

liding2019@ia.ac.cn, yongqiang.tang@ia.ac.cn, zzzhang@cs.ecnu.edu.cn

## Abstract

*Self-supervised skeleton-based action recognition enjoys a rapid growth along with the development of contrastive learning. The existing methods rely on imposing invariance to augmentations of 3D skeleton within a single data stream, which merely leverages the easy positive pairs and limits the ability to explore the complicated movement patterns. In this paper, we advocate that the defect of single-stream contrast and the lack of necessary feature transformation are responsible for easy positives, and therefore propose a Cross-Stream Contrastive Learning framework for skeleton-based action Representation learning (CSCLR). Specifically, the proposed CSCLR not only utilizes intra-stream contrast pairs, but introduces inter-stream contrast pairs as hard samples to formulate a better representation learning. Besides, to further exploit the potential of positive pairs and increase the robustness of self-supervised representation learning, we propose a Positive Feature Transformation (PFT) strategy which adopts feature-level manipulation to increase the variance of positive pairs. To validate the effectiveness of our method, we conduct extensive experiments on three benchmark datasets NTU-RGB+D 60, NTU-RGB+D 120 and PKU-MMD. Experimental results show that our proposed CSCLR exceeds the state-of-the-art methods on a diverse range of evaluation protocols.*

## 1. Introduction

Skeleton-based action recognition has always attracted considerable research interests in the field of computer vision, as it plays a significant role in many real-world applications, such as smart surveillance, human-machine interaction and mixed reality [12, 25, 35, 43]. It aims to recognize human actions using skeleton keypoints, and shows advantages under dynamic circumstance with complicated background (e.g. clutter scene, light-conditions) [40, 52]. In the past years, most existing skeleton-based action recognition

methods are based on supervised learning paradigm, which requires immense time and manual effort for annotating. Instead, self-supervised learning paradigm avoids such limitations, aiming to learn discriminative spatio-temporal action representations by exploring unlabeled skeleton data.

Several self-supervised approaches formulate the pretext task in the way of Generative Learning, including reconstruction of input skeleton sequence [56], solving jigsaw puzzles [23] and motion prediction [7]. However, the generative pretext tasks force the model to pay excessive attention to low-level detailed joint features, while ignoring the high-level semantic information which is more critical to downstream tasks. Different from the pretext task in generative methods, Contrastive Learning typically leverages the instance discrimination of skeleton sequences in the feature space. Different augmentations are applied to skeleton sequences, thus generating multiple views. Then, contrastive pairs are constructed based on these views, and the inherent consistency constraints is utilized to attract the positive pairs and repel negative pairs simultaneously [3, 15]. Based on the high-level semantic features, the contrastive learning enables the model own its advantage in downstream tasks.

Despite the success of contrastive learning in this area, previous methods rely on imposing invariance to augmentations of 3D skeleton, and these positive pairs still could be quite similar in the feature space, *i.e.*, *easy positives*. As shown in Figure 1(a), easy positives are distributed closely, which results in less and less contribution to the loss as training progresses. Using the easy positive pairs, the contrastive pretext task will be easily-accomplished, and thus limiting the ability to explore the complicated movement patterns [5, 14, 45]. On contrary, compared with easy positives, hard positives have longer distance in feature space and contribute more to the loss. Obviously, the hard positives, which could make contrastive pretext task tougher in turn, are expected to promote learning more robust representation.

In this study, we mainly focus on handling two important issues that lead to high similarity of positive pairs: *the*

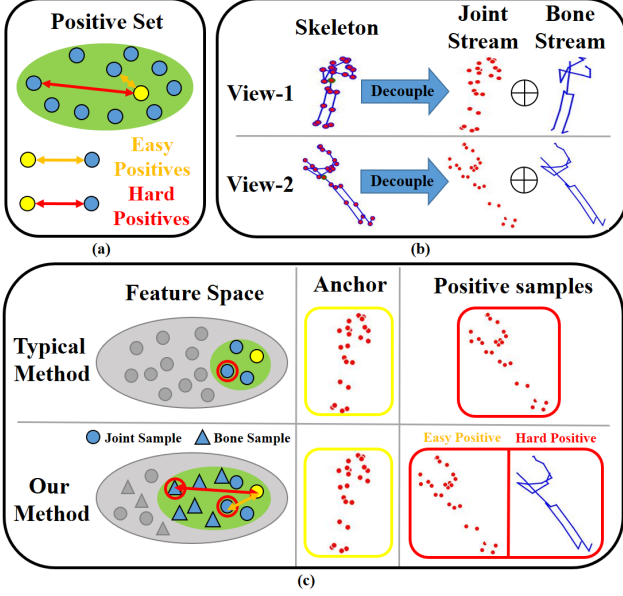


Figure 1. **An illustration of the proposed approach.** (a) Easy positive pairs distribute closely in feature space, while hard positive pairs are relatively far away from each other. (b) A skeleton sequence can be augmented to formulate two views, and then joint and bone stream are decoupled from the two views. (c) Comparing with typical method, CSCLR enforces not only intra-stream correspondence (easy positive), but correspondence between different streams (hard positive), *e.g.* joint stream and bone stream.

*defect of single-stream contrast and the lack of necessary feature transformation.* First, among existing works, only a single skeleton data stream is used for designing contrast pairs, while other streams are underutilized. Actually, based on the raw data, multiple streams (*e.g.* joint, bone and motion stream) can be obtained. There have been some studies dedicating to fuse the prediction results of these streams in the downstream tasks [14, 21, 27, 40, 53]. However, in contrastive pre-training stage, previous methods generally overlook the multi-stream scenario and only construct pairs within a single data stream. The pairs, which are derived from the same stream, share the similar inherent information and tend to be easy positives. Second, conventional methods typically utilize data augmentations to construct positive pairs [14, 21, 37, 44, 53], but such data-level operation of these pairs may not be able to explicitly guarantee their discrepancy in feature level. At the same time, multiple sample views can be designed not only based on input data space, but on feature space as well [10, 49, 58]. Compared with data augmentation, feature transformation offers an explicit solution to design more effective pairs for training. Thus, by involving the feature-level manipulation, the similarity of positive pairs are expected to be further reduced.

To remedy the issues mentioned above, we propose

**CSCLR**: a cross-stream contrastive learning framework with features extracted from multiple data stream. For the first issue, CSCLR exploits the correspondence between different skeleton data streams, so as to constructively learn transferable skeleton representations that benefit to downstream tasks. To be specific, as shown in Figure 1(b), we generate bone stream accompanied with joint stream. And then, different from typical methods that only conduct intra-stream contrastive learning, our CSCLR additionally constructs inter-stream contrast pairs, *i.e.*, the two embeddings in each pair come from different source data stream. Since the inherent information in other data streams are different from the joint stream, thus the inter-stream contrast pairs are expected to act as hard positives to formulate a better representation learning (see Figure 1(c)). For the second issue, we propose the Positive Feature Transformation (PFT) strategy, which aims to increase the variance of features in positive set through feature extrapolation. Inspired by Mixup [55] and Manifold Mixup [49], PFT applies linear extrapolation on the paired query and key features and generate synthetic positive features for contrastive learning. After PFT, the hardness of these positive samples are enhanced, leading to evident gains in representation learning.

The main contributions of this paper can be summarized as follows:

1. We propose a novel cross-stream contrastive learning model named CSCLR for self-supervised skeleton-based action recognition, which contrasts the pairwise features extracted from different data stream. Compared with existing methods only using intra-stream contrast, CSCLR is proposed to introduce more hard samples from other streams, thus resulting in stronger effect of contrastive learning and better generalization performance in downstream tasks.
2. To obtain more robust skeleton feature, we design a positive feature transformation strategy, which manipulates the positive pairs to increase their hardness in feature-level. A further enhancement of contrastive learning is achieved with the generated synthetic positive features, enabling the skeleton encoder to learn more effective representations.
3. We conduct extensive experiments and ablation study on three popular benchmark datasets, *i.e.*, NTU-RGBD-60, NTU-RGBD-120 and PKU-MMD. As a result, our CSCLR achieves the state-of-the-art performance under variety of evaluation protocols, including linear evaluation, semi-supervised evaluation, and finetuned evaluation.

The rest of this article is organized as follows. Section 2 introduces related works. In Section 3, we introduce our method with detailed description. Experimental analysis

and comparison results are shown in Section 4 to verify our method. Finally, we conclude this paper in Section 5.

## 2. Related Work

In order to indicate our proposed method, we will review related advances in three research areas in this section, including contrastive self-supervised representation learning, supervised skeleton-based action recognition and self-supervised skeleton-based action recognition.

### 2.1. Contrastive Self-Supervised Representation Learning

Contrastive self-supervised learning aims to learn feature representations from unlabeled data with contrastive loss. DIM [16] maximizes the mutual information between a region of input to the encoder and its output. MoCo [5, 15] builds a dynamic dictionary with a queue and a moving-averaged encoder for computing the contrastive loss, which enables building a large and consistent dictionary on-the-fly that facilitates contrastive unsupervised learning. SimCLR [3, 4] does not adopt memory bank and introduces a non-linear transformation between the representation and the loss function. PIRL [28] develops pretext-Invariant method that learns invariant representations based on pretext tasks, and substantially improves the semantic quality of the learned image representations. InfoCL [50] claims that non-shared task-relevant information cannot be ignored and proposes to increase the mutual information between the representation and input as regularization to approximately introduce more task-relevant information. Cross-Point [1] constructs a joint training objective which combines the feature correspondences within and across modalities, thus ensembles a rich learning signal from both 3D point cloud and 2D image modalities in a self-supervised fashion. [59] visualizes the similarity score distributions of pairwise samples in contrastive learning, and generates more effective contrast pairs for training. Contrastive learning with multiple inputs, *e.g.* image + audio [2, 29, 32], video + sentence [9, 36, 38], has also achieved favorable advances recently. The advances in this research area lay a solid foundation for our work, and show a promising direction for us to delve into hard contrast pairs based on different streams of 3D skeleton data.

### 2.2. Supervised Skeleton-Based Action Recognition

The input of skeleton-based action recognition is 2D/3D skeleton data, which is robust to illumination change, scene variation, and complex backgrounds [40, 52]. Traditional methods adopt hand-crafted features from raw skeleton sequences, such as [47, 48, 51]. Inspired by the success of deep learning, numerous methods based on deep neural network are carried out to extract discriminative skeleton features. Among these deep learning methods, the most widely-used

networks are recurrent neural networks (RNN), convolutional neural networks (CNN) and graph convolutional networks (GCN). RNN-based models [11, 22, 41] usually concatenate the coordinates of all joints in each frame as a vector and then take the sequence of vectors as input of RNN. Due to the gradient vanishing in RNN-based methods [17], researchers would pay attention to CNN-based models. CNN-based methods [19, 20] usually transfer the raw skeleton sequence into image form, and then adopt convolution operations on this image to capture effective representation.

Recently, GCN-based methods [6, 8, 27, 40, 52] are proposed to model the skeleton data as graph with nodes and edges, resulting in great success in this task. ST-GCN [52] proposes a generic graph-based formulation for modeling dynamic skeletons and introduces several principles in designing convolution kernels to meet the specific demands. 2s-AGCN [40] is proposed to model first-order and the second-order information simultaneously with joint and bone data stream, which shows notable improvement for the recognition accuracy. We choose to use the widely-used ST-GCN as the encoder in this paper, and the easily-obtained data streams (*e.g.* joint, bone, motion) to facilitate the multi-stream learning greatly.

### 2.3. Self-Supervised Skeleton-Based Action Recognition

Inspired by the success of self-supervised learning in image and video tasks [3, 15, 18, 26, 33], many self-supervised methods are proposed to capture effective skeleton representations. Previous methods can be divided into two categories by the input data: single-stream and multi-stream.

The single-stream methods usually only use joint-stream data as input. For example, LongT GAN [56] proposes a conditional skeleton inpainting architecture for learning a fixed-dimensional representation, and utilizes the encoder-decoder framework to regenerate the skeleton sequence with a adversarial strategy. MS<sup>2</sup>L [23] integrates multiple pretext tasks to learn more general features. P&C [42] proposes a novel training strategy which weakens the decoder and forces the encoder to learn a more informative representation. AS-CAL [37] exploits different augmentations of unlabeled skeleton sequences to learn action representations. ISC [44] integrates information between multiple forms of encoders to learn better features. ST-CL [13] explores the pretext task with different spatio-temporal observation scenes and devises a efficient action encoder. Se-BiReNet [30] considers both the kinematic and geometric dependencies and design a sequential bidirectional recursive network.

Multi-stream methods fuse the predictions of multiple data streams in downstream tasks, thus achieving better performance. SkeletonCLR [21] adopts the MoCo pipeline and

designs a simple baseline for this task. CrosSCLR [21] proposes a cross-stream knowledge mining strategy to enlarge the positive set for contrastive learning, where potential positive samples are mined with the reference of feature similarity in other data streams. AimCLR [14] introduces the extremely-augmented skeleton sequences and proposes a nearest neighbor mining to discover the potential positive sample in the memory bank. However, both single-stream methods and multi-stream methods are dragged down by easy positives. For one thing, the positive pairs in these methods are still constructed within a single data stream, indicating the correspondence between different data streams is overlooked in contrastive learning. For another, these works lack of necessary manipulation to enhance the hardness of positive features.

### 3. Methods

In this section, we first introduce preliminaries in Section 3.1. Next, the pipeline of Intra-Stream Contrastive Learning for Skeleton Representation (IntraCLR) and Inter-Stream Contrastive Learning for Skeleton Representation (InterCLR) are described in Section 3.2. Then, we demonstrate the Positive Feature Transformation (PFT) in Section 3.3. Finally, more details of CSCLR are introduced in Section 3.4. Briefly, we take a two-stream case (stream- $u$  and stream- $v$ ) as an example and illustrate the overview of our method in Figure 2.

#### 3.1. Preliminaries

Initially, we are supposed to be given a skeleton sequence  $x \in \mathbb{R}^{T \times C \times V}$  as input, where  $T$  is the temporal length,  $C$  is the number of channels, and  $V$  represents the number of keypoints of human body. Based on the raw skeleton data  $x$ , sample  $x^u$  and sample  $x^v$  in stream  $u$  and  $v$  are generated by method in Section 3.4. Our aim is to train a skeleton encoder  $f$  in self-supervised manner to be effectively transferable to down-stream tasks, *e.g.* skeleton-based action recognition. Inspired by the great success of contrastive learning, we develop the self-supervised skeleton representation learning framework based on the recent advanced practice MoCo-v2.

The contrastive learning framework takes skeleton samples as input, and consists of three components: *Data Augmentation*, *Feature Extraction*, and *Model Training*. We will introduce the first two components in this subsection, and the model training of IntraCLR and InterCLR will be described in the next subsection. In the common contrastive learning pipeline, a skeleton sample is initially transformed into different augments, and these augments can be seen as positive sample pairs. Besides, other augmented samples transformed by different skeletons are regarded as negative samples. After extracting features of these samples, an InfoNCE [31] loss for instance discrimination is introduced

for model training. Below, we take the skeleton data of stream- $u$   $x^u$  as example for elaboration.

**Data Augmentations.** We utilize a data augmentation module to transform the raw skeleton sequence  $x^u$  into different augments  $x_q^u$  and  $x_k^u$ , where  $x_q^u$  represents the query sample, and  $x_k^u$  represents the key sample. Due to the same source of these two augments,  $x_q^u$  and  $x_k^u$  are considered to be positive pairs for training. Both spatial and temporal augmentations are utilized for randomly transforming the input skeleton data, and the augmentation details are introduced in Section 4.2.

**Encoder.** Two GCN-based skeleton encoders  $f_q^u$  and  $f_k^u$  are constructed for extracting deep features of  $x_q^u$  and  $x_k^u$  respectively, depicted as  $h_q^u = f_q^u(x_q^u; \theta_q^u)$ ,  $h_k^u = f_k^u(x_k^u; \theta_k^u)$ . We use ST-GCN [52] as the backbone network in practice. In the training process, momentum update is adopted when optimizing key encoder  $f_k^u$ , and only parameters in query encoder  $f_q^u$  is updated with gradient backpropagation. Denoting the parameters of  $f_q^u$  as  $\theta_q^u$  and those of  $f_k^u$  as  $\theta_k^u$ ,  $\theta_k^u$  is updated by:

$$\theta_k^u \leftarrow m\theta_k^u + (1 - m)\theta_q^u, \quad (1)$$

where  $m \in [0, 1)$  is a momentum coefficient. Due to the momentum update, encoder  $f_k^u$  is able to avoid the rapid change and maintain the consistency of key representation. In the following of skeleton encoder, a simple MLP layer  $g$  with ReLU is utilized to project the hidden vector  $h$  into a low-dimension feature space. For the query and key sample, the corresponding deep skeleton feature are computed as  $z_q^u = g_q^u(h_q^u)$ ,  $z_k^u = g_k^u(h_k^u)$ .

#### 3.2. Intra-Stream and Inter-Stream Contrastive Learning for Skeleton Representation

In IntraCLR, only a single data stream is taken as input, the extracted feature of the two augments  $z_q^u$  and  $z_k^u$  are defined as positive pairs. As for negative pairs, similar to MoCo [15], a dynamic memory bank  $M_u = \{m_i^u\}_{i=1}^M$  is introduced to store negative samples. To enlarge the amount of negative samples, the memory bank consider a dictionary as a queue, where samples  $x_k^u$  enqueue and dequeue in each iteration. To facilitate model training, contrastive loss is used for instance discrimination. If the query data  $x_q^u$  is similar to its positive key  $x_k^u$ , the contrastive loss would be low. Otherwise, if the query data  $x_q^u$  is similar to its negative key in the memory, this loss would be high. Thus, InfoNCE loss is used to pull the positive pairs close in the feature space, while push the negative pairs away. The loss of IntraCLR can be formulated as follows:

$$\mathcal{L}_{intra}^u = -\log \frac{\exp(z_q^u \cdot z_k^u / \tau)}{\exp(z_q^u \cdot z_k^u / \tau) + \sum_{i=1}^M \exp(z_q^u \cdot m_i^u / \tau)}, \quad (2)$$



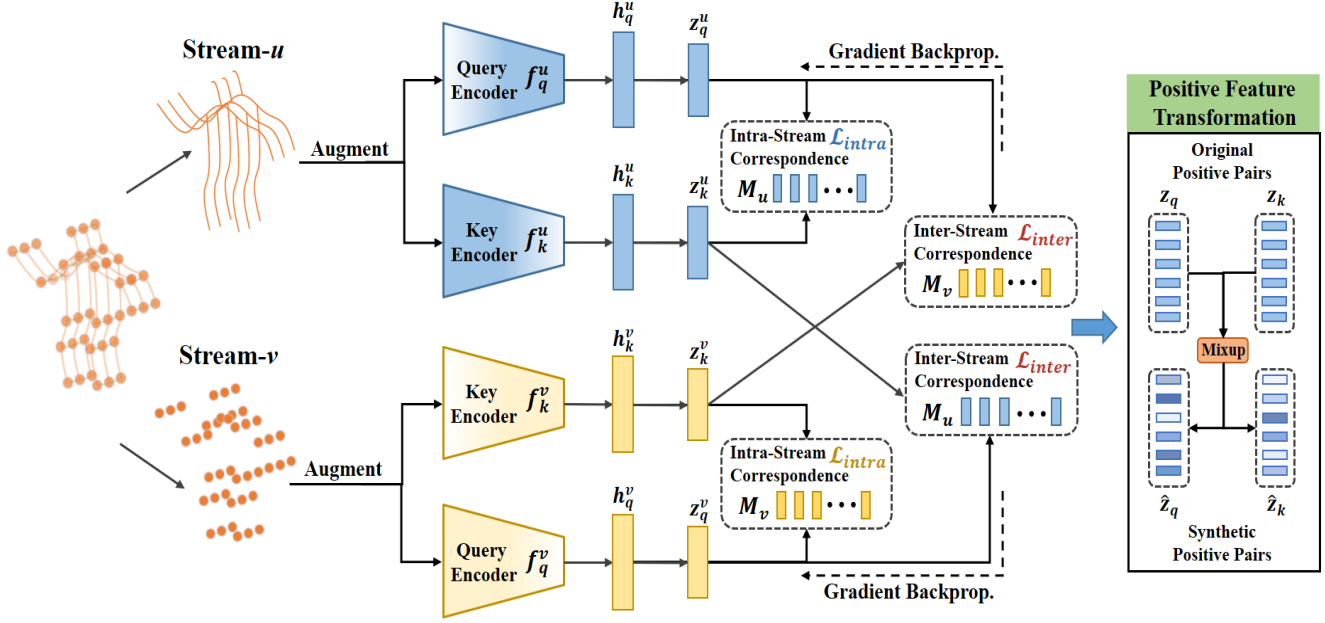


Figure 2. The Framework of our Cross-stream Contrastive Learning for Skeleton-based Action Recognition. The model takes skeleton data with different data streams as input, then extracts the query and key features. Next, intra-stream contrastive learning and inter-stream contrastive learning are adopted for training. Finally, the positive set is adjusted with feature transformation. Synthetic positive pairs are generated and features are updated to be more dissimilar (different color shades), which enable the model to learn more robust features for downstream tasks.

where  $m_i^u \in M_u$ ,  $\tau$  is temperature coefficient, and we use dot product as our similarity function.

In addition to IntraCLR, we demonstrate another contrastive objective InterCLR based on the way of inter-stream learning. To elaborate InterCLR, we first introduce another stream of skeleton data (stream- $v$ ) and extract GCN feature as IntraCLR, denoted as  $x^v$  and  $z^v$  respectively. Similarly, both query feature  $z_q^v$  and key feature  $z_k^v$  are generated for further training. When multiple streams of skeleton input are involved, the aim of contrastive learning would be stream-invariant. Specifically, the similarity of positive pairs from different streams (*e.g.* stream- $u$  and stream- $v$ ) should be maximized, since they both correspond to the same raw skeleton data from the source. Compared with the paired sample within a single stream ( $z_q^u$  and  $z_k^u$ ), the inter-stream alignment brings in harder positive samples, and the increasing variance between inter-stream positive pairs ( $z_q^u$  and  $z_k^v$ ) will lead to more implicit information for contrastive learning. With sample pairs from different streams, the query encoder is able to learn more robust skeleton feature and capture spatial-temporal information of input skeleton sequence more effectively, which benefits the knowledge transfer to downstream tasks. After the feature extraction in both stream- $u$  and stream- $v$ , the inter-stream

loss can be formulated as:

$$\mathcal{L}_{u \rightarrow v} = -\log \frac{\exp(z_q^u \cdot z_k^v / \tau)}{\exp(z_q^u \cdot z_k^v / \tau) + \sum_{i=1}^M \exp(z_q^u \cdot m_i^v / \tau)}, \quad (3)$$

$$\mathcal{L}_{v \rightarrow u} = -\log \frac{\exp(z_q^v \cdot z_k^u / \tau)}{\exp(z_q^v \cdot z_k^u / \tau) + \sum_{i=1}^M \exp(z_q^v \cdot m_i^u / \tau)}. \quad (4)$$

As shown in equation 3 and equation 4, the bi-directional inter-stream loss considers features from different streams as input. Due to the discrepancy of inherent information in different streams, positive pairs with lower similarity are constructed in this way, which facilitates more robust representation learning.

**Multi-stream Scenario.** When a set of streams  $\mathcal{S} = \{S_i\}_{i=1}^{N_s}$  are taken as input, where  $N_s$  represents the number of data streams, we need to combine both IntraCLR and InterCLR losses from all used streams. For IntraCLR, losses are combined as:

$$\mathcal{L}_{intra} = \sum_{u \in \mathcal{S}} \mathcal{L}_{intra}^u. \quad (5)$$

Also, the objective of InterCLR can be computed as:

$$\mathcal{L}_{inter} = \sum_{u \in \mathcal{S}} \sum_{v \in \mathcal{S}} \mathcal{L}_{u \rightarrow v}, \quad (6)$$

where  $u \neq v$ .

Finally, we obtain the overall loss function as the combination of  $\mathcal{L}_{intra}$  and  $\mathcal{L}_{inter}$ , where the former imposes invariance to data augmentations within a single stream, and the latter injects the correspondence between different streams.

$$\mathcal{L} = \mathcal{L}_{intra} + \mathcal{L}_{inter}. \quad (7)$$

### 3.3. Positive Feature Transformation

Except for generating positives with data augmentation pipeline, we propose to adjust the positive set by Positive Feature Transformation (PFT). PFT is designed to generate hard positives through feature extrapolations, which increases the variance of the positive set. In the guarantee of stable and smooth score distribution and gradient, the harder positives can be beneficial the transfer performance of downstream tasks [14, 59].

Figure 3 (a) shows two positive pairs in feature space. Similarly, we take the input of stream- $u$  as example. In order to encourage the model to generate more robust features, we manipulate the positive pair  $z_q^u$  and  $z_k^u$  to moderately increase the view variance between them. In other words, we add reasonable perturbations to the features and generate synthetic harder positives for contrastive learning.

Inspired by the design of Mixup [55] and Manifold Mixup [49], we utilize weighted sum to integrate the positive pairs. As shown in Figure 3 (b), the newly generated synthetic feature  $\hat{z}_q^u$  and  $\hat{z}_k^u$  are computed as:

$$\begin{aligned} \hat{z}_q^u &= \lambda z_q^u + (1 - \lambda) z_k^u, \\ \hat{z}_k^u &= \lambda z_k^u + (1 - \lambda) z_q^u. \end{aligned} \quad (8)$$

To address our aim of generating harder positives, we should guarantee that the similarity of  $\hat{z}_q^u$  and  $\hat{z}_k^u$  will be lower than that of  $z_q^u$  and  $z_k^u$ . Formally, this means  $\hat{z}_q^u \hat{z}_k^u \leq z_q^u z_k^u$  holds, when taking dot product as similarity metric. Based on Equation (8), we can obtain:

$$\begin{aligned} \hat{z}_q^u \hat{z}_k^u &= [\lambda z_q^u + (1 - \lambda) z_k^u] [\lambda z_k^u + (1 - \lambda) z_q^u] \\ &= 2\lambda(1 - \lambda)(1 - z_q^u z_k^u) + z_q^u z_k^u, \end{aligned} \quad (9)$$

where  $0 \leq z_q^u z_k^u < 1$ . From the Equation (9), we can observe that if  $\lambda \geq 1$ , then  $2\lambda(1 - \lambda) \leq 0$  and  $\hat{z}_q^u \hat{z}_k^u \leq z_q^u z_k^u$  holds. Thus, we formulate a beta-like distribution  $F(\alpha, \mu)$  as follows and sample the value of  $\lambda$  from it,  $\lambda \sim F(\alpha, \mu)$ .

$$F(\alpha, \mu) = \text{Beta}(\alpha, \alpha) \times \mu + 1, \quad (10)$$

where  $\alpha$  is the parameter of *Beta* distribution and  $\mu$  is a parameter to limits the value of  $\lambda$ . Consequently, this makes  $\lambda \in [1, 1 + \mu]$ . As shown in Figure 3 (c), a subtle direction shift is applied to the original positive pairs  $z_q^u$  and  $z_k^u$ . After that, the angle between transformed feature  $\hat{z}_q^u$  and  $\hat{z}_k^u$  are

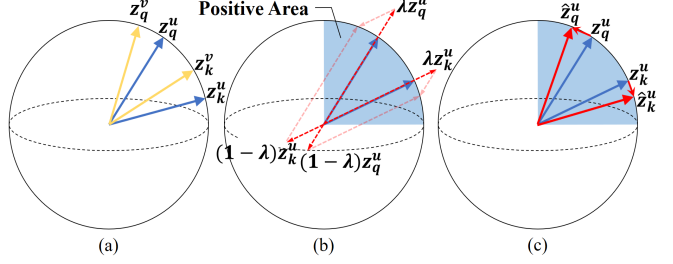


Figure 3. An illustration of our proposed PFT. After PFT, the original positive pairs are repelled, resulting in increasing view variance in positive set and more robust self-supervised representation learning.

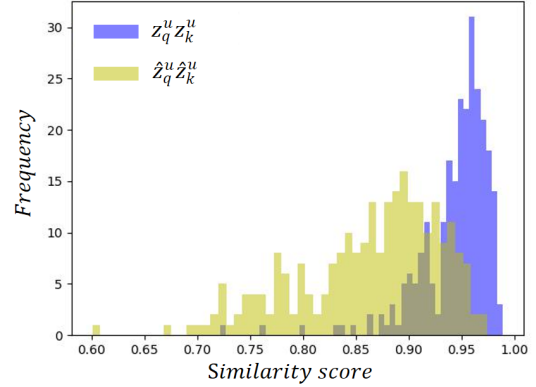


Figure 4. The comparison of similarity score distribution  $z_q^u z_k^u$  and  $\hat{z}_q^u \hat{z}_k^u$ . After PFT, the hard positives are generated, and the variation of similarity in the positive set is evidently increased.

larger than the angle between original feature  $z_q^u$  and  $z_k^u$ , indicating the larger variance in the positive set.

In contrastive learning, the similarity scores of positive pairs are defined to be non-negative, while that of negative pairs are non-positive. Intuitively, the angle between transformed positive feature should be  $90^\circ$  at most, which corresponds to the positive area in Figure 3. In case that the similarity score of a few positive pairs are excessively decreased to negative values after the feature transformation, we do not adopt PFT on those pairs. Figure 4 shows the similarity score distribution of  $z_q^u z_k^u$  and  $\hat{z}_q^u \hat{z}_k^u$  in a randomly-selected mini-batch. First, the similarity scores between positive pairs become lower, and the variances in the positive set get higher after PFT. Second, there is no negative value of  $\hat{z}_q^u \hat{z}_k^u$  after adopting the strategy above, which averts the contradiction with the definition of positive sample pairs.

### 3.4. Model Details

**Different Data Streams of 3D Skeleton.** Except for the 3D joint coordinates  $x \in \mathbb{R}^{T \times C \times V}$ , multiple data streams of skeleton are generated as [39, 40], *e.g.* bone, motion. Bone is represented as a vector pointing to its target joint

from its source joint  $(x_{:,v_2} - x_{:,v_1})$ , which contains not only the length information, but also the direction information. And motion is obtained by computing the displacement between adjacent frames  $(x_{t_2,:} - x_{t_1,:})$ , which brings in the temporal shift information. In this paper, we use three streams: joint, bone and motion in experiments.

**Nearest Neighbors Mining (NNM).** Similar to AimCLR [14], we also apply NNM and take the mined samples in memory queue as positives for training. Although these selected neighbors are stored in the memory queue, they are quite similar with query  $z_q^u$ . With these mined positives, the intra-stream objective of contrastive learning could be formulated as follows:

$$\mathcal{L}_{N-intra}^u = -\log \frac{\exp(z_q^u \cdot z_k^u / \tau) + \sum_{i \in N_+^u} \exp(z_q^u \cdot m_i^u / \tau)}{\exp(z_q^u \cdot z_k^u / \tau) + \sum_{i=1}^M \exp(z_q^u \cdot m_i^u / \tau)}, \quad (11)$$

where  $N_+^u$  is the index set of selected neighbors of  $z_q^u$  in  $M_u$ .

**Multi-stage Training Strategy.** In the earlier epochs, the unstable model is not able to provide effective features for NNM and PFT, so we perform multi-stage training for CSCLR. The whole training process is divided into three stages: Basic training, Basic training + NNM, Basic training + NNM + PFT. In the first stage, we train our model with the combination of  $\mathcal{L}_{intra}$  and  $\mathcal{L}_{inter}$ . Next, we adopt NNM to increase the number of positive pairs with reliable neighbors. Finally, we adopt PFT to generate hard positives and improve the robustness of representation learning.

## 4. Experiments and results

In this section, we first introduce datasets in Section 4.1. Next, the experimental settings for a fair comparison are described in Section 4.2. Then, the comparison with the state-of-the-art methods are shown in Section 4.3. Finally, we demonstrate the results of ablation study in Section 4.4.

### 4.1. Dataset and Evaluation Metric

**NTU-RGB+D 60.** The dataset consists of 56,578 skeleton sequences with 3D joint coordinate for skeleton-based action recognition, and these sequences are labeled with 60 action categories. In each skeleton graph, 25 joints throughout the body are set as nodes, and their 3D coordinates are obtained by Kinect V2 cameras. There are two suggested protocols. The first one is Cross-Subject (xsub), where skeleton sequences in training set and validation set are collected from different subjects. And the other one is Cross-View (xview), where skeleton sequences in training set and validation set are collected from different camera views.

**NTU-RGB+D 120.** The dataset [24] is an extension of NTU-60, and this large-scale benchmark dataset contains 113,945 skeleton sequences in 120 action categories. There

are two suggested protocols. The first one is Cross-Subject (xsub), where skeleton sequences in training set and validation set are collected from different subjects. And the other one is Cross-Setup (xsetup), where skeleton sequences in training set and validation set are collected from different setup IDs.

**PKU-MMD Dataset.** It contains almost 20,000 action sequences covering 51 action classes. It consists of two subsets. Part I is an easier version for action recognition, while part II is more challenging with more noise caused by view variation. We conduct experiments under the cross-subject protocol on the two subsets.

### 4.2. Experiment Settings

All the experiments are conducted on PyTorch [34]. For data pre-processing, we follow AimCLR [14], SkeletonCLR [21] and CrosSCLR [21] for a fair comparison.

**Data Augmentation.** We use the same data augmentations as AimCLR [14], which includes *Normal Augmentations* and *Extreme Augmentations*. Here, we focus on the cross-stream contrastive learning, so we only introduce these augmentations briefly.

The Normal Augmentations consists of one spatial augmentation *Shear* and one temporal augmentation *Crop*. The Extreme Augmentations includes four spatial augmentations: *Shear*, *Spatial Flip*, *Rotate*, *Axis Mask* and two temporal augmentations: *Crop*, *Temporal Flip* and two spatio-temporal augmentations: *Gaussian Noise*, *Gaussian Blur*.

**Self-supervised Pre-training.** We use the same setting for contrastive learning as that in AimCLR, SkeletonCLR and CrosSCLR, the batch size is set to 128, and the size of memory queue is 32768. In the training process, we use SGD with momentum (0.9) and weight decay (0.0001). For the multi-stage training strategy mentioned in Section 3.4, the model is trained for 150 epochs in the basic training stage, and then trained for 150 epochs with NNM involved. Finally, the model is trained for 200 epochs with PFT involved. The learning rate is set as 0.1, and decreases to 0.01 at epoch 250. In NNM, we only take the top-1 nearest neighbor in the memory queue as positive sample. For a fair comparison, we use the weights of [0:6; 0:6; 0:4] for fusing the three-stream predictions like other multi-stream GCN methods.

**Linear Evaluation Protocol.** To verify our model, we adopt linear evaluation for the action recognition task. To be specific, we train a linear classifier (a fully-connected layer followed by a softmax layer), and the encoder is frozen in optimization. The model is trained for 100 epochs with learning rate 3.0 (decrease to 0.3 at epoch 80).

**Finetuned Evaluation Protocol.** We append a linear classifier to the trained encoder, and the whole model (encoder and classifier) is optimized for the action recognition task, to compare it with fully supervised methods. The

model is trained for 100 epochs with learning rate 0.1 and weight decay 0.0001.

**Semi-supervised Evaluation Protocol.** The encoder is pre-trained with all data in self-supervised manner, and the whole model is finetuned with only 1% or 10% randomly selected labeled data.

**Competitors.** In addition to the self-comparison experiments, we compare the proposed CSCLR method with the state-of-the-art methods, including single-stream and multi-stream works. The single-stream setting includes LongT GAN [56], MS<sup>2</sup>L [23], P&C [42], SeBiReNet [30], AS-CAL [37], ST-CL [13] and MG-AL [54], the performance of joint stream in SkeletonCLR [21], CrosSCLR [21] and AimCLR [14] are also demonstrated for single-stream comparison. For the multi-stream setting, we mainly include 3s-SkeletonCLR, 3s-CrosSCLR and 3s-AimCLR, where predictions of three streams (joint, bone and motion) are fused.

Table 1. Linear evaluation accuracy comparisons with the state-of-the-art methods on NTU-60 dataset. “3s” represents three stream fusion. “†” means using cross-stream knowledge mining strategy proposed in 3s-CrosSCLR. The best results is in bold face.

Category	Method	Year	NTU-60 (%)	
			xsub	xview
Single Stream	LongT GAN	AAAI’18	39.1	48.1
	MS <sup>2</sup> L	MM’20	52.6	-
	P&C	CVPR’20	50.7	76.3
	SeBiReNet	ECCV’20	-	79.7
	SkeletonCLR	CVPR’21	68.3	76.4
	AS-CAL	Info. Sci.’21	58.5	64.8
	ST-CL	TMM’21	68.1	69.4
	MG-AL	TCSVT’22	64.7	68.0
	AimCLR	AAAI’22	74.3	79.7
	<b>CSCLR (Ours)</b>	-	<b>75.7</b>	<b>81.3</b>
Multiple Stream	3s-SkeletonCLR	CVPR’21	75.0	79.8
	3s-CrosSCLR	CVPR’21	77.8	83.4
	3s-AimCLR <sup>†</sup>	AAAI’22	78.6	82.6
	3s-AimCLR	AAAI’22	78.9	83.8
	<b>3s-CSCLR (Ours)</b>	-	<b>80.1</b>	<b>85.2</b>

#### 4.3. Comparison with the State-of-the-art

We compare the proposed method with state-of-the-art methods on NTU-60, NTU-120 and PKU-MMD datasets with the corresponding protocol.

**Linear Evaluation Protocol Results.** Results on NTU-60, NTU-120 and PKU-MMD are shown in Table 1, Table 2 and Table 3 respectively. For a fair comparison, we report both single-stream (joint) and multi-stream fusion results.

As can be seen from Table 1, the result of CSCLR single stream significantly outperforms the previous methods

Table 2. Linear evaluation accuracy comparisons with the state-of-the-art methods on NTU-120 dataset. The best results is in bold face.

Category	Method	Year	NTU-120 (%)	
			xsub	xsetup
Single Stream	P&C	CVPR’20	42.7	41.7
	AS-CAL	Info. Sci.’21	48.6	49.2
	ST-CL	TMM’21	54.2	55.6
	SkeletonCLR	CVPR’21	56.8	55.9
	MG-AL	TCSVT’22	46.2	49.5
	AimCLR	AAAI’22	63.4	63.4
	<b>CSCLR (Ours)</b>	-	<b>64.5</b>	<b>64.3</b>
Multiple Stream	3s-SkeletonCLR	CVPR’21	60.7	62.6
	3s-CrosSCLR	CVPR’21	67.9	66.7
	3s-AimCLR <sup>†</sup>	AAAI’22	68.0	68.7
	3s-AimCLR	AAAI’22	68.2	68.8
	<b>3s-CSCLR (Ours)</b>	-	<b>69.2</b>	<b>70.2</b>

Table 3. Linear evaluation accuracy comparisons with the state-of-the-art methods on PKU-MMD dataset. The best results is in bold face.

Category	Method	Year	PKU-MMD (%)	
			part I	part II
Single Stream	LongT GAN	AAAI’18	67.7	26.0
	MS <sup>2</sup> L	MM’20	64.9	27.6
	SkeletonCLR	CVPR’21	80.9	-
	AimCLR	AAAI’22	83.4	-
	<b>CSCLR (Ours)</b>	-	<b>85.3</b>	<b>31.8</b>
Multiple Stream	3s-SkeletonCLR	CVPR’21	84.9	21.2
	3s-AimCLR <sup>†</sup>	AAAI’22	87.4	39.5
	3s-AimCLR	AAAI’22	87.8	38.5
	<b>3s-CSCLR (Ours)</b>	-	<b>89.3</b>	<b>45.1</b>

with joint-stream data on NTU-60. For the fusion results, our CSCLR still maintains the advantage over the previous methods under both x-sub and x-view protocols. Compared with 3s-CrosSCLR, the CSCLR gains 2.3% and 1.8% accuracy improvements under x-sub and x-view respectively. When comparing with 3s-AimCLR, CSCLR outperforms it by 1.2% and 1.4% respectively. It is worth mentioning that the performance of 3s-AimCLR are decreased when using cross-stream knowledge mining strategy in 3s-CrosSCLR, while better performance are boosted when adopting the proposed CSCLR. The improved performance proves that our method is able to take the advantage of correlating information between different data streams and inherit knowledge in positive set.

In Table 2, our CSCLR also enjoys evident accuracy im-



provement over existing self-supervised methods on NTU-120. Under single-stream setting, CSCLR-joint significantly surpasses the advanced ST-CL (64.5% vs 54.2% on xsub and 64.3% vs 55.6% on xsetup). For fusion results, the accuracy of CSCLR is also higher than that of 3s-AimCLR (69.2% vs 68.2% on xsub and 70.2% vs 68.8% on xsetup). The advantages in Table 2 indicates the competitiveness of CSCLR when conducting experiments on large-scale datasets.

As demonstrated in Table 3, the proposed CSCLR outperforms all compared state-of-the-art self-supervised methods on PKU-MMD. Specifically in part I and part II subsets, our method achieves accuracy of 85.3% and 31.8% with single data stream, and gains accuracy of 89.3% and 45.1% respectively with multi-stream fusion. Notably, CSCLR leads 3s-AimCLR 6.6% under the part II subset, which indicates that the trained CSCLR model significantly benefits the downstream task on more challenging and noisy dataset.

Table 4. Finetuned evaluation accuracy comparisons with the state-of-the-art methods on NTU-60 and NTU-120 dataset. The best results is in bold face.

Method	Year	NTU-60 (%)		NTU-120 (%)	
		xsub	xview	xsub	xsetup
SkeletonCLR	CVPR’21	82.2	88.9	73.6	75.3
AimCLR	AAAI’22	83.0	89.2	76.4	76.7
<b>CSCLR (Ours)</b>	-	<b>84.7</b>	<b>90.4</b>	<b>76.9</b>	<b>77.6</b>
3s-ST-GCN	AAAI’18	85.2	91.4	77.2	77.1
3s-CrossCLR	CVPR’21	86.2	92.5	80.5	80.4
3s-AimCLR	AAAI’22	86.9	92.8	80.1	80.9
<b>3s-CSCLR (Ours)</b>	-	<b>87.0</b>	<b>93.6</b>	<b>80.2</b>	<b>81.6</b>

**Finetuned Evaluation Protocol Results.** We compare the CSCLR with other methods under finetuned evaluation protocol, and both single-stream and multi-stream fusion results are represented in Table 4. All the competitors are developed based on the same encoder, ST-GCN. Similar to results in linear evaluation protocol, our CSCLR outperform ST-GCN and other self-supervised methods on both NTU-60 and NTU-120 dataset, indicating the efficacy of our proposed method.

**Semi-supervised Evaluation Protocol Results.** To further evaluate the self-supervised model, we demonstrate the semi-supervised evaluation results with 1% and 10% labeled data on NTU-60 and PKU-MMD. As shown in Table 5, our CSCLR performs better than the previous competitors in most cases. The improvements on semi-supervised evaluation also indicate that CSCLR has positive effect of increasing the robustness of representation learning with extremely limited labeled data.

**Qualitative Results.** To evaluate our proposed method qualitatively, we apply t-SNE [46] with fixed settings to show the embedding distribution of the advanced AimCLR and our CSCLR on NTU-60-xview dataset. Figure 5 shows the comparison between results of CSCLR and AimCLR, it can be seen that the features from CSCLR performs better to enlarge the margin among the inter-class samples and reduce the intra-class distance at the same time, which demonstrates CSCLR is conducive to learn more discriminative features when transferring to downstream tasks.

#### 4.4. Ablation Study

To verify the effectiveness of the proposed method, we conduct ablation studies and provide the experimental results as follows. We follow the self-supervised pre-training, the linear evaluation and finetuned evaluation protocol.

**The Effectiveness of CSCLR.** To verify the effectiveness of the proposed CSCLR, we first reproduce the AimCLR and conduct multiple experiments in all used datasets. Both reproduced results and the results reported in the paper are introduced for comparison, three streams (joint, bone and motion) are utilized for experiments. As shown in Table 6, the proposed CSCLR achieves better performance than both AimCLR and AimCLR (Repro.), which validates the effectiveness of the CSCLR model. For single-stream results, CSCLR outperforms the AimCLR in all used stream, and especially achieves significant improvements in bone and motion stream. For the multi-stream fusion results, CSCLR also performs better than AimCLR in all used dataset.

**The Effectiveness of Inter-stream Contrastive Learning.** To further verify the proposed components in CSCLR, we conduct ablation studies on NTU-60 dataset and show the three-stream fusion results. As shown in Table 7, when only utilizing the intra-stream contrastive learning (w/ Intra.), we can achieve the recognition accuracy of 78.4% and 83.1% on xsub and xview respectively. After adding the inter-stream contrastive learning (w/ Inter.), the performances are boosted, reaching 79.2% and 84.1%. This illustrates that the inter-stream alignment of hard positive pairs benefits the transfer accuracy of downstream recognition task. Similar to linear evaluation results, the performances of finetuned evaluation are also improved. The baseline accuracies (w/ Intra.) are 85.4% and 91.2%, then they are boosted up to 86.2% and 91.9% respectively. Besides, the specific improvements on different streams and the fusion result on NTU-60-xview dataset are shown in Figure 6. When using the combination of intra-stream and inter-stream contrastive learning, the linear evaluation accuracy of each stream is obviously boosted, the improvements on these streams are 2.7%, 2.1% and 5.5% respectively. The general improvements on all used streams confirm the effectiveness of the proposed inter-stream contrastive learning

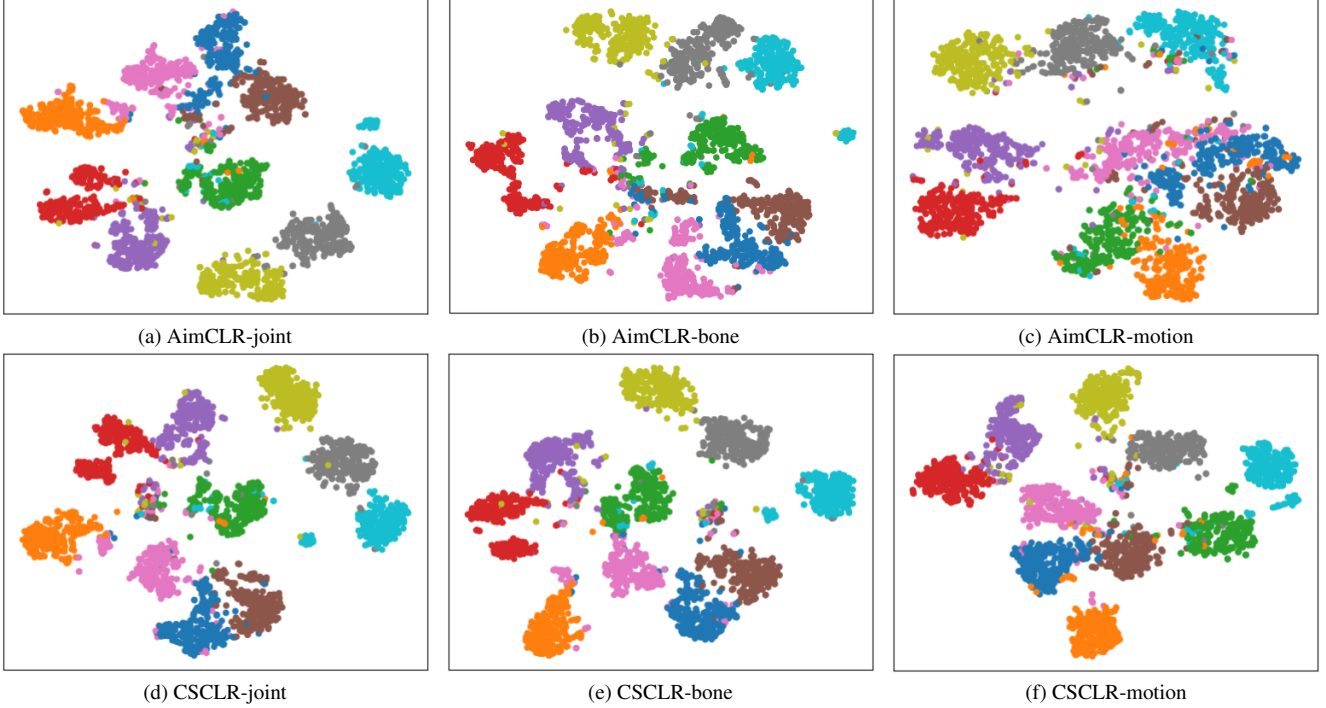


Figure 5. The t-SNE visualization of embedding distributions on NTU-60-xview dataset. Embeddings from 10 categories are sampled and visualized with different colors. AimCLR and our CSCLR results in all used streams (joint, bone, motion) are showed from left to right respectively.

Table 5. Semi-supervised evaluation results on NTU-60 and PKU-MMD dataset. 1% and 10% indicate that we use 1% and 10% labeled data for training the action recognition model. The best results is in bold face.

Method	Year	NTU-60-xsub		NTU-60-xview		PKU-MMD I		PKU-MMD II	
		1%	10%	1%	10%	1%	10%	1%	10%
LongT GAN	AAAI'18	35.2	62.0	-	-	35.8	69.5	12.4	25.7
MS <sup>2</sup> L	MM'20	33.1	65.2	-	-	36.4	70.3	13.0	26.1
ISC	MM'21	35.7	65.9	38.1	72.5	37.7	72.1	-	-
3s-CrosSCLR	CVPR'21	51.1	74.4	50.0	77.8	49.7	82.9	10.2	28.6
3s-Colorization	ICCV'21	48.3	71.7	52.5	78.9	-	-	-	-
3s-AimCLR	AAAI'22	54.8	78.2	54.3	81.6	57.5	86.1	15.1	33.4
<b>3s-CSCLR (Ours)</b>	-	<b>55.4</b>	<b>78.6</b>	<b>57.1</b>	<b>81.8</b>	<b>58.0</b>	<b>86.1</b>	<b>18.0</b>	<b>38.6</b>

method.

#### The effectiveness of Positive Feature Transformation.

From Table 7, we can observe that when adopting positive feature transformation (w/ PFT), the linear evaluation accuracy on xsub and xview are further improved by 0.9% and 1.1%. A similar improvement can be observed when using the finetuned evaluation protocol, the accuracy are improved by 0.8% and 1.7%. Also, the specific improvements of PFT on each used skeleton joint stream are shown in Figure 6. The improvements on joint, bone and motion stream

are 1.2%, 1.2% and 0.3% respectively. Due to that the original positive pairs in motion stream have already been dissimilar, the gains of PFT in this stream is slightly inferior to that in other streams. The improvement shows that the proposed PFT can enforces model to learn more robust feature for downstream tasks by generating harder positive samples with reasonable feature transformation.

**Skeleton Activation Map.** To show how CSCLR works, we utilize CAM (Class Activation Map) [57] to calculate the activation map of skeleton sequence, and the activated

Table 6. Linear evaluation results compared with AimCLR on NTU-60, PKU-MMD I and NTU-120 dataset. “Repro.” means the reproduced results. The best results is in bold face.

Method	Stream	NTU-60(%)		PKU-MMD(%)		NTU-120(%)	
		xsub	xview	part I	part II	xsub	xsetup
AimCLR	joint	74.3	79.7	83.4	-	63.4	63.4
AimCLR (Repro.)	joint	74.5	77.4	83.1	30.2	63.3	63.9
<b>CSCLR (Ours)</b>	joint	<b>75.7</b>	<b>81.3</b>	<b>85.3</b>	<b>31.8</b>	<b>64.5</b>	<b>64.3</b>
AimCLR	bone	73.2	77.0	82.0	-	62.9	63.4
AimCLR (Repro.)	bone	72.3	76.8	81.9	28.2	61.2	64.7
<b>CSCLR (Ours)</b>	bone	<b>76.3</b>	<b>80.1</b>	<b>83.3</b>	<b>39.2</b>	<b>64.8</b>	<b>65.1</b>
AimCLR	motion	66.8	70.6	72.0	-	57.3	54.4
AimCLR (Repro.)	motion	64.5	71.2	72.2	29.9	53.5	55.9
<b>CSCLR (Ours)</b>	motion	<b>72.3</b>	<b>76.7</b>	<b>81.2</b>	<b>31.0</b>	<b>59.2</b>	<b>59.6</b>
3s-AimCLR	joint+bone+motion	78.9	83.8	87.8	38.5	68.2	68.8
3s-AimCLR (Repro.)	joint+bone+motion	78.6	83.2	86.9	35.0	67.9	69.5
<b>3s-CSCLR (Ours)</b>	joint+bone+motion	<b>80.1</b>	<b>85.2</b>	<b>89.3</b>	<b>45.1</b>	<b>69.2</b>	<b>70.2</b>

Table 7. Ablation study results on NTU-60 dataset. The best results is in bold face.

Protocol	w/ Intra.	w/ Inter.	w/ PFT	NTU-60 (%)	
				xsub	xview
Linear Eval.	✓			78.6	83.2
	✓	✓		79.2	84.1
	✓	✓	✓	<b>80.1</b>	<b>85.2</b>
Finetuned Eval.	✓			85.4	91.2
	✓	✓		86.2	91.9
	✓	✓	✓	<b>87.0</b>	<b>93.6</b>

joints in several frames are displayed in Figure 7. From this figure, we can observe that CSCLR is able to concentrate on the informative joints and thus achieve better performance. For example, the arms and hands are relatively informative for the action of brushing hair and hand waving. Compared with AimCLR, CSCLR pays higher attention to these informative body parts. The differences of skeleton activation map verify that CSCLR is conducive to learn effective features for downstream tasks.

**Choice of streams in InterCLR.** To explore the impact of streams used in InterCLR, we further conduct experiments for comparison. We first set the method which only utilizes IntraCLR as a baseline for comparison, and then add InterCLR loss with different streams. The comparison follows linear evaluation protocol, and the detailed experiment results on NTU-60 are shown in Table 8, in which “2s” and “3s” represent two-stream and three-stream re-

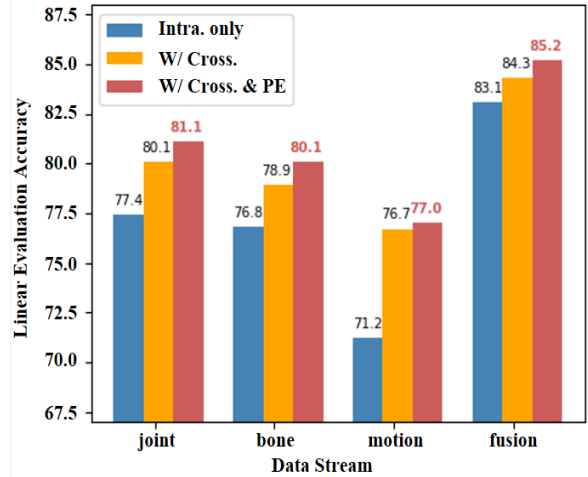


Figure 6. Linear evaluation results with different data stream on NTU-60-xview dataset.

spectively. From this table, we can observe that the accuracy performances are obviously boosted when adding 2s-InterCLR. When replacing 2s-InterCLR with 3s-InterCLR, the accuracy performances can be further improved, reaching 80.1% and 85.2% on xsub and xview respectively. The improvements again demonstrate the effectiveness of our proposed method.

**Choice of Parameters in  $F(\alpha, \mu)$ .** As described in Section 3.4,  $\lambda \sim F(\alpha, \mu)$ ,  $F(\alpha, \mu) = \text{Beta}(\alpha, \alpha) \times \mu + 1$ . To determine the hyper-parameters  $\alpha$ ,  $\mu$ , we study how these parameters impact the performance of contrastive learning.

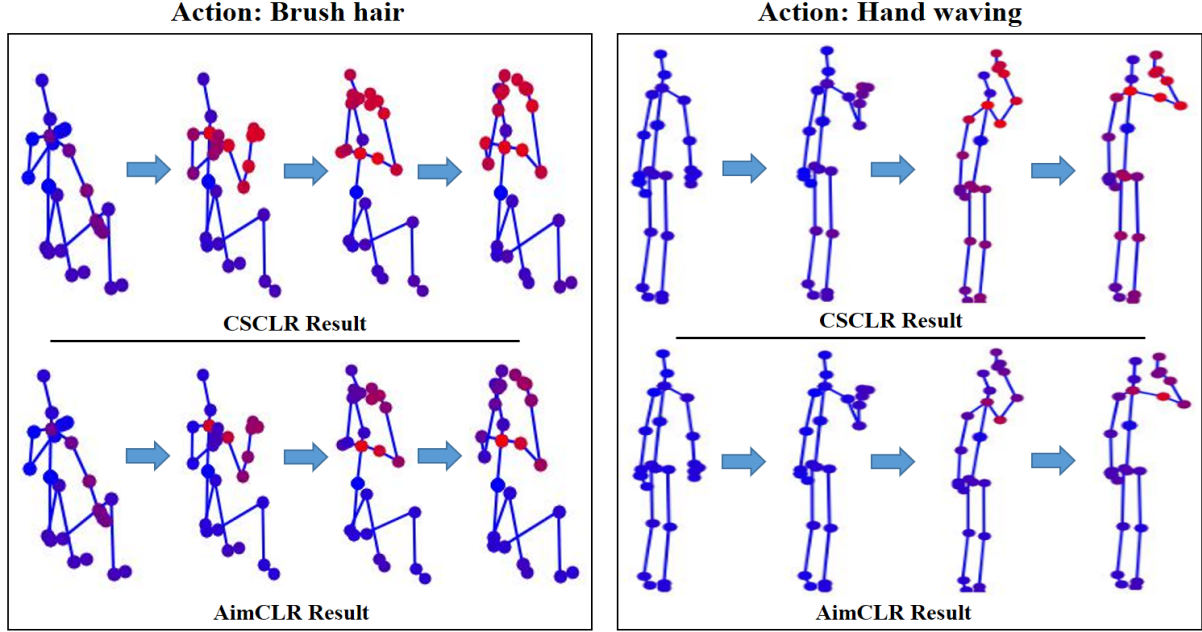


Figure 7. Skeleton Activation Map for sample actions *Hand waving* and *Brush hair*. The red joints are activated joints, and the blue joints are non-activated joints. (Best viewed in color.)

Table 8. Linear evaluation results with different streams which are used in InterCLR.

Method	Stream in Inter.			NTU-60 (%)	
	joint	bone	motion	xsub	xview
Base	-	-	-	78.6	83.2
With 2s-Inter.	✓	✓		79.7	84.8
	✓		✓	79.3	84.5
		✓	✓	79.0	84.1
With 3s-Inter.	✓	✓	✓	<b>80.1</b>	<b>85.2</b>

As shown in Table 9, we found that the best setting is  $\alpha = 2.0$ ,  $\mu = 1.0$ . When  $\alpha = 0.5$ ,  $\lambda$  tends to be close to the upper bound  $\mu$  or lower bound  $1 + \mu$ , the accuracy of recognition is invariant to the choice of  $\mu$ . When  $\alpha = 2.0$ ,  $\lambda$  is sampled to be  $1 + \mu/2$  with high probability, better accuracy are obtained. Therefore, we choose  $\alpha = 2.0$ ,  $\mu = 1.0$  as default setting.

## 5. Conclusion

In this paper, we advocate the significant effect of inter-stream contrast pairs, and therefore propose a Cross-Stream Contrastive Learning framework for skeleton-based action

Table 9. Results of CSCLR-joint with various  $\alpha$ ,  $\mu$  on NTU-60 dataset. The best results is in bold face.

$\alpha$	$\mu$	NTU-60-joint (%)	
		xsub	xview
0.5	1.0	75.0	80.9
	1.2	74.9	80.8
	1.5	74.9	80.9
2.0	1.0	<b>75.7</b>	<b>81.3</b>
	1.2	75.0	81.1
	1.5	75.2	81.0

Representation learning (CSCLR). Except for utilizing easy contrast pairs within a single data stream, CSCLR additionally introduces inter-stream contrast pairs as hard samples to formulate a better representation learning. Besides, to conduct in-depth study on positive samples, we propose a Positive Feature Transformation (PFT) strategy which adopts feature-level manipulation to increase the variance of positive pairs. Extensive experiment results on NTU-RGB+D 60, NTU-RGB+D 120 and PKU-MMD demonstrate that CSCLR achieves superior classification accuracy on linear evaluation, finetuned evaluation and semi-supervised evaluation. The ablation study demonstrates that both the inter-stream hard contrast pairs and the PFT strategy can promote more robust representation learning. In



the future work, we will continue to explore the design of contrast pairs specified for skeleton-based tasks. The unified data augmentations are used for different skeleton data streams, which ignore the discrepancy between them. Thus, with the customized data augmentations for each data stream, the performance of downstream tasks would be further improved.

## References

- [1] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9902–9912, 2022. 3
- [2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 609–617, 2017. 3
- [3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1, 3
- [4] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 3
- [5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 3
- [6] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13359–13368, 2021. 3
- [7] Yi-Bin Cheng, Xipeng Chen, Dongyu Zhang, and Liang Lin. Motion-transformer: self-supervised pre-training for skeleton-based action recognition. In *Proceedings of the 2nd ACM International Conference on Multimedia in Asia*, pages 1–6, 2021. 1
- [8] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20186–20196, 2022. 3
- [9] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11162–11173, 2021. 3
- [10] Xinpeng Ding, Nannan Wang, Xinbo Gao, Jie Li, Xiaoyu Wang, and Tongliang Liu. Kfc: An efficient framework for semi-supervised temporal action localization. *IEEE Trans. Image Proc.*, 30:6869–6878, 2021. 2
- [11] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1110–1118, 2015. 3
- [12] David Freire-Obregón, Paola Barra, Modesto Castrillón-Santana, and Maria De Marsico. Inflated 3d convnet context analysis for violence detection. *Machine Vision and Applications*, 33:1–13, 2022. 1
- [13] Xuehao Gao, Yang Yang, Yimeng Zhang, Maosen Li, Jing-Gang Yu, and Shaoyi Du. Efficient spatio-temporal contrastive learning for skeleton-based 3d action recognition. *IEEE Trans. Multi.*, 2021. early access. 3, 8
- [14] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 762–770, 2022. 1, 2, 4, 6, 7, 8
- [15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 3, 4
- [16] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 3
- [17] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies*. A field guide to dynamical recurrent neural networks. IEEE Press, 2001. 3
- [18] Jing Huang, Yan Huang, Qicong Wang, Wenming Yang, and Hongying Meng. Self-supervised representation learning for videos by segmenting via sampling rate order prediction. *IEEE Trans. Circ. Syst. Video Tech.*, 32(6):3475–3489, 2021. 3
- [19] Qihong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3288–3297, 2017. 3
- [20] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Skeleton-based action recognition with convolutional neural networks. In *IEEE International Conference on Multimedia & Expo Workshops*, pages 597–600, 2017. 3
- [21] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 3d human action representation learning via cross-view consistency pursuit. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4741–4750, 2021. 2, 3, 4, 7, 8
- [22] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5457–5466, 2018. 3
- [23] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2490–2498, 2020. 1, 3, 8

- [24] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans. Pat. Anal. Mach. Intel.*, 42(10):2684–2701, 2019. 7
- [25] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Trans. Image Proc.*, 27(4):1586–1599, 2017. 1
- [26] Yang Liu, Keze Wang, Lingbo Liu, Haoyuan Lan, and Liang Lin. Tcgl: Temporal contrastive graph for self-supervised video representation learning. *IEEE Trans. Image Proc.*, 31:1978–1993, 2022. 3
- [27] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 143–152, 2020. 2, 3
- [28] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 3
- [29] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021. 3
- [30] Qiang Nie, Ziwei Liu, and Yunhui Liu. Unsupervised 3d human pose representation with viewpoint and pose disentanglement. In *Proceedings of the European Conference on Computer Vision*, pages 102–118, 2020. 3, 8
- [31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4
- [32] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision*, pages 631–648, 2018. 3
- [33] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11205–11214, 2021. 3
- [34] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems*, pages 8026–8037, 2017. 7
- [35] Liliana Lo Presti and Marco La Cascia. 3d skeleton-based human action classification: A survey. *Pattern Recognition*, 53:130–147, 2016. 1
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 3
- [37] Haocong Rao, Shihao Xu, Xiping Hu, Jun Cheng, and Bin Hu. Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. *Information Sciences*, 569:90–109, 2021. 2, 3, 8
- [38] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *Proceedings of the European Conference on Computer Vision*, pages 153–170. Springer, 2020. 3
- [39] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7912–7921, 2019. 6
- [40] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2019. 1, 2, 3, 6
- [41] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4263–4270, 2017. 3
- [42] Kun Su, Xiulong Liu, and Eli Shlizerman. Predict & cluster: Unsupervised skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9631–9640, 2020. 3, 8
- [43] Ning Sun, Ling Leng, Jixin Liu, and Guang Han. Multi-stream slowfast graph convolutional networks for skeleton-based action recognition. *Image Vis. Comput.*, 109:1014–1030, 2021. 1
- [44] Fida Mohammad Thoker, Hazel Doughty, and Cees GM Snoek. Skeleton-contrastive 3d action representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1655–1663, 2021. 2, 3
- [45] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *arXiv preprint arXiv:2005.10243*, 2020. 1
- [46] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Jour. Mach. Learn. Res.*, 9(11):2579–2605, 2008. 9
- [47] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2014. 3
- [48] Raviteja Vemulapalli and Rama Chellappa. Rolling rotations for recognizing human actions from 3d skeletal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4471–4479, 2016. 3
- [49] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, Aaron Courville, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. *arXiv preprint arXiv:1806.05236*, 2018. 2, 6
- [50] Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on*

- Computer Vision and Pattern Recognition*, pages 16041–16050, June 2022. 3
- [51] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1297, 2012. 3
  - [52] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7444–7452, 2018. 1, 3, 4
  - [53] Siyuan Yang, Jun Liu, Shijian Lu, Meng Hwa Er, and Alex C Kot. Skeleton cloud colorization for unsupervised 3d action representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13423–13433, 2021. 2
  - [54] Yang Yang, Guangjun Liu, and Xuehao Gao. Motion guided attention learning for self-supervised 3d human action recognition. *IEEE Trans. Circ. Syst. Video Tech.*, 32(12):8623–8634, 2022. 8
  - [55] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2, 6
  - [56] Nenggan Zheng, Jun Wen, Risheng Liu, Liangqu Long, Jianhua Dai, and Zhefeng Gong. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2644–2651, 2018. 1, 3, 8
  - [57] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 10
  - [58] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. 2
  - [59] Rui Zhu, Bingchen Zhao, Jingen Liu, Zhenglong Sun, and Chang Wen Chen. Improving contrastive learning by visualizing feature transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10306–10315, 2021. 3, 6