https://eprints.gla.ac.uk/293121/

Deposited on: 1 March 2023

Focalized Contrastive View-invariant Learning for Self-supervised Skeleton-based Action Recognition

Qianhui Men[a,1,*], Edmond S. L. Ho[b], Hubert P. H. Shum[c], Howard Leung[a]

[a]*Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China*
[b]*School of Computing Science, University of Glasgow, Glasgow, G12 8RZ, United Kingdom*
[c]*Department of Computer Science, Durham University, Durham, DH1 3LE, United Kingdom*

Abstract

Learning view-invariant representation is a key to improving feature discrimination power for skeleton-based action recognition. Existing approaches cannot effectively remove the impact of viewpoint due to the implicit view-dependent representations. In this work, we propose a self-supervised framework called Focalized Contrastive View-invariant Learning (FoCoViL), which significantly suppresses the view-specific information on the representation space where the viewpoints are coarsely aligned. By maximizing mutual information with an effective contrastive loss between multi-view sample pairs, FoCoViL associates actions with common view-invariant properties and simultaneously separates the dissimilar ones. We further propose an adaptive focalization method based on the pairwise similarity to enhance the contrastive learning for a clearer cluster boundary in the learned space. Different from many existing self-supervised representation learning work that rely heavily on supervised classifiers, FoCoViL performs well on both unsupervised and supervised classifiers with superior recognition performance. Extensive experiments also show that the proposed contrastive-based focalization generates a more discriminative latent representation.

Keywords:
self-supervised learning, skeleton-based action recognition, contrastive learning

---

*Corresponding author
Email address: `qianhui.men@eng.ox.ac.uk` (Q. Men)
[1]Present address: Department of Engineering Science, University of Oxford, OX1 3PJ, UK. Work done while Qianhui Men was at City University of Hong Kong.

1. Introduction

Self-supervised skeletal human action recognition (HAR) aims at automatically detecting a robust representation to cluster and identify actions from class-agnostic skeletal data. Compared to supervised models that heavily rely on action labels [1, 2, 3], recognition without manually labeling is considered more efficient and more comprehensive to learn representative features with large-scale data. A few unsupervised attempts [4, 5, 6, 7, 8] have recently achieved classification results comparable to supervised models, which indicates that label information may not be necessary for extracting useful representations for discriminating action dynamics. In this work, we consider the challenging domain of self-supervised action recognition from multi-view features, where the action sequences are captured under different viewpoints. The diverse view appearance introduces large intra-class variations in the feature representation that substantially impacts the clustering performance. Unlike supervised view-invariant learning that benefits from action labels, learning without label guidance is more challenging that usually requires detecting implicit consistency between viewpoints.

Existing skeleton-based HAR works [2, 6] learn view-invariant features from the skeleton descriptions enriched by multiple views. A simple yet effective pre-processing scheme is to align the body key joints with a local coordinate system [9, 10]. However, since this view-invariant transformation is sensitive to the quality of the posture captured from different viewpoints, such as different levels of self-occlusions, the transformed multi-view actions are still mismatched with many inherent view-specific representations [11]. Later on, deep neural networks are utilized to automatically search for the optimal viewpoints for every skeleton sequence [2, 12], which requires strong supervision to guide this additional training. In unsupervised learning, an adversary view-aware classifier is introduced in [13] to discard view information from RGB and depth data. Another attempt [6] learns the view-variant and view-invariant features from spatial and temporal skeletal representations respectively, while the recognition performance is less satisfactory on the multi-view actions. So far, removing the viewpoint impact in the self-supervised skeleton recognition is still an open problem.

In this paper, we propose FoCoViL, the focalized contrastive view-invariant learning framework, for view-independent and discriminative self-supervised action recognition. FoCoViL consists of two complementary components, namely contrastive view-invariant learning (CoViL) and focalization. Figure 1 shows the effect of FoCoViL on learning a view-invariant latent space, where the action representations of the same scene are more closely distributed compared to CoViL.

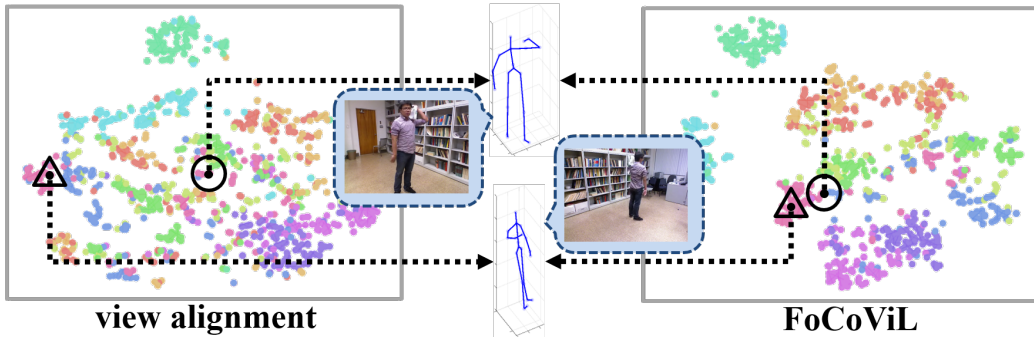**view alignment**   **FoCoViL**

Figure 1: Visualizing the latent space when comparing an action pair of *carrying* from the same scene but different viewpoints. Different colors in the t-SNE visualization refer to the attributes of real classes. The view alignment cannot handle the inherent heterogeneity between viewpoints, such as different levels of self-occlusions shown in the middle skeletons, yielding a less satisfactory space distribution. With FoCoViL, the actions under the same scene but different viewpoints (highlighted in circles and triangles) are geometrically closer in the latent space, which achieves better view invariance.

First, CoViL explores the implicit relationship between viewpoints. With the facing directions aligned [10], CoViL works as a refinement scheme to group the multi-view actions by maximizing their mutual information under different viewpoints, such that the learned representation is robust to view changes. Specifically, under a self-supervised auto-encoder backbone, we propose to maximize the agreement of the actions under the same scene but different viewpoints (*i.e.* positive pairs "+"), which helps extract the common features among them that are view-invariant. Meanwhile, we propose to enlarge the disagreement of actions under different scenes (*i.e.* negative pairs "−"), which benefits the clustering with a sparse latent space. The two goals are jointly achieved by a close form of contrastive loss [14] for its superior ability to find and compare the similarity and dissimilarity in the self-supervised representations. With CoViL, we construct a latent space that is more robust to view dynamics compared to the one generated by the low-level viewpoint alignment [5].

Second, we propose to enhance the latent space by wrapping a novel focalization method around contrastive learning in CoViL. This is to solve the imbalanced training data issue inherent in many existing self-supervised systems - the hard samples that dominate the misclassification are not fully investigated, leading to an ambiguous sample distribution in the latent space. To mitigate imbalance in contrastive learning, several works that are highly related to ours mainly focus on

3

mining hard negatives, such as synthesizing new samples [15, 16] or using class labels as priors [17]. In contrast, our method considers adaptively "focalizing" both the hard positives and negatives under the learned representative similarity. We take advantage of the effective pairwise similarity estimation in CoViL to dynamically identify and re-balance the easy and hard multi-view action pairs. This is done by defining the hard pairs as either sparse positive pairs (same scenes that are far away) or dense negative pairs (different scenes that are close) in the projected latent space, and the easy ones the other way round. The proposed focalization reduces the weightings of easy pairs that provided limited information while focusing on pushing hard negative pairs away and pulling hard positives closer, thereby enforcing a more distinct decision boundary in the latent space.

Experimental results show that FoCoViL outperforms state-of-the-art self-supervised models on five benchmark 3D action datasets including Northwestern-UCLA (N-UCLA) [18], NTU RGB+D 60 [19], NTU RGB+D 120 [20], UWA 3D Multiview Activity II (UWA3D) [21], and PKU-MMD [22]. Unlike some self-supervised representation learning approaches [23, 8] rely heavily on supervise classifiers, FoCoViL performs well with both supervised and unsupervised classifiers. The extensive experiments on representation space evaluation also indicate that the proposed FoCoViL produces a more robust latent space.

The main contributions are summarized in three folds:

- We propose a self-supervised framework to progressively learn a discriminative skeleton-based action representation that is robust for both supervised and unsupervised evaluation protocols. The source code is publicly available: https://drive.google.com/file/d/1VKRF2S3-LrOiXV4BLSjxIlCUewMnxH_W

- We propose contrastive view-invariant learning, which maximizes the mutual information between multi-view action pairs by adapting contrastive learning, aiming to refine the latent representations with high-level view-invariant features.

- As a novel attempt of applying focalization to contrastive learning, we have demonstrated its feasibility of learning a more robust and unbiased representation with the action recognition task.

The rest of this paper is organized as follows. Section 2 reviews the related background research. Section 3 presents the proposed FoCoViL framework for self-supervised action recognition. The experiments and analysis are conducted in

Section 4 with quantitative recognition and reconstruction results, and qualitative latent space evaluations. Finally, we conclude this paper in Section 5.

2. Related Work

2.1. Self-supervised Action Representation Learning

Learning of action representation has been proposed for many years in computer vision applications. The learned latent representation usually includes action semantics which is feasible for multiple downstream tasks, such as action classification [24, 25] and motion generation [26, 27, 28]. Among many self-supervised feature extraction baselines, the auto-encoder is usually adopted to learn the representation space with its superior ability to denoise the action information. Holden et al. [29] first proposed a convolutional auto-encoder to construct the latent space from the encoder output. By operating the high-level features, the model is functional in many areas, such as action interpolation and comparison. Later on, with a hierarchical RNN auto-encoder in both spatial and temporal domains, Wang et al. [27] developed a high-quality representation space that is motivated for precise action modeling.

As a vision-based learning task, the effectiveness of self-supervised representation is frequently explored in RGB-based action understanding. To learn the contextual coherence in action representation, Lai and Xie [30] matched pixel-wise correspondence from the spatial-temporal color information. Han et al. [31] exploited action representations from multiple modalities of RGB streams and optical flow. However, these RGB-based action recognition models usually learn contrastive representation from background or visual consistency [32]. With the visual information unavailable, the challenge of learning self-supervised skeleton-based action representation mainly comes from the diverse pose information under different view observations [6, 23].

In self-supervised skeleton-based action representation learning, existing works such as [4, 33, 34, 5] mainly focus on preserving the action-dependent features as much as possible to identify samples [35]. For example, an adversarial discriminator is used to assist the auto-encoder to rectify the reconstructed action for a more distinctive representation [34, 4]. Since the encoder is dominant in disclosing the action features, Su et al. [5] proposed to strengthen the encoder by exploiting different auto-encoder structures (P&C), such as fixing the encoded state or the decoder weights. Apart from an action-level auto-encoder, they also designed an additional feature-level auto-encoder to reduce the dimensionality of the learned representations, which results in a two-round training process. Because

of the lack of communication within latent space, the derived action representation of these works is not robust to large intra-class variations. Other prior works considered learning feature representations by modeling actions with denoised poses (Denoised-LSTM [33]), different temporal patterns (MS^{2}L [36], MCAE-MP [37]), different spatial-temporal augmentations (AS-CAL [38], ST-CL [39]), or within group activities [40]. However, the learned space is still underestimated with diverse view representations and imbalanced sample distribution, leading to a less satisfactory clustering.

Recently, several studies also show that pre-training the self-supervised representation benefits the supervised [7, 41] or semi-supervised learning [42] in action recognition tasks. In this paper, we investigate self-supervised representations with view invariance to improve action recognition. This is done by an end-to-end framework with balanced pairwise learning based on the performed viewpoints, such that the learned representations of the same scene are more clearly grouped with fewer errors (*i.e.* higher purity and recognition accuracy).

2.2. View-invariant Human Action Recognition

A robust action recognition model requires the learned representations to be less sensitive to viewpoints. In RGB, depth, or optical flow videos, it is common for people to remove the view-dependent backgrounds by learning the view-specific focuses in different viewpoints [18, 13, 43]. In contrast, a natural advantage of the 3D skeleton is that the view-invariant features are more easily extracted with the body joint positions. However, the commonly used view-independent representations, including the statistics-based histogram of joint orientations [44] or geometry-based pairwise joint distance [11] will discard some semantic information that is useful for recognizing action patterns.

Another branch of works [10, 39, 45] employed coordinate transformation (*i.e.* rotations, translations) to align or synthesize multi-view actions. For example, Gao et al. [39] compared the action pairs augmented from arbitrary viewpoints by a contrastive loss. Paoletti et al. [45] utilized gradient reversing to fool a viewpoint regressor that predicts the rotation of the transformed action. However, with the self-occlusions in different directions, the transformed skeleton is not accurate enough to imitate the new viewpoint. Moreover, Zhang et al. [2] automatically learned the view-invariant adaption per action and achieved promising results compared with pre-defined transformations. It further convinces that the recognition ability can be dramatically affected by the inconsistency between viewpoints that cannot be removed manually.

6

Recently, Gao et al. [46] exploited fisher contrastive learning to extract view semantics from different scales of body parts. By assembling multiple spatial features, Guan et al. [47] proposed a feature enhanced approach that is robust to view variation. Nie et al. [6] proposed SeBiReNet that models the view variant and invariant features from the geometric poses and temporal dynamics, respectively, to better denoise the skeleton data. Instead of specifying the view-invariant features, we purify the latent representations through the implicit correlations learned between multi-view samples, which yields a better recognition performance.

2.3. *Contrastive Learning*

Contrastive learning [48, 49, 14] is a self-supervised representation learning method that differentiates between individual instances based on their pairwise similarity. As a label-agnostic approach, contrastive learning is purely based on the feature-level correlations of samples, which is very popular in large-scale visual tasks. Chen et al. [14] augmented the real-world image such as cropping, rotations, or blurring, as positive instances to extract common properties in contrastive learning, where they further used a large minibatch size to increase the capacity of contrastive learning that achieves superior classification performance. However, the above methods rely on heavy data accommodation that requires either argumentation or multi-modal representations.

There are also several works adopting contrastive learning in skeleton-based action recognition. Rao et al. [38] learned the self-supervised action representation with contrastive learning, where they exploited similar augmentation strategies of images [14] onto skeleton sequences. Lin et al. [36] constrained the contrastive loss with multi-task learning from motion prediction and classification. Since the learned embeddings from contrastive loss are hard to discriminate, effective selection strategies [50, 49, 51] are proposed for positive or negative samples to enhance the contrastive metric. As a remarkable work, Momentum Contrast (MoCo) [50] performed contrastive learning by selecting negative samples from a memory bank that is updated dynamically with an extra momentum encoder, which is later adopted in many recent work [23, 52, 53, 54, 8] to improve skeleton-based action recognition. For example, Wang et al. [53] proposed a contrast-reconstruction representation network (CRRL) to contrast between the spatial postures and motion velocity to enhance the action representation learning. Li et al. [23] proposed CrosSCLR to enrich the feature receptive field by contrasting augmented skeleton sequences from bone, joint, and motion features. Their model feasibility was further generalized by Guo et al. [8] (AimCLR) with more
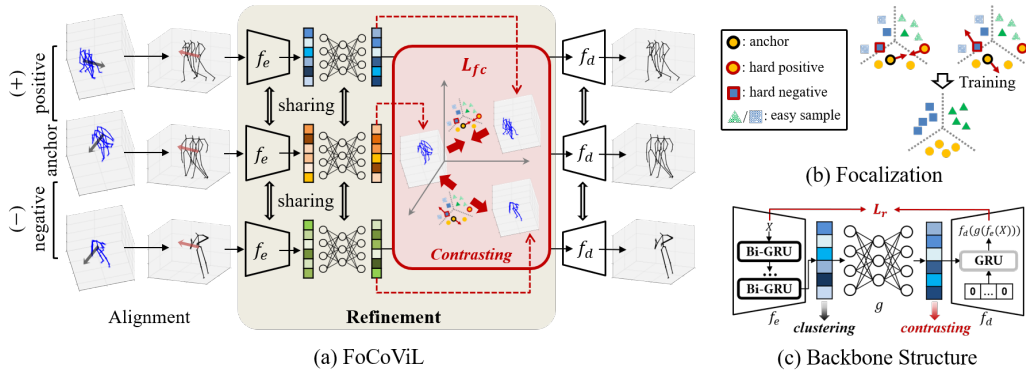
Figure 2: The proposed Focalized Contrastive View-invariant Learning (FoCoViL) framework. (a) FoCoViL aims at progressively extracting the view-invariant action representations. The multi-view actions are initially aligned *w.r.t.* the facing direction. In the refinement step, the proposed multi-view contrastive learning discards the implicit view-specific appearance by enlarging the agreement of the same scene under different viewpoints ("+" pairs), and facilitates clustering with a sparse space by enlarging different scenes ("−" pairs), which is further enhanced by an adaptive focalization. (b) The motivation of focalization. For each target action, focalization focuses on adjusting the hard action pairs that dominate the misclassifications. (c) The architecture of the shared auto-encoder backbone. The clustering is conducted before the encoded representation passes into the projection net g.

diverse spatial and temporal augmentations. However, their learned representations were restricted by the augmented patterns and the cross-modality settings greatly increase the model complexity. Since the natural correlation between different camera viewpoints is still underexplored in self-supervised 3D action understanding, in this paper, we investigate how the viewpoint affects the learned representation by contrasting multi-view information without augmenting the total data size.

3. Methodology

We aim at progressively learning an effective representation space consisting of view-invariant action features for a discriminative clustering. To achieve this goal, we propose FoCoViL, which removes the view influence on the feature-level representation extracted by an RNN-based auto-encoder backbone, as demonstrated in Fig. 2. We first conduct a coarse-level transformation to align different viewpoints. Then, in the refinement step, we disentangle the remaining view-specific features from the latent space by finding the inherent correlations of the same action scene under different viewpoints. Since the contributions to the

8

space learning vary from sample to sample with the *hard positives* (*i.e.* intra-class variation) and *hard negatives* (*i.e.* inter-class similarity) contributing more, we further propose a focalized contrastive loss to cope with the imbalanced learning complexity by adaptively adjusting their training intensity, thereby promoting the quality of the converged latent space.

Problem Formulation In the training set $\hat{\mathbf{X}}$ with V viewpoints, we represent the ith scene of human action $\hat{X}_i^u \in \hat{\mathbf{X}}$ as a sequence of poses, *i.e.* $\hat{X}_i^u = \{\hat{x}_{i,1}^u, \hat{x}_{i,2}^u, ..., \hat{x}_{i,T}^u\}$, under a specific viewpoint $u \in V$. Each pose $\hat{x}_{i,t}^u \in \mathbb{R}^{3 \times N}$ at frame t contains N joint locations under 3D skeleton, and T is the maximum timestamp. As a self-supervised classification task, we tend to learn a view-invariant mapping f_e without the guidance of action label.

3.1. View Alignment

The skeletons are misaligned under different viewpoints, which brings difficulty in recognizing actions. Following [5], we transform the views for aligning the multi-view actions to the same facing direction, resulting in a more comparable representation space for the latter refinement phase.

As 3D joint coordinates are demonstrated with different scopes under different camera points, we first translate them into a local coordinate system with the origin as the root joint $\hat{x}_{i,0}^u(root)$ at the initial frame, thereby removing the dependency of the camera position and the global displacement. We then match the directions of the translated actions using a rotation matrix $R = [\hat{r}_0, \hat{r}_1, \hat{r}_2]$ with:

$$
\begin{aligned}
&r_0 = \hat{x}_{i,0}^u(spine) - \hat{x}_{i,0}^u(root), \\
&r_1 = \tilde{r}_1 - \tilde{r}_1 \cdot \hat{r}_0, \text{ and } \tilde{r}_1 = \hat{x}_{i,0}^u(lhip) - \hat{x}_{i,0}^u(rhip), \\
&r_2 = r_0 \times r_1,
\end{aligned}
\tag{1}
$$

where $\hat{x} = \frac{x}{\|x\|}$ denotes the unit vector. Here, r_0 points from the root to the spine $\hat{x}_{i,0}^u(spine)$, r_1 is the orthogonal projection of the vector between left $\hat{x}_{i,0}^u(lhip)$ and right hip $\hat{x}_{i,0}^u(rhip)$ on r_0.

The obtained joint $n \in N$ in the pose $x_{i,t}^u$ is transformed by:

$$
x_{i,t}^u(n) = R^{-1}(\hat{x}_{i,t}^u(n) - \hat{x}_{i,0}^u(root)).
\tag{2}
$$

Therefore, the action sequence after alignment is represented by $X_i^u = \{x_{i,1}^u, x_{i,2}^u, ..., x_{i,T}^u\}$ within the transformed training set $\mathbf{X}$.

9

3.2. Contrastive View-invariant Learning (CoViL)

We propose a Contrastive View-invariant Learning (CoViL) approach to automatically refine the view-invariant features by contrasting the multi-view representations under the same and different scenes. This is to tackle the problem that many implicit view-specific appearances, such as the inferred joint positions from different directions of self-occlusions, cannot be aligned by the coarse transformation. In a closely related work [6], the view-independent and view-dependent features are being processed as pose and temporal dependencies, respectively. However, the two types of features are not domain-specific and thus are non-trivial to be explicitly grouped. In contrast, we rely on the pairwise action correlations. By maximizing the mutual information of the compressed representations across views, CoViL can better suppress the view-specific factors and derive a highly view-invariant latent space.

With the observation that the same scene should have closer similarity than different scenes, CoViL discards the scene-invariant (*i.e.* view-variant) information by associating the same scene together. The objective of CoViL is to maximize the agreement of the same scene under different viewpoints (*i.e.* "+" positive sample pairs), as well as the disagreement of different scenes (*i.e.* "−" negative sample pairs) by contrastive learning. With the same action, compulsively correlating the "+" pairs will reinforce the co-occurrences that are only related to the underlying action content. In addition, enlarging the differences between "−" pairs will avoid an overly compact representation space while constricting the "+" pairs, such that the dissimilar actions are sparsely distributed to facilitate self-supervised clustering.

Particularly, we select positive and negative pairs based on a minibatch of I anchor samples that are randomly picked. For each anchor X_i^u in the minibatch, we propose to increase the similarity between X_i^u and its corresponding positive sample X_i^v ($v \neq u$), which reduces the motion variations caused by viewpoints. Note that X_i^u and X_i^v are from the same scene but have different viewpoints. We also propose to maximize the dissimilarity between X_i^u and its negative samples from other scenes which consist of two batches: $\{X_j^u\}_{j=1, j \neq i}^I$ under the same viewpoint u, and $\{X_j^v\}_{j=1, j \neq i}^I$ under the different viewpoint v. The far distributed negative pairs will ensure the sparsity of the resulted space.

We integrate our proposed positive and negative pair design via the batch contrastive loss based on InfoNCE loss function [14] due to its superior capacity in modeling the pairwise correlations. The multi-view contrastive loss defined on an

10

anchor X_i^u is given by:

$$L_c(X_i^u) = -\sum_{v \in V \setminus u} \log \frac{S(X_i^u, X_i^v)}{\sum_{j \neq i}(S(X_i^u, X_j^u) + S(X_i^u, X_j^v))}, \qquad (3)$$

where the proximity S is the similarity measurement between a pair of samples. Note that the loss is summed for all the viewpoints in V except u, and it is not bounded by only two views despite the proximity S is counted based on the pairwise manner. The proposed multi-view contrastive loss uses viewpoint information as the indicator to group positive samples under the general form of contrastive loss [14]. Empirically, if more viewpoints are included in training, the learned representation space is more informative by disentangling multi-view features. More specifically, S between sample pair X and Y is determined by their encoded distance r as:

$$S(X, Y) = \exp\left(\frac{g(f_e(X)) \cdot g(f_e(Y))}{\tau \| g(f_e(X)) \| \cdot \| g(f_e(Y)) \|}\right) = \exp\left(\frac{r(X, Y)}{\tau}\right), \qquad (4)$$

where τ is the temperature parameter [14, 48] that controls the scale of the similarity. We have evaluated τ with several choices $\{0.1, 0.5, 1, 2\}$, among which 0.5 performs the best. g is a projection network consisting of two fully-connected layers to integrate features, and we use the cosine distance r as the similarity metric following the general form of contrastive learning [14]. Here, the proposed multi-view contrastive loss is conducted on the projected feature to facilitate a more informative latent space for $f_e(X)$ that benefits the recognition.

3.3. Focalization

We propose to adaptively balance the easy and hard samples via dynamic focalization on the proposed CoViL, thereby increasing model robustness and reducing misclassifications. This is particularly challenging in unsupervised models, as the hard samples cannot be explicitly mined due to the lack of a label-guided distribution.

Here, we focus on rebalancing the representations within the scope of self-supervised contrastive learning, while the vast majority of other works focus on solving imbalance in supervised cross-entropy from true sample distributions. Instead of recognizing individual instances, in contrastive learning, we balance easy and hard samples via the pairwise sample similarity inherited from CoViL. As an effective solution, focal loss [55] aims at detecting and emphasizing the hard

11

instances from the probability outcome. While similar in purpose, our method attempts to solve the imbalanced similarity of sample pairs based on contrastive learning, from the observation that contrastive learning lacks a mechanism to maintain balanced training. By entangling and disentangling sample pairs in terms of their representation similarity, the focalized CoViL (FoCoViL) further enhances the latent space learning with a clearer decision boundary between clusters (see Fig. 2(b)). Note that the proposed focalized contrastive learning is not a simple reweighting scheme but balancing and improving the distributions in representation space by self-supervised hard sampling with adjustable hardness, which has more generalizable advantages and consistent performance improvements (see Table 8).

In particular, we propose a dynamic-scaled focal loss based on the geometric distance of the contrastive representations. Inspired by the evidence that the same scene should have similar feature expressions, we consider a "+" pair as hard if they are too far distributed. Analogous to positive pairs, we define the hard negatives if the "−" pair is too close. FoCoViL intuitively pulls the same scene with very different representations closer while pushing the different scenes with similar representations apart. Numerically, we monotonously increase the weight to the "+" pair X_i^u and X_i^v if their cosine similarity r is getting close to -1, and increase the weight to the "−" pair if r is near 1. The dynamic weight w_+ for positive pair and w_- for negative pair are defined by:

$$w_+ = \sigma(1 - r(X_i^u, X_i^v)),$$
$$w_- = \sigma(\frac{1}{2I-2}\sum_{j \neq i}[(1+r(X_i^u, X_j^u))+(1+r(X_i^u, X_j^v))]). \tag{5}$$

The modulating factors $1-r(X,Y)$ and $1+r(X,Y)$ are added as pair weightings for the positive and negative samples, respectively, which adaptively differentiate between the easy and hard pairs based on the pairwise similarity. The *sigmoid* activation $\sigma(\cdot)$ with the common form $\sigma(x) = \frac{1}{1+e^{-x}}$ is to incorporate nonlinearity to contrastive loss, and the scaling term $\frac{1}{2I-2}$ is to balance the quantities of positive and negative pairs.

By decomposing L_c in Eq. 3, the proposed focalized multi-view contrastive

loss L_{fc} is defined as:

$$L_{fc}(X_i^u) = -\sum_{v \in V \setminus u} [w_+ \log S(X_i^u, X_i^v)$$

$$- w_- \log \sum_{j \neq i} (S(X_i^u, X_j^u) + S(X_i^u, X_j^v))]. \quad (6)$$

Compared with the inliers that stay very close to the cluster centers, the hard sample pairs usually include outliers that are scattered near the cluster boundaries. By focusing on the hard pairs, we establish a more robust latent space with fewer misclassified outliers. Note that the proposed focalized contrastive loss is heuristic and can be extended to other models adopting contrastive learning.

3.4. The Multi-view Auto-encoder Backbone

To maintain the action representation, we employ an effective sequential auto-encoder as the backbone network sharing among the multi-view actions (see Fig. 2(c)). From [56], the encoder usually plays a more important role than the decoder to integrate representative features. We thus consider a portable decoder by feeding the empty frame (zero vector) to every step of the decoder, such that the model only focuses on the hidden representation delivered from the encoded output. Instead of a two-stage encoder for feature extraction [5], our FoCoViL makes full use of a single auto-encoder in an end-to-end fashion that already achieves promising results.

The structure consists of a three-layer bi-directional encoder f_e to derive the latent representation, a linear projection net g that is specifically designed for contrastive learning, and a single-layer uni-directional decoder f_d for reconstruction purposes. Both f_e and f_d are under the Gated Recurrent Unit (GRU) architecture to process frame-wise information. For each action X_i^u, the reconstruction loss L_r is defined as:

$$L_r(X_i^u) = \frac{1}{T} \sum_{t=1}^{T} \parallel f_d(g(f_e(X_i^u)) - X_i^u \parallel, \quad (7)$$

where T is the total number of frames in the action video. Since the sequential auto-encoder reconstructs T frames of action, the loss is counted based on every frame and then averaged.

3.5. Training and Classification

The final objective of the proposed FoCoViL is given by $\alpha L_{fc} + \beta L_r$, where the α and β are the trade-offs between two losses. By optimizing the combination of L_{fc} and L_r, the whole network will search for the optimal representation space for the downstream classification task, where $f_e(X)$ is used for evaluation. Note that we do not cluster on the compressed output $g(f_e(X))$, since it may discard some information that is necessary for classification [14].

4. Experiments

4.1. Datasets and Experimental Setup

4.1.1. Datasets

To test the robustness of our model, we evaluate five benchmark 3D action datasets with diverse scales and properties, *i.e.* N-UCLA [18], NTU RGB+D 60 [19], NTU RGB+D 120 [20], PKU-MMD [22], and UWA3D [21]. The adopted datasets were all captured using Kinect with diverse self-occlusions under a multi-view environment. N-UCLA contains 10 types of human daily activities from three different viewpoints. PKU-MMD, NTU RGB+D 60, and NTU RGB+D 120 are large-scale action datasets with around 20,000, 56,880, and 114,480 clips covering 51, 60, and 120 types of human activities, respectively, where NTU RGB+D 120 is the largest benchmark for skeletal action recognition. UWA3D is more challenging due to four distinct action directions captured from the front, left, right, and top views.

4.1.2. Implementation Details

For pre-processing, the raw skeleton is initially normalized to $[-1, 1]$, and before feeding in the model, all action clips are interpolated to a fixed length with 50 frames. Our FoCoViL is trained under the combination of L_{fc} and L_r, where we set $\alpha = \beta = 1$ for N-UCLA, NTU RGB+D, and PKU-MMD, and $\alpha = 1, \beta = 2$ for UWA3D because of its noisy skeletons. τ is set to 0.5 for the similarity measurement in Eq. 4. Inside the auto-encoder, 1024 hidden units are used in the GRU cell for each layer, and the unit sizes in the projection net are 512 and 1024 respectively. The training batch size is 128 for NTU RGB+D and 64 for the other three datasets, and we adopt Adam optimizer with a learning rate of 0.0001 and a decay rate of 0.95. Following [5] and [36], we conduct cross-view (CV) evaluations on N-UCLA, NTU RGB+D 60, and UWA3D, and cross-subject (CS) evaluations on NTU RGB+D 120 and PKU-MMD. Unlike [5] only tested two views on UWA3D, we test on all evaluation combinations with any two

Table 1: Performance comparisons (%) on N-UCLA with supervised (Linear) and unsupervised (1-NN) evaluation protocol.

Method	N-UCLA
Linear Classifier	
LongT GAN [4]	74.3
Denoised-LSTM [33]	76.8
MS^2L [36]	76.8
SeBiReNet [6]	80.3
AS-CAL [38]	75.6
ST-CL [39]	81.2
MCAE-MP [37]	83.6
CRRL [53]	83.8
FoCoViL	**84.2**
1-Nearest Neighbour (1-NN) Classifier	
P&C [5]	84.9
MCAE-MP [37]	79.1
CRRL [53]	86.4
CoViL	86.7
FoCoViL	**88.3**

viewpoints for training and the rest for testing, which results in 12 experimental trials in total.

4.1.3. Evaluation Protocols for Classification

For a fair comparison, we adopt both supervised and unsupervised classifiers to evaluate the encoded representation for the action recognition task. **Linear Classifier**: as in [23, 8], a fully-connected layer (together with a *softmax* activation) is trained on the top of the fixed encoder as the supervised evaluator. **1-Nearest Neighbor (1-NN)**: as adopted in [5], the test label is assigned from its top nearest neighbour in the training samples in a non-parametric fashion, where the representation similarity is measured by cosine distance. Unlike the supervised linear classifier, 1-NN is used as an unsupervised evaluator that does not require extra training to assign the label.

4.2. Recognition Comparisons with the SOTAs

We first compare the proposed FoCoViL with the state-of-the-art (SOTA) unsupervised approaches based on 3D skeleton, including regression-based repre-

Table 2: Performance comparisons (%) on NTU RGB+D 60 with supervised (Linear) and unsupervised (1-NN) evaluation protocols.

Method	NTU RGB+D 60
Linear Classifier	
LongT GAN [4]	48.1
SeBiReNet [6]	79.7
AS-CAL [38]	64.8
ST-CL [39]	69.4
MCAE-MP [37]	74.7
CRRL [53]	73.8
TSL [54]	76.3
CrosSCLR [23]	76.4
3s-CrosSCLR [23]	83.4
AimCLR [8]	79.7
3s-AimCLR [8]	83.8
3s-SCC [7]	83.1
ISC [52]	**85.2**
FoCoViL	83.2
1-Nearest Neighbour (1-NN) Classifier	
P&C [5]	76.1
CrosSCLR [23]	63.5
3s-CrosSCLR [23]	65.2
AimCLR [8]	70.1
3s-AimCLR [8]	69.3
CRRL [53]	75.2
MCAE-MP [37]	**82.4**
CoViL	79.4
FoCoViL	80.2

Table 3: Comparison of model complexity based on NTU RGB+D 60.

Method	CRRL [53]	3s-CrosSCLR [23]	3s-AimCLR [8]	FoCoViL
Inference Time	-	**0.69ms**	0.70ms	1.31ms
#params	7.4M	2.51M	2.51M	**1.62M**

Table 4: Performance comparisons (%) on NTU RGB+D 120 with supervised (Linear) and unsupervised (1-NN) evaluation protocols.

Method	NTU RGB+D 120
Linear Classifier	
AS-CAL [38]	48.6
MCAE-MP [37]	52.8
CRRL [53]	56.2
TSL [54]	59.1
ISC [52]	67.1
CrosSCLR [23]	67.1
3s-CrosSCLR [23]	67.9
AimCLR [8]	63.4
3s-AimCLR [8]	**68.2**
FoCoViL	62.3
1-Nearest Neighbour (1-NN) Classifier	
P&C [5]	39.5
MCAE-MP [37]	42.3
ISC [52]	50.6
CrosSCLR [23]	**52.5**
FoCoViL	51.0

Table 5: Performance comparisons (%) on PKU-MMD with unsupervised (1-NN) evaluation protocol.

Method	Phase 1	Phase 2
P&C [5]	70.4	38.4
CrosSCLR [23]	68.9	21.2
AimCLR [8]	72.0	39.5
FoCoViL	**75.2**	**43.3**

Table 6: Performance comparisons (%) on UWA3D under different training partitions.

Training Views	V1&V2		V1&V3		V1&V4		V2&V3		V2&V4		V3&V4		Average
Testing Views	V3	V4	V2	V4	V2	V3	V1	V4	V1	V3	V1	V2	
AS-CAL [38]	25.1	22.8	21.3	19.7	22.4	25.5	21.6	19.5	23.9	21.1	21.2	19.7	22.0
SeBiReNet [6]	53.9	61.6	54.1	58.6	51.5	52.0	**71.5**	56.0	**72.3**	51.3	**68.9**	51.5	58.6
P&C [5]	**59.9**	63.1	57.1	62.7	58.7	58.3	63.5	58.3	64.3	53.8	66.3	**55.2**	60.1
FoCoViL	58.3	**63.1**	**59.5**	**64.7**	**59.1**	**59.9**	67.5	**61.5**	69.8	**56.7**	67.5	52.8	**61.7**

sentation learning models LongT GAN [4], Denoised-LSTM [33], SeBiReNet [6], P&C [5], MCAE-MP [37], and SCC [7], and contrastive learning-based models MS²L [36], AS-CAL [38], ST-CL [39], CRRL [53], TSL [54], ISC [52], CrosS-CLR [23], and AimCLR [8], where action class labels of all models are not used during training.

As shown in Table 1, our FoCoViL achieves superior recognition results on N-UCLA dataset for both evaluation protocols compared to other SOTA models, including the contrastive-based method CRRL. FoCoViL yields a significant 3.2% accuracy increase over P&C on N-UCLA, which shows that the disparity of the same scene from different viewpoints can greatly affect the classification results. Furthermore, we also achieve consistent improvements from CoViL to FoCoViL, showing that focalization improves contrastive learning with better feature representation.

For NTU RGB+D 60 in Table 2 and NTU RGB+D 120 in Table 4, the overall accuracies are slightly lower than N-UCLA for all methods since the datasets contain highly similar classes such as *drinking water* and *eating meal*, as well as local-scale movements such as *thumb up* and *thumb down*. We first observe that FoCoViL is more advantageous than non-contrastive learning-based approaches like SeBiReNet and 3s-SCC. When comparing with MoCo-based approaches, FoCoViL outperforms CRRL and TSL, and performs comparably with CrosSCLR, AimCLR, and ISC under supervised evaluations. Under the more challenging unsupervised protocol, FoCoViL outperforms ISC under NTU RGB+D 120, and outperforms both single- and multi-stream CrosSCLR and AimCLR under NTU RGB+D 60 with over 10% performance improvement. This shows that CrosSCLR and AimCLR heavily rely on the supervised classifier.

As shown in Table 3, we also test the computational resources by comparing the processing time for each action clip and the number of parameters used for [23, 8] and FoCoViL, where FoCoViL has comparable classification performance but much fewer parameters. In general, FoCoViL is less spatially complex compared to other contrastive learning-based methods, specifically 3s-CrosSCLR [23] and 3s-AimCLR [8], since they are multi-stream methods that fuse three skeleton features (*i.e.*, joint, bone, and motion). In terms of the inference time, FoCoViL takes longer. However, since the action sequence is relatively short (usually around 1s for each trial), the model complexity will not be heavily affected by the recurrent times of GRU.

When comparing PKU-MMD in Table 5, FoCoViL also outperforms CrosS-CLR and AimCLR for both phases 1 and 2, where phase 2 is a noisy version dataset with more diversities in terms of facing directions and action performance.

Note that the cross-subject evaluations are conducted on PKU-MMD, where three viewpoints are used for training that validates the feasibility of our method under multiple viewpoints (≥ 2).

For UWA3D in Table 6, we significantly outperform AS-CAL [38] at all partitions of different training and testing viewpoints. Since the UWA3D dataset is challenging with large variations appearing in different viewpoints, the performance may vary in different training combinations. However, on average our method performs better than SeBiReNet and P&C. This is evidenced by having most of the best performances achieved using our approach, which shows the generality of our model across a variety of viewpoints.

We further compare the confusion matrices of P&C and FoCoViL under all types of actions of N-UCLA. In Fig. 3, we observe that five classes reach 100% recognition accuracy in the proposed FoCoViL compared to three in P&C. In addition, FoCoViL discriminates better between the actions like *doffing vs. throw* and *donning vs. carry*. In particular, these two pairs of actions are quite similar in some viewpoints, thereby hard to be distinguished. By correlating different viewpoints, FoCoViL can get a comprehensive understanding of action features from various angles to provide a more discriminative classification for these ambiguous classes.

The confusion matrix comparison on NTU RGB+D 60 is given in Fig. 4. To compare the recognition performance with large numbers of classes, we calculate the difference by subtracting the confusion matrix of the most recent method AimCLR from FoCoViL. From Fig. 4(c), we observe that most of the diagonal values are positive values (red), and there are lots of negative values (blue) that appear in the off-diagonal part, which indicates that compared to FoCoViL, AimCLR is more likely to have non-zero values that are misclassified as other classes.

*4.3. Latent Space Evaluation*

4.3.1. Purity & ARI

Following [57], we also test two common metrics, Purity and Adjusted Rand Index (ARI), to *quantitatively* evaluate the quality of our learned latent space. *Purity* measures to what extent samples in a cluster belong to the true class:

$$Purity = \frac{1}{|X|} \sum_k \max_l \omega_{kl} \tag{8}$$

where $|X|$ is the total number of test samples. ω_{kl} is the number of samples in the kth predicted cluster that belongs to the lth ground-truth class. *ARI* measures the
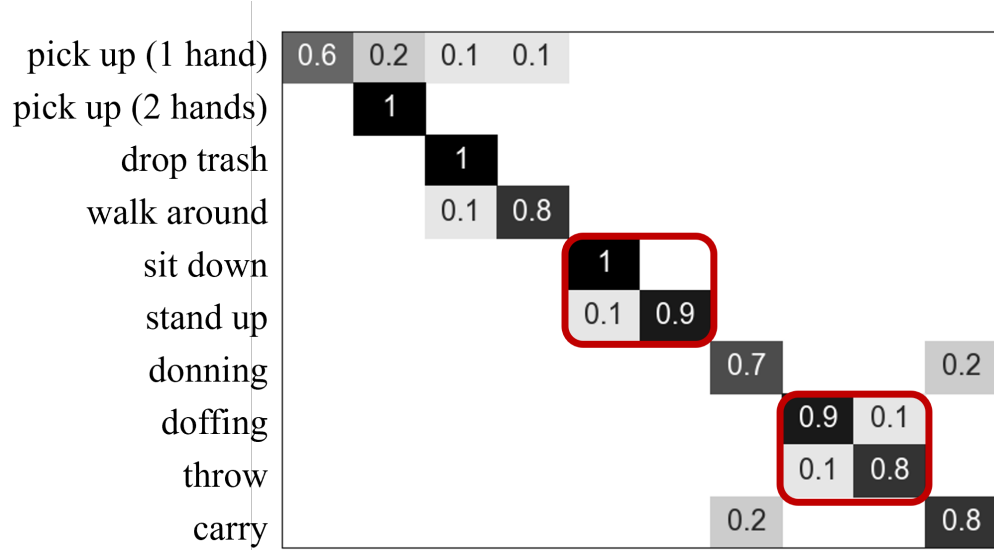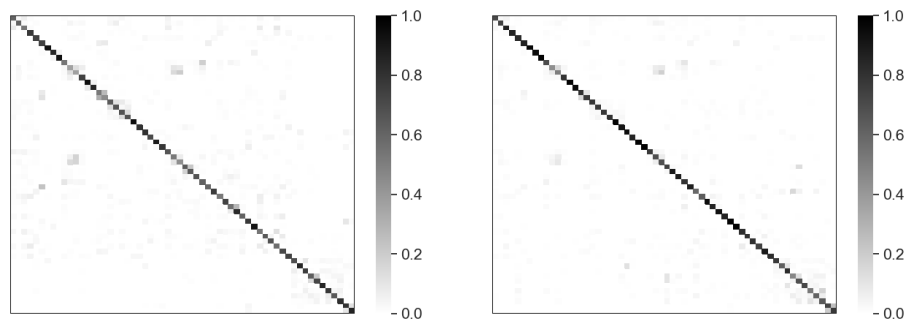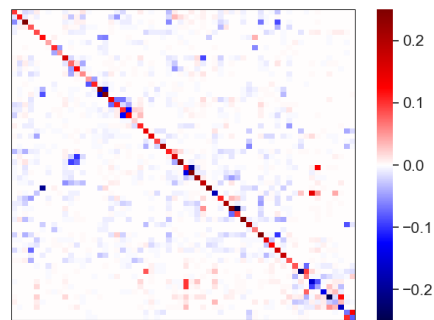
(a) P&C [5]
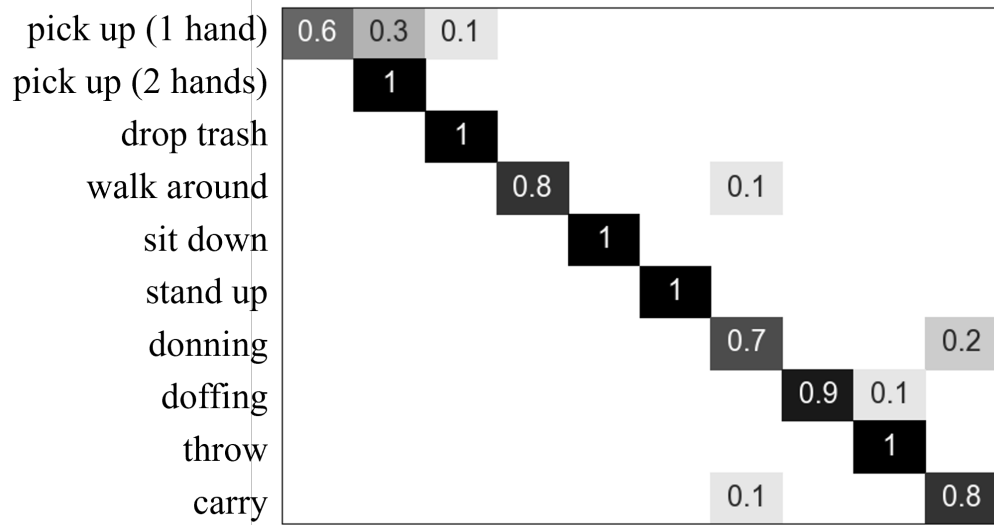
(b) FoCoViL

Figure 3: Confusion matrices on N-UCLA.

(a) AimCLR

(b) FoCoViL

(c) FoCoViL-AimCLR

Figure 4: Confusion matrices on NTU RGB+D 60 on all 60 classes. (a) and (b) represent the confusion matrices of AimCLR and our FoCoViL, respectively, and (c) is the difference confusion matrix between the two methods, *i.e.* (b)-(a). In (c), The diagonal value in red and the off-diagonal element in blue indicate the better classifications from FoCoViL.

Table 7: Quantitative evaluation of clustering quality. For both Purity and ARI, the higher value indicates better clustering.

Dataset	Method	Purity		ARI	
		GMM	K-Means	GMM	K-Means
N-UCLA	SeBiReNet [6]	0.513	0.527	0.280	0.299
	P&C [5]	0.512	0.592	0.260	0.412
	FoCoViL	**0.605**	**0.618**	**0.412**	**0.478**
NTU RGB+D 60	SeBiReNet [6]	0.131	0.125	0.071	0.053
	P&C [5]	0.246	0.249	0.129	0.137
	FoCoViL	**0.294**	**0.311**	**0.170**	**0.172**
PKU-MMD	P&C [5]	0.418	0.409	0.237	0.237
	FoCoViL	**0.483**	**0.501**	**0.329**	**0.350**
UWA3D	P&C [5]	0.405	0.445	0.172	0.221
	FoCoViL	**0.469**	**0.485**	**0.255**	**0.272**

correctness of classification concerning the mutual information between clusters:

$$ARI = \frac{\sum_{kl}\binom{\omega_{kl}}{2} - (\sum_{k}\binom{\omega_k}{2}\sum_{l}\binom{\omega_l}{2})/\binom{|X|}{2}}{\frac{1}{2}(\sum_{k}\binom{\omega_k}{2} + \sum_{l}\binom{\omega_l}{2}) - (\sum_{k}\binom{\omega_k}{2}\sum_{l}\binom{\omega_l}{2})/\binom{|X|}{2}} \tag{9}$$

where $\omega_k = \sum_l \omega_{kl}$, $\omega_l = \sum_k \omega_{kl}$ is the number of samples in the kth cluster or the lth class, respectively. Both measurements reveal the quality of clustering from different aspects with the maximum value of 1 if each sample gets its cluster. In addition to the 1-nearest neighbour, we adopt two common unsupervised clustering methods, Gaussian Mixture Model (GMM) and K-Means, on the spanned latent space. The number of clusters is set the same as the number of real classes. The corresponding results are compared with SeBiReNet [6] and P&C [5] presented in Table 7. It is worth noting that in all the compared datasets, we achieve the highest scores with significant improvements on both clustering metrics.

4.3.2. Visualization

We further visualize t-SNE of the learned features in Fig. 5 and 6 to *qualitatively* compare the latent space. The results clearly show that compared to P&C, CrosSCLR, and AimCLR, FoCoViL generates clusters with less overlapping since it learns a more sparse and discriminative latent space by generally enlarging the distance between negative samples. Meanwhile, the representation is more compact within each cluster. This highlights that the system can better
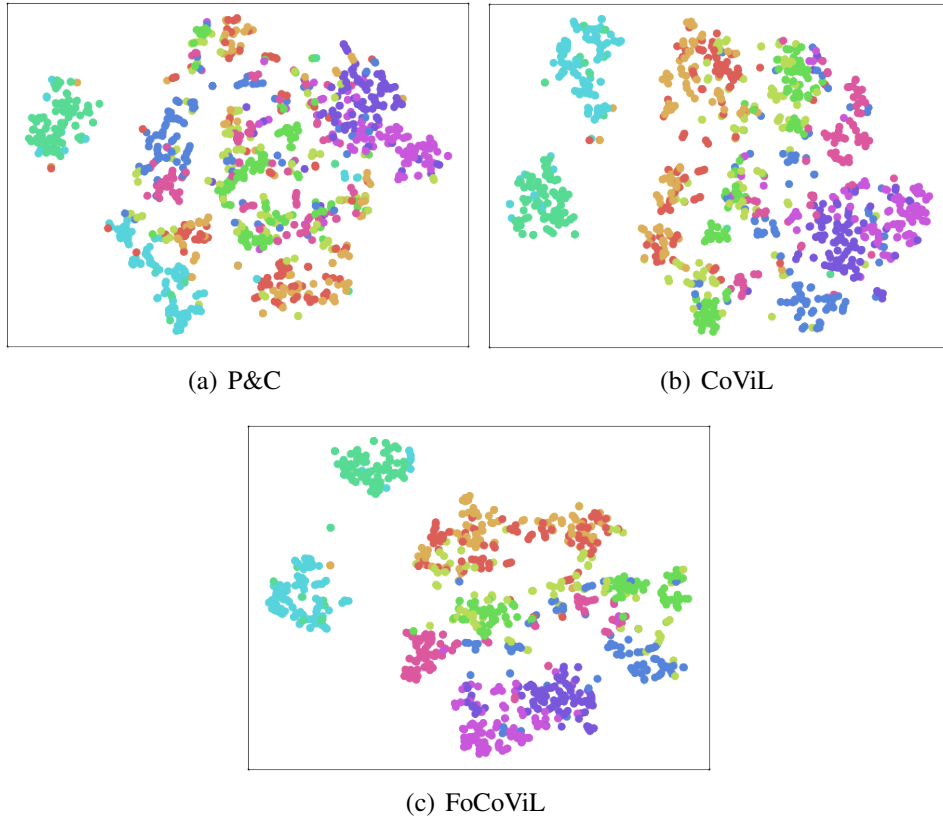
(a) P&C

(b) CoViL

(c) FoCoViL

Figure 5: T-SNE comparisons of P&C, our CoViL, and FoCoViL on 10 classes of N-UCLA.

group the actions with common properties by removing the view interference. We also visualize the effectiveness of focalization by comparing (b) and (c) of the two figures, where the latent space is improved by having a clearer margin between clusters.

4.4. Ablation Study

4.4.1. Network Structure

We also verify the effectiveness of the main components in our network structure. The ablation results based on the N-UCLA dataset are provided in Table 8. In general, the designed modules consistently improve the performance from the reconstruction baseline (*i.e.* Reconst.). We first observe that the view alignment (denoted as Ali.) can largely increase the recognition accuracy (Reconst. *vs.* Ali. reconst.). Then, by adding back the fine-level multi-view contrastive loss, the

(a) CrosSCLR

(b) AimCLR
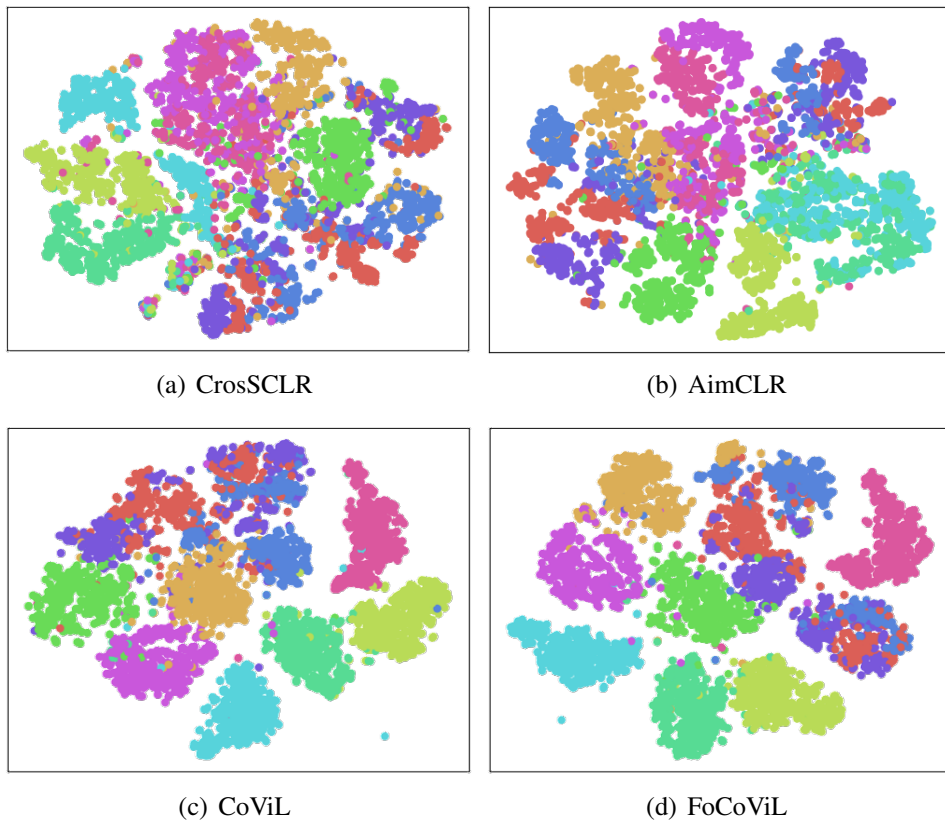
(c) CoViL

(d) FoCoViL

Figure 6: T-SNE comparisons of CrosSCLR, AimCLR, our CoViL, and FoCoViL on 10 selected categories of NTU RGB+D 60.

Table 8: The ablation tests of recognition accuracy and purity (GMM) on the proposed FoCoViL.

Method	Accuracy (%)	Purity
Reconst.	74.4	0.413
Ali. reconst.	83.8	0.457
CoViL w/o g	84.0	0.531
CoViL w/o "+"	84.4	0.488
CoViL w/o "−"	85.1	0.533
CoViL	86.7	0.569
FoCoViL	**88.3**	**0.605**

recognition performance improves by 2.9% (Ali. reconst. *vs.* CoViL), showing that the discrimination power is increased by refining the view-invariant representation. There is also a significant improvement in purity score (0.457 *vs.* 0.569), proving that CoViL contributes a lot to shaping the clustering space by modeling the mutual distance between samples. Finally, by adding the focalization, FoCoViL converges to a better latent space with a clearer distribution of clusters, thus further boosting the purity score from 0.569 to 0.605.

As in Table 8, we observe a large improvement in Purity by comparing FoCoViL (0.457 *vs.* 0.605) and singly considering view alignment (0.413 *vs.* 0.457). The performance gain indicates that FoCoViL contributes more to shaping the representation space to boost the clustering compared to view alignment. Although view alignment is necessary for aligning the facing directions, the resulted space is still view-dependent with large motion variations because of the view-specific self-occlusions (see Fig. 1), which explains why it is necessary to learn a view-invariant space after view alignment. By constraining the mutual distances between the same or different scenes, FoCoViL learns a better representation space with clearer cluster distributions to improve the recognition.

In addition to the main structures in FoCoViL, we also evaluate the sub-structures of the projection net g, CoViL w/o "+" (i.e. only including negative pairs regardless of viewpoints), and CoViL w/o "−" (i.e. only including positive pairs by supplementing the anchor samples from different views). We first notice a large performance boost on both metrics by including g when comparing CoViL w/o g and CoViL. This extensively reflects the importance of the projection net to contrastive learning. Then we find that increasing the agreement of "+" pairs (CoViL w/o "+" *vs.* CoViL) and the disagreement of "−" pairs (CoViL w/o "−" *vs.* CoViL) are both essential to CoViL for a robust clustering ability, as the two factors complement each other with "+" pairs generating compact clusters by cor-

Table 9: Recognition accuracy under different structures of g.

	#input→#FC1→#FC2	Accuracy (%)
1 Layer	1024→1024	85.1
2 Layers	1024→256→1024	85.7
	1024→512→1024	**88.3**
	1024→1024→1024	87.3
	1024→2048→1024	87.3

Table 10: Evaluation of training sample size.

Data Proportion	*10%*	*50%*	*70%*	*100%*
Acc. (%)	77.3	84.9	85.3	88.3

relating the same scene together, where "−" pairs enable a sparse distribution to avoid an over dense representation space.

4.4.2. Projection Net Configuration

As a key component of FoCoViL, we also conduct a detailed evaluation of the projection net g. We test the recognition performance under different combinations of layer and unit in g as given in Table 9. Note that the vector dimension is 1024 for both the encoder and the decoder output. The results show that using 2 fully-connected layers by first going through a squeeze operation (1024→512) in the first layer, and following an excitation operation (512→1024) in the second layer will broadcast a more powerful structure to the projection net.

4.5. The Impact of Training Size

In Table 10, we show how the sample size would affect the model performance by selecting 10%, 50%, 70%, and 100% (the entire training split) of data for training. When training on a small subset (10%) of data, the performance already reaches 77.3% with most of the actions being recognized correctly. When increasing the training proportion, the model observes more sample patterns that benefit the contrastive learning.

5. Conclusion

We propose FoCoViL for cross-view self-supervised skeleton-based action recognition in this work. By maximizing the mutual information of the multi-view actions, FoCoViL better clusters the actions with common properties in

the latent space. This is done by contrasting pairwise similarity of latent representations under the same and different scenes to refine the space with high-level view-invariant features. An adaptive focalization on the contrasted sample pairs further converges FoCoViL to a more discriminative latent space with fewer misclassifications. Experiments on five benchmark 3D datasets demonstrate that our method achieves state-of-the-art recognition performance with a high-quality view-invariant space for action clustering, which has more generalization benefits. The performance also demonstrates the compatibility of FoCoViL with different scales of data size.

Other than multi-view features, contrastive learning with focalization is also extensible to other modalities to improve the sample representation learned by contrastive loss, such as contrasting between color, depth, or textual features [49, 58]. Another future direction is that at the focalization stage, it is of interest to explore the imbalanced similarity of the negative pairs as well to further improve the classification performance.

In this work, we use an RNN-based auto-encoder to learn motion dynamics. However, the proposed focalized contrastive learning is also feasible to convolutional-based auto-encoder backbones, such as Residual 3D Convolutions (R3D) [24] or Pseudo-3D Residual network (P3D) [59] that are usually adopted in RGB-based action recognition tasks, to detect more visual variations appearing in the action images.

Acknowledgment

References

[1] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[2] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, N. Zheng, View adaptive neural networks for high performance skeleton-based human action recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (8) (2019) 1963–1978.

[3] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, N. Zheng, Semantics-guided neural networks for efficient skeleton-based human action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 1112–1121.

[4] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, Z. Gong, Unsupervised representation learning with long-term dynamics for skeleton based action recognition, in: Thirty-Second AAAI conference on Artificial Intelligence, 2018.

[5] K. Su, X. Liu, E. Shlizerman, Predict & cluster: Unsupervised skeleton based action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 9631–9640.

[6] Q. Nie, Z. Liu, Y. Liu, Unsupervised 3d human pose representation with viewpoint and pose disentanglement, in: European Conference on Computer Vision, 2020, pp. 102–118.

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

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

[9] M. Liu, H. Liu, C. Chen, Enhanced skeleton visualization for view invariant human action recognition, Pattern Recognition 68 (2017) 346–362.

[10] I. Lee, D. Kim, S. Kang, S. Lee, Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks, in: IEEE International Conference on Computer Vision, 2017, pp. 1012–1020.

[11] Q. Nie, J. Wang, X. Wang, Y. Liu, View-invariant human action recognition based on a 3d bio-constrained skeleton model, IEEE Transactions on Image Processing 28 (8) (2019) 3959–3972.

[12] X. Liu, Y. Li, R. Xia, Adaptive multi-view graph convolutional networks for skeleton-based action recognition, Neurocomputing 444 (2021) 288–300.

[13] J. Li, Y. Wong, Q. Zhao, M. S. Kankanhalli, Unsupervised learning of view-invariant action representations, Advances in Neural Information Processing Systems 31 (2018) 1254–1264.

[14] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, 2020, pp. 1597–1607.

[15] W. Dai, K. Ng, K. Severson, W. Huang, F. Anderson, C. Stultz, Generative oversampling with a contrastive variational autoencoder, in: 2019 IEEE International Conference on Data Mining (ICDM), 2019, pp. 101–109.

[16] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, D. Larlus, Hard negative mixing for contrastive learning, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 21798–21809.

[17] B. Kang, Y. Li, Z. Yuan, J. Feng, Exploring balanced feature spaces for representation learning, in: International Conference on Learning Representations, 2021.

[18] J. Wang, X. Nie, Y. Xia, Y. Wu, S.-C. Zhu, Cross-view action modeling, learning and recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2649–2656.

[19] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1010–1019.

[20] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, A. C. Kot, Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding, IEEE transactions on pattern analysis and machine intelligence 42 (10) (2019) 2684–2701.

[21] H. Rahmani, A. Mahmood, D. Q. Huynh, A. Mian, Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition, in: European Conference on Computer Vision, 2014, pp. 742–757.

[22] L. Chunhui, H. Yueyu, L. Yanghao, S. Sijie, L. Jiaying, Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding, ACM Multimedia workshop (2017).

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

[24] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A closer look at spatiotemporal convolutions for action recognition, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459.

[25] A. Piergiovanni, A. Angelova, M. S. Ryoo, Evolving losses for unsupervised video representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 133–142.

[26] J. Butepage, M. J. Black, D. Kragic, H. Kjellstrom, Deep representation learning for human motion prediction and classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6158–6166.

[27] H. Wang, E. S. Ho, H. P. Shum, Z. Zhu, Spatio-temporal manifold learning for human motions via long-horizon modeling, IEEE Transactions on Visualization and Computer Graphics (2019).

[28] Q. Men, E. S. Ho, H. P. Shum, H. Leung, A quadruple diffusion convolutional recurrent network for human motion prediction, IEEE transactions on circuits and systems for video technology 31 (9) (2020) 3417–3432.

[29] D. Holden, J. Saito, T. Komura, T. Joyce, Learning motion manifolds with convolutional autoencoders, in: SIGGRAPH Asia 2015 Technical Briefs, 2015, pp. 1–4.

[30] Z. Lai, W. Xie, Self-supervised learning for video correspondence flow, in: British Machine Vision Conference, 2019.

[31] T. Han, W. Xie, A. Zisserman, Self-supervised co-training for video representation learning, Advances in Neural Information Processing Systems 33 (2020) 5679–5690.

[32] J. Wang, Y. Gao, K. Li, Y. Lin, A. J. Ma, H. Cheng, P. Peng, F. Huang, R. Ji, X. Sun, Removing the background by adding the background: Towards

background robust self-supervised video representation learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 11804–11813.

[33] G. G. Demisse, K. Papadopoulos, D. Aouada, B. Ottersten, Pose encoding for robust skeleton-based action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 188–194.

[34] J. N. Kundu, M. Gor, P. K. Uppala, V. B. Radhakrishnan, Unsupervised feature learning of human actions as trajectories in pose embedding manifold, in: IEEE winter conference on applications of computer vision, 2019, pp. 1459–1467.

[35] R. Yue, Z. Tian, S. Du, Action recognition based on rgb and skeleton data sets: A survey, Neurocomputing (2022).

[36] L. Lin, S. Song, W. Yang, J. Liu, Ms2l: Multi-task self-supervised learning for skeleton based action recognition, in: ACM International Conference on Multimedia, 2020, pp. 2490–2498.

[37] Z. Xu, X. Shen, Y. Wong, M. S. Kankanhalli, Unsupervised motion representation learning with capsule autoencoders, Advances in Neural Information Processing Systems 34 (2021) 3205–3217.

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

[39] X. Gao, Y. Yang, S. Du, Contrastive self-supervised learning for skeleton action recognition, in: NeurIPS Workshop on Pre-registration in Machine Learning, 2021, pp. 51–61.

[40] C. Bian, W. Feng, S. Wang, Self-supervised representation learning for skeleton-based group activity recognition, in: ACM International Conference on Multimedia, 2022, pp. 5990–5998.

[41] Y. Su, G. Lin, Q. Wu, Self-supervised 3d skeleton action representation learning with motion consistency and continuity, in: Proceedings of the IEEE international conference on computer vision, 2021, pp. 13328–13338.

[42] C. Si, X. Nie, W. Wang, L. Wang, T. Tan, J. Feng, Adversarial self-supervised learning for semi-supervised 3d action recognition, in: European Conference on Computer Vision, 2020, pp. 35–51.

[43] D. Wang, W. Ouyang, W. Li, D. Xu, Dividing and aggregating network for multi-view action recognition, in: European Conference on Computer Vision, 2018, pp. 451–467.

[44] L. Xia, C.-C. Chen, J. K. Aggarwal, View invariant human action recognition using histograms of 3d joints, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 20–27.

[45] G. Paoletti, J. Cavazza, C. Beyan, A. Del Bue, Unsupervised human action recognition with skeletal graph laplacian and self-supervised viewpoints invariance, in: British Machine Vision Conference, 2021.

[46] L. Gao, Y. Ji, Y. Yang, H. Shen, Global-local cross-view fisher discrimination for view-invariant action recognition, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 5255–5264.

[47] S. Guan, H. Lu, L. Zhu, G. Fang, Afe-cnn: 3d skeleton-based action recognition with action feature enhancement, Neurocomputing 514 (2022) 256–267.

[48] Z. Wu, Y. Xiong, S. X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3733–3742.

[49] Y. Tian, D. Krishnan, P. Isola, Contrastive multiview coding, arXiv preprint arXiv:1906.05849 (2019).

[50] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[51] Y. Yang, Z. Xu, Rethinking the value of labels for improving class-imbalanced learning, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 19290–19301.

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

[53] P. Wang, J. Wen, C. Si, Y. Qian, L. Wang, Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition, IEEE Transactions on Image Processing 31 (2022) 6224–6238.

[54] A. Ben Tanfous, A. Zerroug, D. Linsley, T. Serre, How and what to learn: Taxonomizing self-supervised learning for 3d action recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2696–2705.

[55] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[56] Z. Yang, Z. Hu, R. Salakhutdinov, T. Berg-Kirkpatrick, Improved variational autoencoders for text modeling using dilated convolutions, in: International Conference on Machine Learning, 2017, pp. 3881–3890.

[57] Q. Nie, Y. Liu, View transfer on human skeleton pose: Automatically disentangle the view-variant and view-invariant information for pose representation learning, International Journal of Computer Vision 129 (1) (2021) 1–22.

[58] P. Hu, X. Peng, H. Zhu, L. Zhen, J. Lin, Learning cross-modal retrieval with noisy labels, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 5403–5413.

[59] Z. Qiu, T. Yao, T. Mei, Learning spatio-temporal representation with pseudo-3d residual networks, in: proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5533–5541.