Skeleton-Contrastive 3D Action Representation Learning

Thoker, F.M.; Doughty, H.; Snoek, C.G.M.

DOI
10.1145/3474085.3475307

Publication date
2021

Document Version
Final published version

Published in
MM '21

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/in-the-netherlands/you-share-we-take-care)

Citation for published version (APA):
Skeleton-Contrastive 3D Action Representation Learning

Fida Mohammad Thoker
University of Amsterdam
f.m.thoker@uva.nl

Hazel Doughty
University of Amsterdam
hazel.doughty@uva.nl

Cees G.M. Snoek
University of Amsterdam
cgmsnoek@uva.nl

ABSTRACT
This paper strives for self-supervised learning of a feature space suitable for skeleton-based action recognition. Our proposal is built upon learning invariances to input skeleton representations and various skeleton augmentations via a noise contrastive estimation. In particular, we propose inter-skeleton contrastive learning, which learns from multiple different input skeleton representations in a cross-contrastive manner. In addition, we contribute several skeleton-specific spatial and temporal augmentations which further encourage the model to learn the spatio-temporal dynamics of skeleton data. By learning similarities between different skeleton representations as well as augmented views of the same sequence, the network is encouraged to learn higher-level semantics of the skeleton data than when only using the augmented views. Our approach achieves state-of-the-art performance for self-supervised learning from skeleton data on the challenging PKU and NTU datasets with multiple downstream tasks, including action recognition, action retrieval and semi-supervised learning. Code is available at https://github.com/fmthoker/skeleton-contrast.

CCS CONCEPTS
• Computing methodologies → Activity recognition.

KEYWORDS
skeleton action recognition; contrastive learning; self-supervision

ACM Reference Format:

1 INTRODUCTION
The goal of this paper is to learn a latent feature space suitable for 3D human action understanding. Different from traditional RGB frames [2, 15], skeleton data consists of 3D coordinates representing the major joints of each person in a video [5, 20, 32]. It offers a light-weight representation that can be processed faster and in a privacy-preserving manner providing application potential in video surveillance, assisted living, gaming and human-computer interaction. Moreover, when compared to RGB, such a representation is robust to changes in background and appearance [23, 46]. However, learning a good feature space for 3D actions requires large amounts of labeled skeleton data [7, 12, 35, 36, 44–46], which is much harder to obtain than large amounts of labeled RGB video. To address this major shortcoming, we propose a new self-supervised contrastive learning method for 3D skeleton data.

Several previous works also considered self-supervised learning for 3D skeleton data [19, 27, 39, 49]. These works design pretext tasks, such as learning to reconstruct masked input [49] and motion prediction [19], which still require the features to represent variations such as the viewpoint and skeleton scale, rather than focusing on higher-level semantic features relevant to downstream tasks. Instead, we take inspiration from recent self-supervised literature for RGB images, which aims to learn the high-level similarity between augmented forms of the same image and the dissimilarity between these and other images [3, 11, 29]. At the core of such contrastive learning is the nature of the RGB data, where each sample contains abundant pixel information, allowing for augmentations like spatial-cropping and color-jittering to easily generate subtly different versions of an image without changing its semantic content. However, skeleton sequences are much more sparse than RGB data and the augmentations commonly applied to images would not change the estimated skeleton of a person. Thus, for contrastive learning with skeleton sequences, we need skeleton-specific augmentations to encourage the learned features to encode information relating to spatio-temporal dynamics of the joints. We also want to enrich the input space which can be sampled from, to increase the variety of samples with the same semantic content, and thus increase the difficulty of the contrastive learning task.

We make three contributions. Our first contribution is to leverage multiple input-representations of the 3D-skeleton sequences. In...
particularly, we propose inter-skeleton contrast to learn from a pair of skeleton-representations in a cross-contrastive fashion, see Figure 1. This allows us to enrich the sparse input space and focus on the high-level semantics of the skeleton data rather than the nuances of one specific input representation. Second, we introduce several skeleton-specific spatial and temporal augmentations for generating positive pairs which encourage the model to focus on the spatio-temporal dynamics of skeleton-based action sequences, ignoring confounding factors such as viewpoint and the exact joint positions. Finally, we provide a comprehensive evaluation of our learned feature space on various challenging downstream tasks, showing considerable improvement over prior methods in all tasks.

2 RELATED WORK

Self-Supervised Learning. Self-supervised learning aims to learn feature representations without human annotation, typically by solving pretext tasks which exploit the structure of unlabeled data. Previous works have proposed a variety of such tasks for learning image representations, e.g., solving spatial jigsaw puzzles [28], rotation prediction [9], spatial context-prediction [6], image inpainting [30] and colorization [47, 48]. Similarly, pre-text tasks have been designed for learning video representations, such as spatio-temporal puzzles [14], prediction of frame-order [8], clip-order [43], speed [1], future [10] and temporal coherence [16]. Such pretext tasks rely on the rich structured nature of RGB data with the hope that by learning to solve these tasks the encoded features will rely on the high-level semantics of the image or video and are thus applicable to the downstream task(s). Unfortunately, these existing RGB-based pretext tasks are not suited for 3D-skeleton sequences which have a simple structure and are less rich in information.

Instead of designing specific pretext tasks, recent self-supervised methods rely on instance discrimination and learn the similarity between sample pairs [3, 11, 26, 29, 40]. A noise contrastive loss learns invariances to certain image or video transformation functions, resulting in good feature representations. For example, Chen et al. [3] show that learning invariance to simple image augmentations, such as color jitter, results in highly discriminative features. He et al. [11] propose a momentum contrast which is able to utilize a large number of negatives for the noise contrast by storing image features from previous batches in a dynamic queue. In this paper, we rely on contrastive estimation for 3D action representation learning. As existing works use augmentations specific to RGB images, we introduce three skeleton-specific augmentations to generate positive pairs for learning the spatio-temporal dynamics of 3D-skeleton sequences. Furthermore, we propose inter-skeleton contrastive learning which additionally aims to learn invariance to the particular input representation of the 3D-skeleton sequences.

Supervised 3D Action Recognition. Numerous methods for supervised 3D action recognition exist. While earlier methods design handcrafted features [25, 41, 42] to model geometric relationships between skeleton joints, recent approaches rely on data-driven deep neural networks. Three skeleton-representations have become popular for deep learning. Sequence-based treats the 3D-skeleton data as a multi-dimensional time-series and models it with a recurrent architecture [21, 22, 32, 35, 46] to learn the temporal dynamics of the joints. Image-based create a pseudo-image representation of the 3D-skeleton data [7, 12, 17, 23, 38] which is encoded by CNN architectures to model the co-occurrence of multiple joints and their motion. Finally, graph-based [4, 13, 18, 24, 31, 33, 37, 44] represents the 3D-skeleton data with a graph consisting of spatial and temporal edges. Graph-convolutional architectures then encode the spatio-temporal motion from the human skeleton graph. Although these methods achieve excellent performance, they are all fully supervised and require time-consuming action class annotations. We propose a self-supervised method for 3D-skeleton data that leverages the diversity of the skeleton-representations to learn highly discriminative features from unlabeled data.

Self-Supervised 3D Action Recognition. Overcoming the need for large amounts of annotations has only recently received attention in the 3D action recognition community. Zheng et al. [49] propose a seq2seq model that learns to reconstruct masked input 3D-skeleton sequences. In particular, a GAN is trained such that the decoder attempts to regenerate the input sequences, while a discriminator measures the quality of the regenerated sequences. Similarly, Nie et al. [27] propose a cross-view reconstruction task that relies on a siamese denoising autoencoder to reconstruct the correct version of corrupted and rotated input skeletons. Su et al. [39] also propose a seq2seq model that regenerates input skeleton sequences. To encourage the encoder to learn better latent features, the decoder is weakened by fixing its weights.

Lin et al. [19] take a different approach and propose multi-task self-supervised learning for the sequence-based skeleton representation. Their framework solves multiple pretext tasks simultaneously, such as motion prediction and skeleton-jigsaw. Si et al. [34] propose an adversarial self-supervised learning approach that couples the self-supervised learning and the semi-supervised scheme via neighbor relation exploration and adversarial learning.

Different from all these works, we do not rely exclusively on a sequence-based skeleton-representation and pretext tasks such as input-reconstruction and motion prediction. Instead, we propose to exploit the diversity of skeleton-representations in an inter-contrastive learning regime and design skeleton-specific spatial and temporal augmentations for use in this contrastive method.

3 SKELETON-CONTRASTIVE LEARNING

In this section we present our inter-skeleton contrast approach for self-supervised learning of 3D action features. Contrastive methods aim to learn a good feature space by learning the similarity between augmented views of the same data. Since augmentations in existing contrastive learning works are primarily designed for RGB images [3] they are not suitable for the skeleton data that is considered in this work. Therefore, we first propose several skeleton-specific augmentation functions in Section 3.1. These augmentations enable us to apply existing contrastive learning methods, such as MoCo [11], to skeleton data. We describe this is Section 3.2.

However, contrastive learning can be vulnerable to shortcuts, where simple features, irrelevant to the downstream task, may be enough to identify the different augmented views of the same data. For instance, Chen et al. [3] show that color distributions can be a shortcut to identify different crops from the same image. To avoid such shortcuts and make the contrastive learning task more difficult, we additionally contrast pairs of different input
skeleton representations with each other. We call this inter-skeleton contrastive learning and detail our approach in Section 3.3.

3.1 Skeleton Augmentations

The goal of contrastive learning is to learn the semantic similarity between items in a dataset without labels. This is usually done by learning the similarity of two augmented views (positive pairs) of a sample $X$. A data augmentation function $D$, composed of a single or multiple transformations, creates the augmented views. Hence, the network learns features for $X$, which are invariant to the transformations in $D$. The nature of the data $X$ and the downstream task determines the appropriate invariances that the learned features should possess. In our case, $X$ is a 3D-skeleton sequence, where each sequence represents a particular spatial configuration of human joints and its motion over a short period of time. Thus, to learn useful representations for 3D-skeleton data, the commonly used RGB augmentations, such as color-distortion and Gaussian blurring [3], are not suitable. Instead, we need to learn invariances to transformations that encode the spatial and temporal dynamics of 3D skeleton action sequences. We introduce multiple spatial and temporal skeleton augmentation techniques to generate positive pairs for 3D-skeleton action sequences: Pose Augmentation, Joint Jittering and Temporal Crop-Resize. We then combine these to create our final spatio-temporal skeleton augmentation. Let us assume each raw action sequence $X \in \mathbb{R}^{T \times J \times 3}$ consists of 3D coordinates of $J$ body joints in $T$ consecutive video frames. We define our individual augmentations $D$ based on $X$.

3.1.1 Spatial Skeleton Augmentations. To apply our learned feature space to downstream tasks such as 3D action recognition, we require the feature encodings to rely on more discriminatory spatial semantics like joint configurations, while being invariant to factors such as viewpoint, camera distance, skeleton scale and joint perturbations. Existing augmentations for RGB images would not achieve this, thus we propose two new skeleton-specific spatial augmentations: pose augmentation and joint jittering. These can be applied to each of the $T$ skeletons in the sequence $X$ so a contrastive learning framework can learn invariance to these augmentations.

Pose Augmentation. With this transformation, we aim to create positive pairs which differ in viewpoint and distance to the camera, while retaining the same pose from the original sequence. To achieve this, we apply a 3D shear on the action sequence $X$:

$$D_{\text{Spatial}_{1}}(X) = X \cdot \begin{bmatrix} 1 & r_{01} & r_{02} \\ r_{10} & 1 & r_{12} \\ r_{20} & r_{21} & 1 \end{bmatrix},$$  \hspace{1cm} (1)$$

where the elements of the augmentation matrix are randomly drawn from a uniform distribution $[-1, 1]$. Figure 2 shows several examples. By applying the same shearing operation to each joint of the skeleton at each time-step in the sequence we are able to simulate changes in camera viewpoint and distance between the subject and camera. Therefore, a contrastive network which learns invariance to this transformation is forced to learn more discriminatory pose semantics of the positive pairs and ignore redundant information such as the viewpoint and proximity to the camera.

Joint Jittering. We also want a contrastive method to be invariant to noise in the estimated skeleton. Therefore we propose joint jittering to create positive pairs where some of the joint connections in $X$ are randomly perturbed. We select $j$ of the $J$ joints at random and move these joints to irregular positions, while keeping other joints in their original position. The transformation is defined as:

$$D_{\text{Spatial}_{2}}(X) = X[:, j] \cdot \begin{bmatrix} r_{00} & r_{01} & r_{02} \\ r_{10} & r_{11} & r_{12} \\ r_{20} & r_{21} & r_{22} \end{bmatrix},$$  \hspace{1cm} (2)$$

where $j$ is a subset of the joints such that $|j| < J$, and the elements of the jitter matrix are randomly drawn from a uniform distribution $[-1, 1]$. The same jitter matrix is applied to each joint in $j$ at each time-step $T$. Examples are shown in Figure 3. To learn invariance to such transformations, the contrastive task is encouraged to rely on the spatio-temporal semantics of the common joint connections and ignore the noise from the irregular joint connections.

3.1.2 Temporal Skeleton Augmentation. Besides the spatial perturbations, a good 3D skeleton feature space should also be robust to temporal modifications of the skeleton sequences, such as the speed of an action and changes to the temporal bounds of the sequence. To this end, we propose temporal crop-resize.

![Figure 2: Spatial pose augmentation examples. A shear operation is applied to the original action so that the augmented pairs differ in viewpoint and camera distance.](image1)

![Figure 3: Spatial joint jittering examples. The augmented pairs contain a subset of common joint connections while other joint connections are randomly moved to an irregular position.](image2)
As we will show in the experiments, learning invariance to spatio-temporal transformations is crucial for contrastive learning. In this transformation, we create positive pairs with varying speed and varying starting and ending points. We sample different parts of the action sequence \( X \) via a random crop and resize this crop over the temporal dimension \( T \):

\[
D_{\text{Temporal}}(X) = \text{Interpolate}(X[L_{\text{start}} : L_{\text{start}} + TL_{\text{ratio}}]).
\]  

The length ratio \( L_{\text{ratio}} \) is first randomly sampled from distribution \([\text{min}, 1.0]\), followed by randomly selecting a starting frame \( L_{\text{start}} \) between \((0, T - TL_{\text{ratio}})\). The sub-sequence \( X[L_{\text{start}} : L_{\text{start}} + TL_{\text{ratio}}] \) is then re-sampled to a fixed length. This re-sampling causes the temporal crop-resize to also alter the speed of a sequence as well as its start and end times; a shorter sub-sequence will effectively have a slower speed once re-sampled. Figure 4 shows examples of this transformation. By including this augmentation the contrastive task is forced to focus on the commonalities of the joint motion dynamics over the sampled temporal periods and be robust to changes in the exact start, end and speed of an action.

**3.1.3 Spatio-Temporal Skeleton Augmentations.** To learn the spatio-temporal dynamics of the skeleton sequences, we propose to combine the above spatial and temporal transformations into a single augmentation function. Such composition results in strong positive pairs which vary in both spatial and temporal dynamics locally, while retaining the high-level semantics of the original action sequence. In particular, we first apply the temporal crop-resize augmentation \( D_{\text{Temporal}} \) on the original action sequence \( X \) followed by a spatial augmentation \( D_{\text{Spatial}} \) to the resulting sequence:

\[
D_{\text{Spatial-Temporal}}(X) = D_{\text{Spatial}}(D_{\text{Temporal}}(X)).
\]  

Here, \( i \) can either be fixed to the pose augmentation or the joint jitter or randomized to select either of the spatial augmentations. As we will show in the experiments, learning invariance to spatio-temporal transformations produces a better 3D action feature space and randomizing the composition further improves the result.

**3.2 Intra-Skeleton Contrast**

Before describing our proposed inter-skeleton method, we first describe how the above augmentations can be incorporated into an existing contrastive method, such as MoCo [11], with a single input skeleton-representation. We call this intra-skeleton contrastive learning. Each raw action sequence \( X \in \mathbb{R}^{T \times J \times 3} \) is first augmented into two different views \( X_q \) and \( X_k \) (called query and key) via a data augmentation function \( D \). Both views of the skeleton data are then instantiated into the same skeleton-representation, be it image-based or sequence-based or graph-based. A contrastive method such as MoCo uses two encoders, one for the query and one for the key. We refer to the query encoder as \( f_q \) and the key encoder as \( f_k \). Let \((Z_q, Z_k) = (f_q(X_q), f_k(X_k))\) be output embeddings of the encoders for the input query-key pair. We then train the contrastive network using the noise contrastive estimation loss InfoNCE [29]:

\[
\mathcal{L}(X) = -\log \frac{\exp(Z_q \cdot Z_k / \tau)}{\exp(Z_q \cdot Z_k / \tau) + \sum_{n \neq k} \exp(Z_q \cdot Z_n / \tau)},
\]  

where \( \tau \) is a temperature softening hyper-parameter and \( N \) is the current set of negatives that are stored in a dynamic queue via previous states of the key encoder \( f_k \) as in [11]. Only the query encoder is actively trained using Equation (5) and the key encoder is updated as a moving average of the query encoder. This trains the framework to learn 3D action features which are invariant to the transformations in \( D \) for the chosen skeleton-representation.

**3.3 Inter-Skeleton Contrast**

Up to this point, our method, like previous contrastive learning approaches [3, 11, 26], learns the similarity between different augmented forms of the same input. We now extend contrastive learning for 3D skeleton data beyond these augmentations and propose inter-skeleton contrast which aims to learn invariance to the input representation of the skeleton sequence. Three 3D-skeleton representations are common: image-based as a \( T \times J \) pseudo-image where the 3D coordinates of each joint are the image channels, sequence-based as a multi-dimensional time series, or graph-based as a spatio-temporal graph. Each requires a different network architecture and encodes different characteristics of the sequence. For example, RNNs treat skeleton sequences as a time series and explicitly model the temporal evolution of joints, while GCNs treat sequences as a graph with both spatial and temporal edges and thus explicitly encode human pose as well as each joint’s temporal motion. While the action depicted by the skeleton sequence is the same, the way the input sequence is represented and encoded is different. To learn invariance to the input representation the contrastive framework has to learn the similarities between the characteristics of these different representations as well as our data augmentations which will result in more discriminative features.

The overall network is depicted in Figure 5. The raw skeleton sequence is first augmented into two views as in Section 3.2. Each view is then represented in two ways, in this case with a graph-based representation and a sequence-based representation. We refer to the different representations of the raw action sequence \( X \) as \( X_{\text{IMG}} \), \( X_{\text{SEQ}} \) for seq-based and \( X_{\text{STG}} \) for graph-based. For the rest of this section we will take the example of the pair \( X_{\text{SEQ}} \) and \( X_{\text{STG}} \) as displayed in Figure 5. We adapt our model to contrast the different input representations by using a pair of momentum contrastive models together, one for each input-representation \( X_{\text{SEQ}} \) and \( X_{\text{STG}} \).
and $X^{STG}$. In particular, the model now consists of two query encoders $f_q^{SEQ}$ and $f_q^{STG}$ and two key encoders $f_k^{SEQ}$ and $f_k^{STG}$. A query-key pair $(X_q, X_k)$ is obtained by augmenting a raw action sequence $X$ with $D$ as before. We instantiate two different skeleton-representation pairs $(X_q^{SEQ}, X_k^{SEQ})$ and $(X_q^{STG}, X_k^{STG})$. Then, for the query in each input representation, we generate the positives and negatives from the key encoder of the other input representation and vice versa. The encoders $(f_q^{SEQ}, f_q^{STG})$ are trained jointly using a cross-contrastive loss function:

$$L(X^{SEQ}, X^{STG}) = L(X^{SEQ}) + L(X^{STG}),$$

$$L(X^{SEQ}) = - \log \frac{\exp(Z^{SEQ}_q, Z^{STG}_k / \tau)}{\exp(Z^{SEQ}_q, Z^{STG}_k / \tau) + \sum_{Z_n \in X^{SEQ}} \exp(Z^{SEQ}_q, Z^{SEQ}_n / \tau)}.$$  

$$L(X^{STG}) = - \log \frac{\exp(Z^{STG}_q, Z^{SEQ}_k / \tau)}{\exp(Z^{STG}_q, Z^{SEQ}_k / \tau) + \sum_{Z_n \in X^{STG}} \exp(Z^{STG}_q, Z^{SEQ}_n / \tau)}.$$  

where $Z^{SEQ}_q, Z^{SEQ}_k$ is the embedding of the sequence-based query and key. $N^{SEQ}$ is the current set of negative sequence-based embeddings. These are defined similarly for the other representations and augmentations of $X$. This formulation serves two purposes. First, the input space of the contrastive task is enriched to learn from multiple representations of the same sequence, in addition to the multiple 'views' the data augmentation $D$ provides. Second, different from Equation (5), the cross-contrastive loss i.e. Equation (6) forces the framework to rely on mutual information between the embeddings of the two skeleton representations. Thus the contrastive framework is encouraged to focus on higher-level semantics and avoid resorting to shortcut solutions to identify the similarity between query-key pairs.

4 EXPERIMENTS

We first describe the datasets and implementation details. We then demonstrate the effectiveness of our contrastive learning approach on several 3D action understanding downstream tasks. Finally, we ablate the effects of our proposed skeleton augmentations and inter-skeleton contrast.

4.1 Datasets and Evaluation

NTU RGB+D 60 [32]. This is the most commonly used dataset for 3D action recognition. All actions are captured in indoor scenes with three cameras concurrently. The dataset contains 40 different subjects and 60 action classes. Each action sequence is performed by an individual or pair of actors with each actor represented by the 3D coordinates of 25 skeleton joints. The dataset consists of 56,880 video samples and is evaluated under the two standard protocols as suggested by [32]. The first is cross-view, where samples from two angles (0°, 45°) are used for training (37,920 samples) and a third angle (−45°) is used for testing (18,960 samples). The second is cross-subject, where the actors in the training and testing sets are different, with 40,320 training and 16,560 testing samples.

NTU RGB+D 120 [20]. This is an extension to NTU RGB-D 60 and is currently the largest benchmark for 3D action recognition.

Figure 5: Inter-skeleton contrast. We learn invariances to input skeleton representations, as well data augmentations, in a cross-contrastive manner. We first augment the input sequence into two different views called the query and key using our proposed spatio-temporal augmentations. Each of these views is then represented with two different input skeleton-representations, here graph-based and sequence-based. We encourage the embedding for the graph-based query to be similar to the embedding of the sequence-based key while being dissimilar to the current set of sequence-based negatives. The same applies for the sequence-based query and graph-based key and negatives.
Table 1: 3D action recognition. Our method learns better 3D-action features from unlabeled data than alternatives, no matter the dataset or evaluation protocol. All results of Zheng et al. and Su et al. obtained with code provided by Su et al.

<table>
<thead>
<tr>
<th></th>
<th>NTU RGB+D 60</th>
<th>NTU RGB+D 120</th>
<th>PKU-MMD I</th>
<th>PKU-MMD II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>x-view</td>
<td>x-sub</td>
<td>x-view</td>
<td>x-sub</td>
</tr>
<tr>
<td>Zheng et al. [49]</td>
<td>56.4</td>
<td>52.1</td>
<td>39.7</td>
<td>35.6</td>
</tr>
<tr>
<td>Lin et al. [19]</td>
<td>–</td>
<td>52.5</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Su et al. [39]</td>
<td>59.3</td>
<td>56.1</td>
<td>44.1</td>
<td>41.1</td>
</tr>
<tr>
<td>Nie et al. [27]</td>
<td>79.7</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>This paper</td>
<td>85.2</td>
<td>76.3</td>
<td>67.9</td>
<td>67.1</td>
</tr>
</tbody>
</table>

with 114,480 samples over 120 action classes. Actions are captured with 106 subjects in a multi-view setting using 32 different setups (varying camera distances and background). Each action sample has 1 or 2 subjects, and each is represented by 25 3D-skeleton joints. The dataset is challenging due to the variation in subject, background, viewpoint and fine-grained actions captured. For evaluation, two recommended protocols [20] are used: cross-setup, where even-numbered setups are used for training (54,471 samples) and odd-numbered setups are used for testing (59,477 samples), and again cross-subject, with 63,026 training and 50,922 testing samples.

PKU-MMD [5]. This dataset was originally proposed for action detection but has also been used for action recognition [19]. It contains 52 human action classes. Each action is represented by the 3D coordinates of the 25 joints of each actor involved in the action. The dataset consists of two parts: PKU-MMD I and PKU-MMD II, with almost 20,000 and 7,000 action instances. Both parts are challenging for action recognition, as the number of action classes is large while the training sets are relatively small, however PKU-MMD II is more challenging due to the large view variation causing more skeleton noise. We split both sets into a training and a testing set using the recommended cross-subject protocol [5]. The training sets of PKU-MMD I & II contain 18,841 and 5,332 samples, while the testing sets contain 2,704 and 1,613 samples.

Evaluation Criteria. For all datasets, protocols and downstream tasks we report the top-1 accuracy.

4.2 Implementation Details

Network Architectures. We instantiate the encoder pairs \( f_q, f_k \) based on the skeleton-representations used. For the sequence-based encoder \( f_{SEQ} \) we rely on a 3-Layer Bi-GRU with \( H=1024 \) units per layer [39]. For the image representation encoder \( f_{IMG} \), we adopt the CNN based Hierarchical Co-occurrence Network (HCN) [17]. For the graph representation encoder \( f_{STG} \), a joint based graph-convolutional network A-GCN [33] is used. We represent each skeleton sequence \( X \) as two people, with the second actor being all zeros for single actor actions. The augmented forms of the raw skeleton sequence \( X (x_q, x_k) \) have temporal length 64. Unless mentioned otherwise we use \( |j| = 15 \) for the joint jitter augmentation and \( l_{min} = 0.1 \) for the temporal crop-resize augmentation.

Self-Supervised Pretraining. Our inter-skeleton contrastive network is based on MOCO [11] and is trained on the training data without any labels. A projection head (an MLP) is appended to each encoder to produce embeddings of a fixed size of 128. The embeddings are L2-normalized before computing the contrastive loss.

We train the whole network with a temperature value of \( T=0.07 \), an SGD optimizer, a learning rate of 0.01 and a weight decay of 0.0001. For NTU RGB+D 60 & 120, the size of the set negatives \( N \) is 16,384 and the model is pre-trained for a total of 450 epochs. For PKU-MMD I & II, the size of \( N \) is set to 8,192 and 2,048, and we pre-train for 600 epochs. The training and evaluation details of the downstream tasks are discussed in the supplementary material.

4.3 Downstream Tasks

In this section, we evaluate the 3D action features learned by our inter-skeleton contrast for various downstream tasks in comparison with the respective state-of-the-art in self-supervised learning. For a fair comparison we follow the setups of prior works and only train and evaluate downstream tasks with the sequence-based input representation \( X_{SEQ} \). In particular, we pre-train our inter-skeleton contrast network with \( X_{SEQ} \) and \( X_{STG} \) skeleton representations as this gives the best result (see Section 4.4) and evaluate only the sequence-based query encoder \( f_q^{SEQ} \). We also show some qualitative results in the supplementary material.

3D Action Recognition. We compare our method to prior works in self-supervised learning for skeleton data by training a linear classifier on top of the frozen features from our inter-skeleton contrast. We compare with the proposed methods of Zheng et al. [49], Su et al. [39] and Nie et al. [27], all of which use reconstruction of the skeleton sequence as a pretext task. We also compare to the multi-task self-supervised method by Lin et al. [19], which uses skeleton-jigsaw and motion prediction as auxiliary tasks.

We present results on the NTU RGB+D 60, NTU-120 and PKU-MMD (I and II) datasets in Table 1. It is evident our inter-skeleton contrast outperforms all methods by a considerable margin on each benchmark. We conclude the self-supervised feature space learned by our method is state-of-the-art for 3D action recognition.

3D Action Retrieval. We follow the setup introduced by Su et al. [39]. We apply the kNN classifier \( (k=1) \) to the pre-trained features of the training set to assign classes. We match each test sample to the most similar training class using cosine similarity. Besides comparison with Su et al. [39], we also compare with Zheng et al. [49], using numbers and code provided by Su et al. We present results for NTU RGB+D 60 and NTU RGB+D 120 in Table 3. For both datasets, our method outperforms the alternatives, especially for the more challenging cross-subject and cross-setup protocols. Both [39, 49] rely on an input reconstruction pretext-task for learning their feature space, which easily captures varying viewpoints. However,
with a simple reconstruction, it is difficult to capture variation with respect to subjects and setups as our inter-skeleton contrast can. **Semi-Supervised 3D Action Recognition.** In semi-supervised setting, a network utilizes both labeled and unlabeled data during the training process. Following prior work for semi-supervised learning in 3D action recognition, we first train our encoder on an unsupervised inter-skeleton contrastive learning task. Then, we fine-tune the final classification layer and the pre-trained encoder together using a portion of the data labeled with the action class. Again, we compare with Zheng et al. [49] and Lin et al. [19] as well as the method of Si et al. [34] on NTU RGB+D 60 and the PKU-MMD I datasets. To compare with prior works, we report results when using 1%, 5%, 10% and 20% of the training data with labels for NTU RGB+D 60 and when using 1% and 10% of the labels for PKU-MMD I. The rest of the training set is used as the unlabeled data.

The results in Table 2 reveal that our method outperforms all previous methods on each benchmark. We also demonstrate a large improvement over supervised only training, i.e. training with only the available labeled data from randomly initialized weights. From these results we can see that our inter-skeleton contrastive learning is especially suited to learn from both unlabeled and labeled skeleton data in order to boost the performance of 3D action recognition. **Transfer Learning for 3D Action Recognition.** To evaluate if knowledge gained from a source dataset generalizes to a different target dataset, we also consider transfer learning. In this setting, an encoder network is first trained on the source dataset for our inter-skeleton contrastive task, followed by jointly finetuning the pretrained encoder and a classifier on a target dataset for action recognition. As in Lin et al. [19], we use NTU RGB+D 60 and PKU-MMD I as the source datasets and PKU-MMD II as the target dataset. Table 4 shows our features are just as or more transferable than those of Zheng et al. [49] and Lin et al. [19], especially for transfer from PKU-MMD I to PKU-MMD II which are from same domain. Thus, the knowledge gained by our method from a source dataset can improve action classification accuracy on a different target set, especially one with a similar domain.

### 4.4 Ablation Studies

We now ablate the effect of each of our skeleton augmentations and demonstrate the effectiveness of our inter-skeleton contrastive learning. These ablations are performed on the cross-view protocol of NTU RGB+D 60 for the downstream task of 3D action recognition. As before, after pre-training the models with our contrastive self-supervision methods, we train a linear classifier with action labels on top of the frozen features of the query encoder $f_q$.

**Benefit of Skeleton Augmentation.** First, we show the benefit of each of the proposed skeleton augmentations when learning from a single input skeleton representation. We choose as skeleton augmentation function $D$, either pose augmentation, joint jitter, temporal crop-resize or combinations thereof, and train an intra-skeleton contrastive model as described in Section 3.2. Table 5 shows the accuracy of our augmentations with each input representation. We find that all of the proposed spatial and temporal skeleton augmentations individually perform better than using no augmentation. Thereby, reinforcing our claim that learning invariances to spatial changes like viewpoints, scale and joint perturbations, or, temporal changes such as delay and speed result in learning good action features. The composition of augmentations further improves the accuracy by a considerable margin for all input representations, with the best combination being the inclusion of all three augmentation functions. For example, the final accuracy with the $X_{IM}$ representation is a $\sim10\%$ increase over using only pose augmentation and $\sim28\%$ over using no augmentation.
Table 5: Benefit of skeleton augmentation. We ablate the effect of our augmentations with 3D action recognition on NTU RGB+D 60. Combining all three augmentations generates strong positive pairs for increased accuracy, no matter the 3D action representation.

<table>
<thead>
<tr>
<th>Augmentations</th>
<th>Downstream Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X_{IMG}</td>
</tr>
<tr>
<td>Temporal</td>
<td>51.0</td>
</tr>
<tr>
<td>Crop-resize</td>
<td>✓</td>
</tr>
<tr>
<td>Pose</td>
<td>✓</td>
</tr>
<tr>
<td>Joint</td>
<td>✓</td>
</tr>
<tr>
<td>Jitter</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 6: Intra-skeleton vs. Inter-skeleton. Training alongside a second input representation in our inter-skeleton contrast results in better features for all input representations, regardless of the pair used. Note that a representation can only be used in the downstream task when it is present in pre-training. Ablation performed on 3D action recognition with NTU RGB+D 60.

We first observe that pre-training with any two skeleton representations side by side in our inter-skeleton contrast is considerably better than only learning with a single representation as in the intra-skeleton contrast. For example, the accuracy with X_{STG} increases by 6% when pre-trained together with X_{SEQ} in our inter-skeleton contrast model. A similar increase of 5% occurs when pre-training alongside X_{IMG}. We find this to be the case with each skeleton representation; regardless of the second representation it is trained alongside in the inter-skeleton contrast, there is an increase in performance. We also tried training all three skeleton representations together. While this does give the best result, the improvement is outweighed by the computational cost of training all three representations simultaneously. Overall, these results reinforce our claim that learning invariance to skeleton augmentations alone leads to sub-optimal features and learning additional invariance to skeleton-representations results in a better feature space.

5 CONCLUSION

In this work, we presented a method for self-supervised learning of 3D skeleton data. We design a contrastive learning framework that relies on novel skeleton augmentations and multiple skeleton-representations to learn spatio-temporal dynamics of the skeleton sequences. Our comprehensive evaluation with different skeleton augmentations and skeleton-representation pairs reveal that learning invariance to our spatio-temporal augmentations and contrasting sequence-based and graph-based representations with each other results in best action features. The final model achieves considerable performance gains and outperforms prior state-of-the-art in self-supervised learning for multiple downstream tasks on NTU RGB+D 60 & 120 and PKU-MMD.

ACKNOWLEDGMENTS

This work is part of the research programme Perspectief EDL with project number P16-25 project 3, which is financed by the Dutch Research Council (NWO) domain Applied and Engineering/Sciences (TTW).
REFERENCES


[33] Lei Shi, Yifan Zhang, Jian Cheng, and Hangming Lu. 2019. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In CVPR.


[43] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In AAAI.


