DuoCLR: Dual-Surrogate Contrastive Learning for Skeleton-based Human Action Segmentation

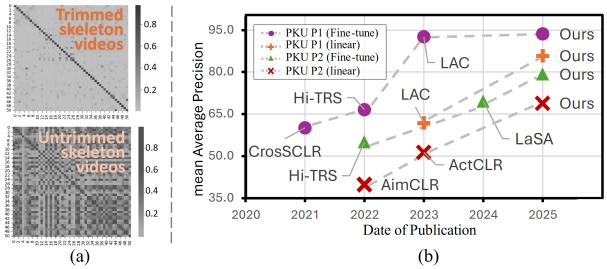
Haitao Tian
University of Ottawa, Canada
htian026@uottawa.ca

Abstract

In this paper, a new contrastive representation learning framework is proposed to enhance action segmentation via pretraining using trimmed (single action) skeleton sequences. Unlike previous representation learning works that are tailored for action recognition and that develop isolated sequence-wise representations, the proposed framework focuses on exploiting multi-scale representations in conjunction with cross-sequence variations. More specifically, it proposes a novel data augmentation strategy, "Shuffle and Warp", which exploits diverse multi-action permutations. The latter effectively assist two surrogate tasks that are introduced in contrastive learning: Cross Permutation Contrasting (CPC) and Relative Order Reasoning (ROR). In optimization, CPC learns intra-class similarities by contrasting representations of the same action class across different permutations, while ROR reasons about inter-class contexts by predicting relative mapping between two permutations. Together, these tasks enable a Dual-Surrogate Contrastive Learning (DuoCLR) network to learn multi-scale feature representations optimized for action segmentation. In experiments, DuoCLR is pretrained on a trimmed skeleton dataset and evaluated on an untrimmed dataset where it demonstrates a significant boost over state-the-art comparatives in both multi-class and multi-label action segmentation tasks. Lastly, ablation studies are conducted to evaluate the effectiveness of each component of the proposed approach.

1. Introduction

Contrastive learning [4, 5, 18] opens a new opportunity for taking advantage of the growing volume of unlabeled data for learning a rich latent feature representation aware of underlying data structures and semantics without human annotations. This approach effectively reduces the need for extensive labeled data in downstream task fine-tuning. In human-centric video understanding, human skeleton data



(a) (b)

Figure 1. Limitation of traditional skeleton-based contrastive learning. (a): Distinct feature similarity matrices obtained from feature extraction on trimmed (upper) vs. untrimmed (lower) skeleton videos. Entry (i, j) represents the feature similarity (Eq. (1)) between action classes i and j. The feature extractor is learned on NTU RGB+D using AimCLR [18]. Trimmed and untrimmed videos are both sampled from PKU Part 1. For untrimmed videos, action-specific features are extracted via post-segmentation in the feature space using frame annotations (detailed in Sec. 3.3). (b): Trend of contrastive learning performance (detailed in Table 1) in skeleton action segmentation by publication year.

[12, 29, 35] has become a widely used modality in contrastive learning as it depicts the trajectories of key body joints and provides a compact and robust depiction of human motion that is resilient to environmental variations. Existing research primarily involves trimmed (single action) skeleton sequences into contrastive learning, where the model treats each sequence as an instance and learns intra-action temporal representations (i.e., *similarities*) through instance discrimination [6, 9, 46, 47]. The learned representations have been effectively transferred into trimmed-sequence-based downstream tasks such as *action recognition* [1, 9, 17, 20, 28, 38, 42, 48, 54] and *action retrieval* [13, 18, 24, 34, 49].

However, complex human-centric video understanding applications such as *action segmentation* often involve untrimmed long skeleton videos exhibiting multiple actions and complex action correlations. While focusing on sequence-wise instance discrimination tasks, traditional contrastive learning generally neglects the exploitation of inter-sequence temporal dependencies (i.e., *contexts*) that

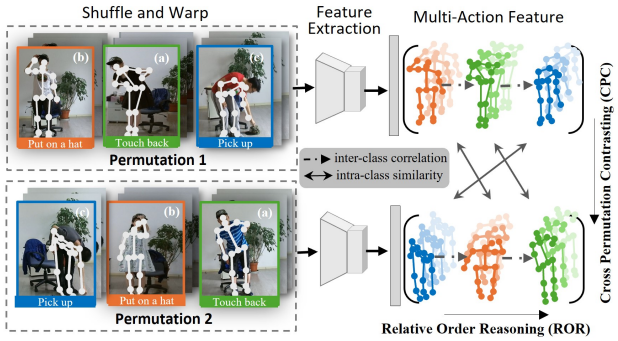


Figure 2. The proposed contrastive learning framework for action segmentation utilizes multi-action permutations. For example, six single-action skeleton sequences are randomly shuffled and warped into two sequences, Permutation 1 and Permutation 2, with three shared actions. Such a data augmentation introduces dynamic cross-sequence variations that enhance learning in two ways: i) CPC contrasts intra-class instance across different permutations, promoting permutation-invariant representations despite changing contexts (e.g., "Put on a hat" appears in both permutations but in different contexts); ii) ROR captures inter-class dependencies by mapping one permutation to another (e.g., learning the transformation from (b, a, c) to (c, b, a)).

are critical for action segmentation. To highlight this limitation, Fig. 1a experimentally verifies that a state-of-the-art contrastive learning paradigm [18], which learns isolated sequence-wise representations, may lead to significantly different feature representations when applied to trimmed and untrimmed videos. We hypothesize that such a discrepancy results from the extraction of inconsistent action representations from dynamic action contexts permuted in untrimmed sequences, which leads to suboptimal transfer learning performance to action segmentation tasks (as illustrated in Fig. 1b).

In this work, a new contrastive learning framework is proposed to improve action segmentation on *untrimmed* videos by pretraining on *trimmed* skeleton videos. **First, the framework leverages multi-action permutations to fulfill cross-sequence data augmentation.** Previous single-action augmentations use simple sequence-wise modifications, e.g., spatial transformation [9, 18] and temporal scaling [25, 49], that generate low-level context-free data variations. This work rather introduces "Shuffle and Warp" to exploit high-level contextual variations: (1) "Shuffle" treats trimmed skeleton sequences as modular action elements that can be sampled and shuffled, introducing dynamic cross-sequence variations; and (2) "Warp" leverages geometrical transformations to concatenate the shuffled sequences into a view-consistent permutation of actions.

Second, two new surrogate tasks ensure multi-scale representation learning. To fully exploit the temporal variations present in multi-action permutations, this work introduces two contrastive learning surrogate tasks, Cross Permutation Contrasting (CPC) and Relative Order Reasoning (ROR). As conceptually illustrated in Fig. 2, CPC learns *permutation-invariant intra-class similarities* via instance discrimination where the representation of the same action class should be similar regardless of context variations arising from their neighboring segments. ROR focuses on the *permutation-aware inter-class contexts* by predicting the temporal mapping between action permutations, encouraging the model to learn relative positioning among actions, thereby adding a complementary regularization to the feature space.

Lastly, a hierarchical network facilitates sliding-window-free transfer learning. We amalgamate the key heuristics into a **Du**al-Surr**o**gate **C**ontrastive **L**ea**R**ning (DuoCLR) network that can be trained end-to-end. Unlike previous works [7, 28, 48, 50] relying on a single visual encoder, DuoCLR employs a cascaded structure [16] composed of a visual encoder and a temporal encoder. The additional temporal encoder enables a large temporal receptive field, allowing the model to capture long-range action dependencies without requiring sliding-window operations [7, 11, 48] on untrimmed skeleton videos. As a result, DuoCLR surpasses state-of-the-art methods in action segmentation (Fig. 1b). After pretraining, both encoders can be jointly fine-tuned as a feature extractor for downstream tasks.

The original contributions of the work are threefold. First, a new contrastive learning framework for skeleton data-based representation learning enables trimmed sequences to learn a fine-grained representation for action segmentation. Second, a new data augmentation strategy, "Shuffle and Warp", for skeleton data introduces variations in between data sequences to improve feature learning. Third, the framework builds upon a dual-contrastive approach by combining CPC and ROR. It focuses on instance coherence and instance correlation, both critical for action segmentation. The project is available at: https://htian026.github.io/DuoCLR, which contains code implementation.

2. Related work

Skeleton-based Action Segmentation. Compared to skeleton action recognition [6, 14, 25, 44, 46, 47] that aims at sequence-wise action classification, skeleton action segmentation [2, 3, 16, 26, 31] is inherently more challenging as the model must aggregate both action patterns and contexts among multiple actions permutated in a single sequence. To tackle the difficulty, recent research [16, 21, 21, 26] proposes to optimize an end-to-end model that incorporates frame-wise action classification into supervised training. Yet, training an action segmentation model from scratch relies heavily on frame-level

annotations that are costly and time-consuming to generate [12, 29]. In this work, we circumvent the data annotation usage by leveraging trimmed skeleton videos for pretraining, where a contrastive learning network can leverage numerous action permutations to pre-learn a feature extractor relevant to action segmentation, and quick fine-tuning to untrimmed videos.

Skeleton Data Augmentation is primarily achieved by single-action skeleton modifications. For instance, studies [1, 7, 18, 20, 24, 28, 38, 42, 49, 53] delve into designing more effective single-action augmentations, e.g., "Shear", "Crop", "Spatial Flip", "Spatial Rotate", "Spatial Jitter", "Axis Mask", "Gaussian Noise", and "Gaussian Blur", demonstrating the effectiveness of carefully involving multiple augmentations in representation learning [18, 49]. Recent works also leverage multi-action skeleton augmentations, e.g., "Latent Action Composition" [48] and "Skeleton Actionlet" [28] are proposed to mix spatiotemporal statistics of two sequences; "Skeleton-Mix" [8, 19, 28, 30, 43, 51] is proposed to generate a synthetic sequence by merging both data and label from two sequences into a composable sequence. SCS [39] proposes a temporal skeleton stitching scheme to generate multi-action stitched sequences. However, it imposes an offline data generation process requiring time-consuming frame registration operations. In contrast, this work proposes utilizing online multi-action permutations, "Shuffle and Warp", a new data augmentation scheme which will be detailed in Section 3.2.

Skeleton-based Representation Learning [4, 5] is primarily achieved by learning action similarities from augmented skeleton views via effective surrogate tasks such as instance discrimination [1, 7–9, 13, 17–20, 24, 28, 42, 49, 53, 54] , jigsaw puzzles [27, 37], adversarial learning [52], and language-video contrasting [21]. However, prior works mainly produce sequence-level representations, which are effective for simplified downstream tasks like action recognition but may be insufficient for action segmentation, where fine-grained feature representations are crucial. Recent efforts [7, 11, 48] attempt to mitigate this by applying sliding windows to generate fixed-length action clips from untrimmed sequences, enabling the use of standard action recognition models. However, this strategy introduces noisy boundaries, leading to degraded segmentation accuracy (experimentally verified in Section 4.3). These limitations highlight the need for a contrastive learning paradigm tailored for action segmentation.

3. Dual-Surrogate Contrastive Learning

3.1. Overview

Let $X \in \mathbb{R}^{T\times V\times C}$ be a trimmed skeleton video with T skeleton frames V skeleton joints, and C spatial dimensions, which covers the motion of a single action class. A contrastive learning paradigm aims to use trimmed skeleton videos to pretrain a feature extraction network $\mathcal{F}$ by which the features of the same action class should be similar. In previous studies [18, 20, 24], feature similarity between two sequences, X_i andX_j, is calculated as:

$$\text{sim}\big(\mathcal{F}(X_i), \mathcal{F}(X_j)\big) = \frac{\mathcal{F}(X_i)\,\mathcal{F}(X_j)}{|\mathcal{F}(X_i)||\mathcal{F}(X_j)|\cdot\tau} \tag{1}$$

where τ acts as the temperature parameter. This work aims to pretrain a $\mathcal{F}$ using Dual-Surrogate Contrastive Learning (DuoCLR) and customize a task-specific (action segmentation) layer $\mathcal{C}$ for a target dataset composed of a small number of frame-wise labeled and untrimmed videos. Section 3.2 describes the Shuffle and Warp data augmentation technique for generating dynamic multi-action permutations. Section 3.3 details the network structure of $\mathcal{F}$. In Sections 3.4 and 3.5, DuoCLR introduces two surrogate tasks (as conceptually depicted in Fig. 3), Cross Permutation Contrasting (CPC) and Relative Order Reasoning (ROR), respectively.

3.2. Multi-Action Permutations

Capturing temporal contexts is critical for action segmentation as it requires the model to discern the state of actions over a single untrimmed skeleton sequence. Since trimmed skeletal data lacks such context, Shuffle and Warp is introduced to generate dynamic multi-action permutations to support improved contrastive learning for action segmentation.

Temporal Shuffling was previously utilized to generate single-action augmentation where segments of a trimmed sequence are shuffled to solve a jigsaw puzzle task for action recognition [27, 33]. In this work, we rather involve multiple trimmed sequences at each augmentation of temporal shuffling to introduce cross-sequence variations. Specifically, let $\mathcal{X} = \{X^{(a)}, X^{(b)}, X^{(c)}, \dots\}$ be a set of trimmed skeleton sequences sampled from a data batch, we can generate a multi-action permutation as:

$$\textit{Shuffle}(\mathcal{X}, \mathcal{P}) = \textit{concat}\left(X^{\mathcal{P}[1]}, X^{\mathcal{P}[2]}, X^{\mathcal{P}[3]}, \dots\right) \tag{2}$$

where the function "*Shuffle*" first permutes the elements of $\mathcal{X}$ according to the permutation $\mathcal{P}$, e.g.,$\mathcal{P} = (b, c, a, ...)$, and then generates a single sequence using the tensor concatenation operation "*concat*". The proposed data augmentation supports both labeled and unlabeled skeleton data where $a, b, c, ...$ denotes action classes while labeled and sequence indices while unlabeled (discussed in Section 4.3).

Temporal Warping. However, skeleton sequences collected under different imaging configurations may lead to unnecessary data variations in the skeleton scale and poses [35]. A representation learning network may use such intricacies to develop shortcuts to fulfill the surrogate task of

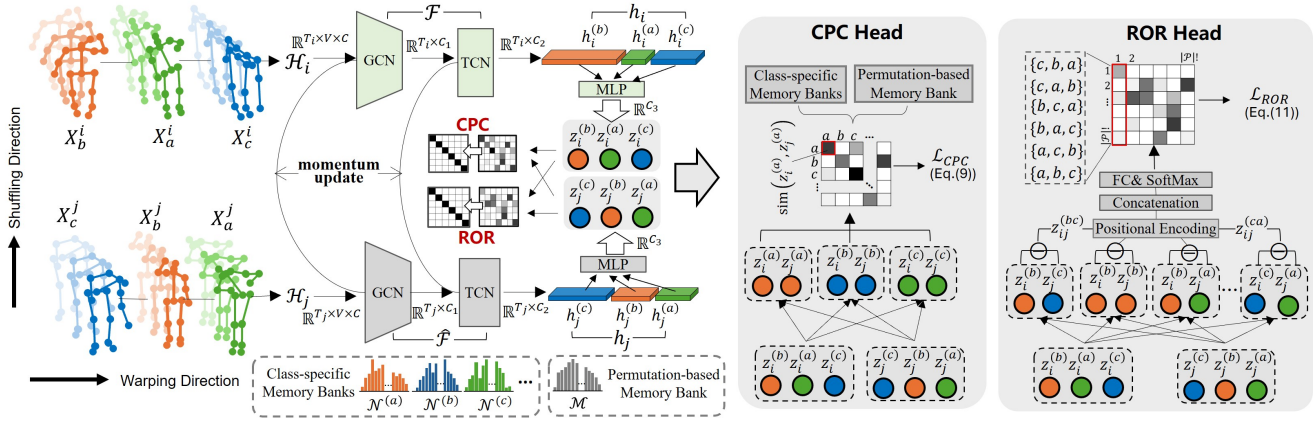


Figure 3. Computational workflow of the proposed DuoCLR approach. Trimmed single-action skeleton videos are augmented by using "Shuffle and Warp" on the input. In DuoCLR, the two surrogate tasks, "ROR" and "CPC", share the feature encodings, $z_i^{(n)}$ and $z_j^{(n)}$, while fulfilling different pipelines for representation learning. By solving simultaneous CPC and ROR, the network learns a fine-grained action representation for action segmentation.

contrastive learning [18, 38]. To tackle the issue, temporal warping is newly proposed to eliminate the unnecessary data variations by using a pairwise sequence concatenation operation. Specifically, given two heterogenous sequences $X_i \in \mathbb{R}^{T_i \times V \times C}$ and $X_j \in \mathbb{R}^{T_j \times V \times C}$, it uses a rotation $\mathcal{R} \in \mathbb{R}^{C \times C}$, a translation $\mathcal{T} \in \mathbb{R}^{C \times 1}$, and a scaling factor $s \in \mathbb{R}$, to map X_j into the camera configuration of X_i. In this way, "$Warp$" is defined as:

$$Warp\,(X_i, X_j) = concat\,(X_i, s\,(X_j * \mathcal{R} + \mathcal{T})) \qquad (3)$$

where $*$ denotes the batched dot product. Details regarding parameters estimation and sequence operation are available in supplementary material.

At last, we use "Shuffle and Warp" to generate the multi-action permutation $\mathcal{H}$ as follows

$$\mathcal{H} = Warp(Shuffle(\mathcal{X}, \mathcal{P})), \qquad (4)$$

which denotes that $\mathcal{H}$ is generated by recursively operating the "*Warp*" operation on shuffled $\mathcal{X}$ under $\mathcal{P}$.

3.3. Feature Extraction

In this work, the feature extraction network $\mathcal{F}$ is composed of a visual encoder and a temporal encoder, as illustrated in Fig. 3.

Visual Encoder. It is common in the research community [13, 17, 20, 25, 28, 28] to use a graph convolutional network (GCN) [46] as a visual encoder to parse intermediate skeleton feature embeddings. In this work, we removed the temporal pooling layer of GCN to keep the feature's temporal resolution, i.e., GCN:$\mathbb{R}^{T \times V \times C} \rightarrow \mathbb{R}^{T \times C_1}$. However, a vanilla GCN may be limited by their small temporal receptive fields [16], thus not exploiting contextual structures that are relevant to action segmentation.

Temporal Encoder. In order to effectively capture action relationships, we stack a temporal encoder onto the visual encoder to constructs $\mathcal{F}$. The temporal encoder adopts a temporal convolution network [15] that utilizes a series of dilated temporal convolutions to aggregate long-range temporal contexts of the hidden state from the visual encoder, i.e., $TCN : \mathbb{R}^{T \times C_1} \rightarrow \mathbb{R}^{T \times C_2}$.

At last, given a skeleton sequence input X, its hidden feature representation h is obtained as:

$$h = \underbrace{(\text{TCN} \circ \text{GCN})}_{\mathcal{F}}(X) \qquad (5)$$

Multi-grained feature projections. DuoCLR differs from prior contrastive learning works that input single-action sequences. DuoCLR rather utilizes multi-action permutations for feature extraction prior to contrastive learning. As illustrated in Fig. 3, DuoCLR uses Shuffle and Warp (Eq. (4)) to generate a pair of multi-action permutations $\mathcal{H}_i$ and $\mathcal{H}_j$ of $\mathcal{X}$ with two random permutations $\mathcal{P}_i$ and $\mathcal{P}_j$. Using Eq. (5), it obtains multi-action representations $h_i = \mathcal{F}(\mathcal{H}_i)$ and $h_j = \bar{\mathcal{F}}(\mathcal{H}_j)$, where $\bar{\mathcal{F}}$ denotes a moving average of the feature extractor $\mathcal{F}$ and we follow the same implementation introduced in [18]. For notation simplicity, let's assume the cardinality of $\mathcal{X}$ is 3, e.g., $\mathcal{X} = X^{(a)}, X^{(b)}, X^{(c)}$, and assume $\mathcal{P}_i = (b, a, c)$ and $\mathcal{P}_j = (c, b, a)$. In this way, DuoCLR generates multi-grained feature projections for contrastive learning by treating the feature representations, h_i and h_j, as concatenations of action-wise encodings: $h_i = concat(h_i^{(b)}, h_i^{(a)}, h_i^{(c)})$ and $h_j = concat\left(h_j^{(c)}, h_j^{(b)}, h_j^{(a)}\right)$. Furthermore, an MLP-based projection head is used to generates *local* and *global* feature projections:

$$z_i^{(n)} = MLP[\phi(h_i^{(n)})]; z_i = MLP[\phi(h_i)] \qquad (6)$$

where n=a,b,c and ϕ denotes the temporal average pooling operation. Likewise, it generates $z_j^{(n)}$ and z_j.

3.4. Cross Permutation Contrasting (CPC)

While treating each action within a permutation as an instance, individual instances of the same action class can have different contextual neighbors (e.g., the action "Put on a hat" in Fig. 2 may appear next to different actions in each permutation). Traditional methods [18, 25, 28, 48] overlook such information and learn isolated sequence-wise representations, thereby achieving suboptimal transferability in action segmentation (as illustrated in Fig. 1). In this work, DuoCLR utilizes multi-action permutations as data augmentation and employs Cross Permutation Contrasting (CPC) for self-supervised representation learning. Since the augmentation of action permutations expose the network to numerous action temporal contexts, CPC learns permutation-invariant temporal coherence where intra-class representations from the different permutations should be pulled together while inter-class representations should be pushed away.

Permutation Bank-based Instance Discrimination. In implementation, CPC employs local feature projections $z_i^{(n)} \in \mathbb{R}^{C_3}$ and $z_j^{(n)} \in \mathbb{R}^{C_3}$ as positive pairs and formulates class-specific memory banks $\mathcal{N} = \{\mathcal{N}^{(m)}\}$ to store negative instances, each updated with $z_j^{(m)}$ where $m \neq n$. Afterwards, it learns short-term invariant representations for each instance regardless of surrounding context changes by optimizing a InfoNCE loss [5]:

$$\mathcal{L}_{\mathcal{N}}^{(n)} = -log\frac{sim\left(z_i^{(n)}, z_j^{(n)}\right)}{sim\left(z_i^{(n)}, z_j^{(n)}\right) + \sum_{z^- \in \mathcal{N}} sim\left(z_i^{(n)}, z^-\right)} \tag{7}$$

where z^- is a negative sample from the memory bank.

Second, CPC also involves a permutation (multi-action) memory bank $\mathcal{M}$ in contrastive learning, which is updated by z_j. When $\mathcal{H}_i$ and $\mathcal{H}_j$ share the same permutation, i.e., $\mathcal{P}_i = \mathcal{P}_j$, it treats each permutation as an instance and encourages the network to learn long-term invariant representations of permutations. It calculates the similarity between permutation-wise positive pair, $z_i \in \mathbb{R}^{C_3}$ and$z_j \in \mathbb{R}^{C_3}$, and optimizes the second InfoNCE loss as:

$$\mathcal{L}_{\mathcal{M}} = -log\frac{sim(z_i, z_j)}{sim\left(z_i, z_j\right) + \sum_{z^- \in \mathcal{M}} sim(z_i, z^-)} \tag{8}$$

Upon the formulation of Eq. (7) and Eq. (8), the final objective for CPC is formulated as:

$$\mathcal{L}_{CPC} = \lambda \sum_{n=a,b,c,...} \mathcal{L}_{\mathcal{N}}^{(n)} + (1-\lambda)\mathcal{L}_{\mathcal{M}} \tag{9}$$

where λ is equals to 1 if $\mathcal{P}_i \neq \mathcal{P}_j$, and 0 otherwise.

3.5. Relative Order Reasoning (ROR)

Order reasoning of action videos is a fundamental task in video representation learning. Traditional approaches primarily focus on chronological order prediction within a single-action class, where models learn temporal coherence by reordering shuffled video frames [23] or clips [27, 33, 45]. However, in multi-action sequences, the order is inherently ambiguous, making traditional approaches not applicable. To address this, we introduce Relative Order Reasoning (ROR), a novel paradigm that leverages multi-action permutations. Instead of predicting a single "correct" order, ROR focuses on "relative" mapping between paired action permutations and learns the capability of discerning the latent state of actions over long skeleton sequences, thus regularizing the feature space from a complementary direction to CPC.

In implementation, DuoCLR solves ROR via classification upon the mapping between $\mathcal{H}_i$ and $\mathcal{H}_j$. Specifically, since each pair of action permutations share the same actions, the relative mapping between the paired permutations is uniquely determined. For example, let $|\mathcal{P}_j|$ (where $\mathcal{P}_j = (c, b, a)$) be the action granularity of $\mathcal{H}_j$, there exist $|\mathcal{P}_j|! = 6$ factorial orders: $\{(c, b, a), (c, a, b), (b, c, a), (b, a, c)...\}$. The mapping from $\mathcal{H}_i$ to $\mathcal{H}_j$ is represented by the one-hot class label $\mathbf{1}(\mathcal{P}_i|\mathcal{P}_j) = [0, 0, 0, 1, 0, 0]^T$ while $\mathcal{P}_i = (b, a, c)$.

Positional Encodings . For optimization, ROR calculates Positional Encodings (PE) of $\mathcal{H}_i$ and $\mathcal{H}_j$. It uses local projections $z_i^{(n)} \in \mathbb{R}^{C_3}$ and $z_j^{(m)} \in \mathbb{R}^{C_3}$ (obtained via Eq. (6)) to calculate pair-wise feature comparison $z_{ij}^{(n,m)} = \left|z_i^{(n)} - z_j^{(m)}\right|$. Second, it generates the concatenation of $\{z_{ij}^{(n,m)} | n \in \mathcal{P}_i, m \in \mathcal{P}_j\}$ to obtain fine-grained position relations between $\mathcal{H}_i$ and $\mathcal{H}_j$:

$$PE\left(\mathcal{H}_i, \mathcal{H}_j\right) = concat(z_{ij}^{(b,c)}, z_{ij}^{(b,b)}, z_{ij}^{(b,a)}, z_{ij}^{(a,c)}, ...). \tag{10}$$

Such an operation allows the network to infer relative positioning of $\mathcal{H}_i$ and $\mathcal{H}_j$ by comparing their temporal structures in feature space, which is also illustrated in Fig 3. Afterwards, ROR utilizes a fully connected layer, $FC : \mathbb{R}^{C_3 \times |\mathcal{P}_j|^2} \rightarrow \mathbb{R}^{|\mathcal{P}_j|!}$, to optimize a Cross-Entropy loss:

$$L_{ROR} = -\mathbf{1}(\mathcal{P}_i|\mathcal{P}_j) \cdot \log\left(FC\left(PE(\mathcal{H}_i, \mathcal{H}_j)\right)\right) \tag{11}$$

DuoCLR Loss. The final loss function for DuoCLR combines CPC and ROR losses, with a weighting factor α to balance their contributions:

$$L = \mathcal{L}_{CPC} + \alpha\mathcal{L}_{ROR} \tag{12}$$

The implementation of DuoCLR upon Eq. (12) leads to an expressive representation $\mathcal{F}$ that is fully transferable to action segmentation and will be comprehensively evaluated in Section 4.

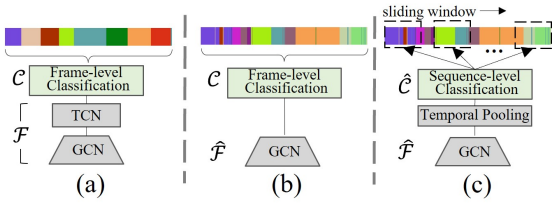


Figure 4. Transfer learning in skeleton-based human action segmentation. (a): Hierarchical structure introduced in Sec. 3.3. (b) and (c): Common structures used in existing works. See more details in supplementary material.

4. Experiments

This section studies the application of DuoCLR in action segmentation. DuoCLR first involves a trimmed dataset for pretraining, followed by an evaluation phase where an untrimmed dataset is used to test the action segmentation performance. The study experimentally demonstrates the effectiveness of the proposed DuoCLR approach in comparison to state-of-the-art methods.

4.1. Implementations

• **Datasets**. **Three trimmed** skeleton datasets are considered for pretraining in DuoCLR: NTU RGB+D (NTU) [35], Toyota Smarthome (TS) [12] and Kinetics-400 (K4) [22]. **Three untrimmed** skeleton datasets are involved for evaluating the pre-trained models in the application of action segmentation, including PKU Multi-Modality Dataset (PKU) [29] (We use both versions: PKU1 and PKU2), Toyota Smarthome Untrimmed (TSU) [11], and Charades (CR) [36]. We report on mean Average Precision at IoU thresholds 0.1 ($\text{mAP}_{0.1}$) and 0.5 ($\text{mAP}_{0.5}$) [48], Top1 Accuracy (Acc) [46], mean Intersection over Union (mIoU) [28], and per-frame mAP (mAP) [11] for cross-view (CV) and cross-subject (CS) test cases of all untrimmed datasets.

• **Pretraining**. We adopt ST-GCN [46] as the skeleton visual encoder and MS-TCN [15] as the skeleton temporal encoder. In the pretraining, we use the entire (both training and testing) set of a trimmed dataset to learn the feature extractor $\mathcal{F}$. We adopt the same hyper-parameter values as in previous works [18, 20, 38] , i.e., C_3 is set as 128 and τ in Eq. (1) as 0.07. The sizes of the two memory banks, $\mathcal{N}^{(m)}$ and $\mathcal{M}$, are set as 684 and 32,768, respectively. More details are available in supplementary material.

• **Evaluation**. We consider two popular downstream tasks: multi-class action segmentation [16, 29] and multi-label (composite) action segmentation [11, 48], in each of which we learn a frame-wise classification (frame annotating) layer $\mathcal{C}$ over the pre-learned $\mathcal{F}$ (as illustrated in Fig. 4a). In multi-class segmentation, each skeleton frame is annotated with one action label. $\mathcal{C}$ is composed of a 1×1 convolution and SoftMax layer. It is optimized by the multi-class Cross-Entropy loss function with one-hot encoded frame annotations of the evaluation dataset. In composite action segmentation, each skeleton frame could be labeled with multiple action labels (e.g., "Cooking" and "Stirring"). $\mathcal{C}$ is composed of a 1×1 convolution and Sigmoid layer and is optimized using the binary Cross-Entropy loss with multi-hot encoded frame annotations of the evaluation dataset. In both applications, we use SGD to update $\mathcal{C}$ with Nesterov momentum 0.9 and learning rate 0.05.

Models	**Method**	**Pretrain**	**Eval.**	**CS%**		**CV%**	
NTU → PKU1:				$\text{mAP}_{.1}$	$\text{mAP}_{.5}$	$\text{mAP}_{.1}$	$\text{mAP}_{.5}$
Baseline-I[46]	Linear	NTU	PKU1	56.0	50.3	56.7	54.9
LAC [48]	Linear	Poetics	PKU1	61.8	-	62.4	-
DuoCLR	Linear	NTU	PKU1	**85.2**	**74.3**	**84.7**	**77.5**
Baseline-II [16]	F-T	-	PKU1	89.6	86.4	92.1	88.9
MS2L [27]	F-T	NTU	PKU1	-	50.9	-	-
CMD [32]	F-T	NTU	PKU1	-	59.4	-	-
CrossSCLR [24]	F-T	NTU	PKU1	-	60.1	-	-
PCM3 [50]	F-T	NTU	PKU1	-	61.8	-	-
Hi-TRS [7]	F-T	NTU	PKU1	-	63.5	-	-
USDRL [41]	F-T	NTU	PKU1	-	75.7	-	-
LAC [48]	F-T	Poetics	PKU1	92.6	**90.6**	94.6	-
DuoCLR	F-T	NTU	PKU1	**94.4**	90.1	**96.8**	**94.8**
NTU → PKU2:				mIoU	Acc	mIoU	Acc
Baseline-I [46]	Linear	NTU	PKU2	30.7	47.2	26.4	44.7
AimCLR [18]	Linear	Poetics	PKU2	-	-	15.7	39.8
ActCLR [28]	Linear	NTU	PKU2	-	-	21.4	51.3
DuoCLR	Linear	NTU	PKU2	**54.5**	**73.9**	**50.9**	**68.8**
Baseline-II [16]	F-T	-	PKU2	50.9	69.3	45.6	66.0
Hi-TRS [7]	F-T	NTU	PKU2	-	-	-	55.0
LaSA [21]	F-T	-	PKU2	-	73.4	-	69.4
DuoCLR	F-T	NTU	PKU2	**66.1**	**82.6**	**56.3**	**79.2**

Table 1. Experimental comparison in the multi-class skeleton-based action segmentation task. Evaluation methods include Linear Evaluation and Fine-tune (F-T) Evaluation.

Models	Methods	Modality	**TS → TSU** CS(%)	CV(%)	**K4 → CR** mAP(%)
MS-TCT [10]	F-T	RGB	33.7	-	25.4
Baseline-I [46]	Linear	Skeleton	18.2	11.4	6.1
LAC [48]	Linear	Skeleton	20.8	18.3	14.3
DuoCLR	Linear	Skeleton	24.1	19.8	19.1
Baseline-II [16]	F-T	Skeleton	28.3	20.3	21.8
LAC [48]	F-T	Skeleton	**36.8**	23.1	28.0
DuoCLR	F-T	Skeleton	35.1	**26.3**	**33.9**

Table 2. Experimental comparison in the composite skeleton-based action segmentation task.

4.2. Results

In the multi-class action segmentation task, we use NTU RGB+D for pretraining while using two datasets, PKU1 and PKU2, as the target for evaluation. A similar setting is considered in composite action segmentation where the trimmed TS and K4 are used as the pretraining domain while TSU and CR are used for evaluation. As with previous works, we use Linear Evaluation [24], where $\mathcal{F}$ is fixed while training $\mathcal{C}$, and Fine-tune (F-T) Evaluation [24], where $\mathcal{F}$ and $\mathcal{C}$ are trained jointly, to evaluate our approach. We report the experimental results in Table 1 and Table 2.

• **Comparison to State-of-the-art Methods.** We first

Models	Methods	5% of PKU1		10% of PKU1	
		CS(%)	CV(%)	CS(%)	CV(%)
Baseline-I	Linear	33.9	36.0	41.6	42.7
DuoCLR	Linear	58.4	41.3	70.0	72.8
Baseline II	Fine-Tune	47.4	49.5	56.4	58.1
LAC[48]	Fine-Tune	73.9	75.4	79.8	81.1
DuoCLR	Fine-Tune	**74.6**	**77.5**	**81.9**	**84.2**

Table 3. Semi-supervised evaluation ($mAP_{0.1}$) with a low proportion of samples from the target domain.

Model	PKU1	PKU2	TSU	Charade
DuoCLR	63%	59%	55%	42%

Table 4. The ratio of labeled data is saved while reaching the same performance with the vanilla model that was trained on 100% data.

compare the proposed DuoCLR to recent state-of-the-art methods[13, 14, 27, 47, 48] that involve a similar "trimmed-to-untrimmed" learning scenario. Experimental results demonstrate that DuoCLR outperforms most state-of-the-art comparatives across both multi-class (Table 1) and composite (Table 2) action segmentation tasks. We conclude that, since prior methods are limited in exploiting sequence-independent representations, the resultant feature space preserves a low transferability to action segmentation. In contrast, our model benefits from two efficient surrogate tasks, leading to better performance.

• **Linear and Fine-tune Evaluation.** A vanilla supervised pretraining paradigm involves learning a feature space $\mathcal{F}$via a conventional end-to-end action recognition model [46] using trimmed skeleton videos without data augmentation. We treat such a paradigm as a baseline model (Baseline-I in Table 1 and Table 2) while making comparison with DuoCLR in Linear Evaluation. Experimental results demonstrate that our model significantly outperforms Baseline-I, which validates the effectiveness of the proposed DuoCLR approach in learning expressive action representations for the action segmentation tasks. We also compare DuoCLR with Baseline-II, an action segmentation model [16] that can be trained end-to-end on untrimmed datasets. DuoCLR, when fine-tuned, achieves substantial performance improvements over Baseline-II, as shown in Table 1 and Table 2 for both multi-class and composite tasks. This outcome highlights the advantage of incorporating DuoCLR's pre-trained features, which enhance the model's adaptability and performance in segmentation.

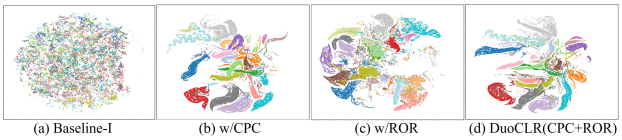
• **Semi-supervised Evaluation** The last experiment examines the transferability of our model while using fewer data samples in evaluation. Specifically, the experiment varies the ratios (5% and 10%) of samples from the target domain involved in ransfer learning (i.e., Linear and Fine-tune Evaluation). Experimental results in Table 3 demonstrate that the model's performance tends to increase monotonically along with the ratio of samples from the target domain. Table 4 suggests that DuoCLR can save up to 63% labeled data to reach the compatible performance compared to full-supervised learning.

NTU to PKU1	Method	CS (%)	CV (%)
Baseline-I	Linear	56.0	60.3
CPC	Linear	83.9	85.1
ROR	Linear	76.7	79.5
CPC+ROR (DuoCLR)	Linear	**85.2**	**84.7**

Table 5. Ablation study of two surrogate tasks: CPC (Cross Permutation Contrasting), ROR (Relative Order Reasoning), and combined DuoCLR. Results are reported as $mAP_{0.1}$.

PKU1 to PKU2	Methods	Pretraining	CS(%)	CV(%)
Baseline-II	Fine-Tune	-	69.3	66.0
DuoCLR	Linear	Trimmed PKU1	**78.1**	76.0
DuoCLR	Fine-Tune	Trimmed PKU1	78.0	**77.6**

Table 6. Experimental results (Acc) of pretraining using untrimmed datasets.



(a) Baseline-I (b) w/CPC (c) w/ROR (d) DuoCLR(CPC+ROR)

Figure 5. Feature clusters visualization by t-SNE [40]

4.3. Ablation studies

• **Effectiveness of Surrogate Tasks.** To assess the independent and combined effectiveness of these tasks, we conduct experiments using each task individually and in combination, pretraining on NTU and evaluating on PKU1. First, either surrogate task demonstrates a significant performance improvement compared to the vanilla supervised pretraining paradigm (Baseline-I) as summarized in Table 5. In Fig. 5b and Fig. 5c, CPC and ROR demonstrate their independent effectiveness at learning more separate feature clusters compared to Baseline-I (Fig. 5a). Second, CPC achieves superior effectiveness over ROR likely because it directly enforces instance-level similarity, which is critical for clustering. Last, DuoCLR achieves its best learning result when combining the two tasks, i.e. 85.2% and 84.7% on the benchmarks of CS and CV, meanwhile demonstrating the best separability in feature clustering (Fig. 5d).

• **Pretraining with Untrimmed Datasets.** It is also effective to apply DuoCLR to an untrimmed dataset for representation learning. In this experiment, we consider PKU1 as the pretraining dataset where we trim each sequence into separate single-action sequences using their associated frame annotations. By applying "Shuffle and Warp" to these trimmed sequences, DuoCLR learns a feature extractor, $\mathcal{F}$, which is then evaluated on the original untrimmed PKU2 dataset. As shown in Table 6, DuoCLR outperforms the vanilla end-to-end supervised model (Baseline-II) in both linear and fine-tune evaluation protocols, showcasing the benefit of pretraining.

NTU to PKU1	Shear, Crop	Shuffle	Warp	Supervised	CS%
DuoCLR		√	√	√	85.2
DuoCLR-I	√			√	58.3
DuoCLR-II		√		√	73.8
DuoCLR-unsup	√	√	√		81.6
DuoCLR-III	√	√	√	√	**87.1**

Table 7. Experimental comparison of DuoCLR and its variants using different skeleton data augmentation strategies.

Encoder	Classifier	Method	Params	Previous	Ours
GCN	Sequence-wise $\hat{\mathcal{C}}$	F-T	837K	63.5 [7]	**77.6**
GCN	Frame-wise $\mathcal{C}$	Linear	13.3K	61.8 [48]	**81.9**
GCN+TCN	Frame-wise $\mathcal{C}$	Linear	13.3K	-	**85.2**

Table 8. Experimental comparison between two types of classifiers. For $\hat{\mathcal{C}}$, the sliding window size is set as 300.

• **Effectiveness of Data Augmentation.** We learn a new model (DuoCLR-I) by incorporating two classical skeleton augmentations, "Shear" and "Crop" [24], with NTU used for pretraining and PKU1 for evaluation. As shown in Table 7, DuoCLR-I achieves an accuracy of 58.3%, suggesting that traditional augmentations alone provide limited transferability for action segmentation tasks. Our "Shuffle and Warp" method, however, is scalable to previous augmentations as we observe an orthogonal performance increase (1.3% over regular DuoCLR in PKU1) with the model DuoCLR-III after combining two types of augmentation. Additionally, we perform an ablation study on the roles of "Shuffle" and "Warp" individually. The DuoCLR-II variant, which uses only "Shuffle" (Eq. (2)), achieves an accuracy of 73.8%. When compared to DuoCLR, which uses both "Shuffle" and "Warp", the results demonstrate the importance of "Warp" which reduces irrelevant variations.

• **Pretraining with Unlabeled Datasets.** DuoCLR also supports unsupervised pretraining. For this approach, labeled proxies are created using sequence subscripts within each data batch. To learn an unsupervised DuoCLR model (DuoCLR-unsup, Table 7), we complement the augmentation with "Shear" and "Crop" to compensate for the lack of explicit labels. As seen in Table 7, regular DuoCLR outperforms the unsupervised variant (81.6% vs. 85.8% accuracy). This performance gap is likely due to the challenge of accurately identifying negative pairs in an unsupervised setting, which can lead to noise to the learning process.

• **DuoCLR is general to different network structures.** Fig. 4 illustrates two network structures commonly used in skeleton action segmentation: 1) Siding-window-free structure (Fig. 4b) composed of a GCN encoder $\hat{\mathcal{F}}$ and a frame-wise classifier $\mathcal{C}$; 2) sliding window based structure (Fig. 4c) composed of a $\hat{\mathcal{F}}$ and a sequence-wise classifier $\hat{\mathcal{C}}$ while using sliding windows to pre-segment input sequences. Fig. 4a illustrates the network proposed in Sec. 3.3. For evaluation, we train three DuoCLR models with different structures and conclude the experimental results in Table 8). It demonstrates that DuoCLR is effective for different structures while suppassing previous works with clear margins.

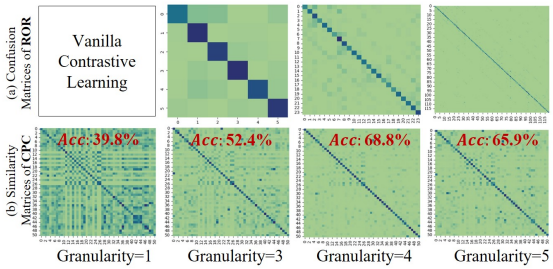


Figure 6. The confusion matrices obtained by ROR on different action granularities.

• **Effectiveness of Action Granularity.** In this study, we treat the action granularity ($\mathcal{G}$) of action permutations as a hyperparameter and learn how it affects optimizations of DuoCLR. In practice, setting a high value of $\mathcal{G}$ will introduce diverse action contexts, however, it can also lead to high computational complexity, e.g., $o(\mathcal{G})!$ for ROR while the size of the FC layer grows. In experiments, we increase $\mathcal{G}$ from 1 to 5 while preserving other implementations parameters as defined in Table 1. We illustrate confusion matrices (showing how ROR is good at mapping permutations) and similarity matrices (showing how CPC is good at learning consistent features) in Fig. 6, where Acc values evaluate the transferability of the resultant (CPC+ROR) network on PKU 2. While $\mathcal{G} = 1$, it degrades to a vanilla contrastive learning model [18]. While $\mathcal{G}$ increases, each model is capable of learning at different action granularity and and the best model is achieved when $\mathcal{G}$ is around 4.

5. Conclusions

This paper introduces a dual-surrogate contrastive learning (DuoCLR) framework while exploiting fine-grained extraction on action representations for skeleton data-based human action segmentation. It applies "Shuffle and Warp" to generate various multi-action permutations where two surrogate tasks, Cross Permutation Contrasting and Relative Order Reasoning, learn permutation-invariant intra-class similarities and permutation-aware inter-class contexts. The learned action representations can be represented by a feature extractor which can be entirely transferred into action segmentation. Experimental results demonstrate that 1) DuoCLR enables training with both unlabeled and labeled trimmed skeleton videos; 2) DuoCLR outperforms state-of-the-art methods in multiple tasks; 3) DuoCLR can save 42%-63% frame-level labeled data to reach the compatible fully-supervised performance in action segmentation; 4) DuoCLR supports learning with RGB videos that can lead to better action segmentation performance compared to previous RGB based approaches.

6. Acknowledgments

This research was supported in part by MITACS Accelerate and NSERC Discovery grants. The authors would like to express their sincere gratitude to Dr. Pierre Payeur for his invaluable contributions to this work. Dr. Payeur provided consistent supervision throughout the research and played a key role in project administration and funding acquisition. His thoughtful input during the review and editing process was instrumental to the successful completion of this work. The authors also gratefully acknowledge the collaboration of the Sensing and Machine Vision for Automation and Robotic Intelligence (SMART) laboratory and Spectronix Inc.

References

[1] Mohamed Abdelfattah, Mariam Hassan, and Alexandre Alahi. Maskclr: Attention-guided contrastive learning for robust action representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18678–18687, 2024. 1, 3

[2] Shurong Chai, Rahul Kumar Jain, Jiaqing Liu, Shiyu Teng, Tomoko Tateyama, Yinhao Li, and Yen-Wei Chen. A motion-aware and temporal-enhanced spatial–temporal graph convolutional network for skeleton-based human action segmentation. *Neurocomputing*, 580:127482, 2024. 2

[3] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9454–9463, 2020. 2

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 3

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 3, 5

[6] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13359–13368, 2021. 1, 2

[7] Yuxiao Chen, Long Zhao, Jianbo Yuan, Yu Tian, Zhaoyang Xia, Shijie Geng, Ligong Han, and Dimitris N Metaxas. Hierarchically self-supervised transformer for human skeleton representation learning. In *European Conference on Computer Vision*, pages 185–202. Springer, 2022. 2, 3, 6, 8

[8] Zhan Chen, Hong Liu, Tianyu Guo, Zhengyan Chen, Pinhao Song, and Hao Tang. Contrastive learning from spatio-temporal mixed skeleton sequences for self-supervised skeleton-based action recognition. *arXiv preprint arXiv:2207.03065*, 2022. 3

[9] Seunggeun Chi, Hyung-gun Chi, Qixing Huang, and Karthik Ramani. Infogcn++: Learning representation by predicting the future for online skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2, 3

[10] Rui Dai, Srijan Das, Kumara Kahatapitiya, Michael S Ryoo, and François Brémond. Ms-tct: Multi-scale temporal convtransformer for action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20041–20051, 2022. 6

[11] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2533–2550, 2022. 2, 3, 6

[12] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 833–842, 2019. 1, 3, 6

[13] Jianfeng Dong, Shengkai Sun, Zhonglin Liu, Shujie Chen, Baolong Liu, and Xun Wang. Hierarchical contrast for unsupervised skeleton-based action representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 525–533, 2023. 1, 3, 4, 7

[14] Haodong Duan, Mingze Xu, Bing Shuai, Davide Modolo, Zhuowen Tu, Joseph Tighe, and Alessandro Bergamo. Skeletr: Towards skeleton-based action recognition in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13634–13644, 2023. 2, 7

[15] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3584, 2019. 4, 6

[16] Benjamin Filtjens, Bart Vanrumste, and Peter Slaets. Skeleton-based action segmentation with multi-stage spatial-temporal graph convolutional neural networks. *IEEE Transactions on Emerging Topics in Computing*, 12(1):202–212, 2022. 2, 4, 6, 7

[17] Luca Franco, Paolo Mandica, Bharti Munjal, and Fabio Galasso. Hyperbolic self-paced learning for self-supervised skeleton-based action representations. *arXiv preprint arXiv:2303.06242*, 2023. 1, 3, 4

[18] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 762–770, 2022. 1, 2, 3, 4, 5, 6, 8

[19] Jinhua Hu, Yonghong Hou, Zihui Guo, and Jiajun Gao. Global and local contrastive learning for self-supervised skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 3

[20] Xiaohu Huang, Hao Zhou, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Jingdong Wang, Xinggang Wang, Wenyu Liu, and Bin Feng. Graph contrastive learning for skeleton-based action recognition. *arXiv preprint arXiv:2301.10900*, 2023. 1, 3, 4, 6

[21] Haoyu Ji, Bowen Chen, Xinglong Xu, Weihong Ren, Zhiyong Wang, and Honghai Liu. Language-assisted skeleton action understanding for skeleton-based temporal action segmentation. In *European Conference on Computer Vision*, pages 400–417. Springer, 2025. 2, 3, 6

[22] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6

[23] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE international conference on computer vision*, pages 667–676, 2017. 5

[24] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 3d human action representation learning via cross-view consistency pursuit. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4741–4750, 2021. 1, 3, 6, 8

[25] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3595–3603, 2019. 2, 4, 5

[26] Yunheng Li, Zhongyu Li, Shanghua Gao, Qilong Wang, Qibin Hou, and Ming-Ming Cheng. A decoupled spatio-temporal framework for skeleton-based action segmentation. *arXiv preprint arXiv:2312.05830*, 2023. 2

[27] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2490–2498, 2020. 3, 5, 6, 7

[28] Lilang Lin, Jiahang Zhang, and Jiaying Liu. Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2363–2372, 2023. 1, 2, 3, 4, 5, 6

[29] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017. 1, 3, 6

[30] Hanchao Liu, Yuhe Liu, Tai-Jiang Mu, Xiaolei Huang, and Shi-Min Hu. Skeleton-cutmix: Mixing up skeleton with probabilistic bone exchange for supervised domain adaptation. *IEEE Transactions on Image Processing*, 2023. 3

[31] Hao Ma, Zaiyue Yang, and Haoyang Liu. Fine-grained unsupervised temporal action segmentation and distributed representation for skeleton-based human motion analysis. *IEEE Transactions on Cybernetics*, 52(12):13411–13424, 2021. 2

[32] Yunyao Mao, Wengang Zhou, Zhenbo Lu, Jiajun Deng, and Houqiang Li. Cmd: Self-supervised 3d action representation learning with cross-modal mutual distillation. In *European Conference on Computer Vision*, pages 734–752. Springer, 2022. 6

[33] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 527–544. Springer, 2016. 3, 5

[34] Anshul Shah, Aniket Roy, Ketul Shah, Shlok Mishra, David Jacobs, Anoop Cherian, and Rama Chellappa. Halp: Hallucinating latent positives for skeleton-based self-supervised learning of actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18846–18856, 2023. 1

[35] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 1, 3, 6

[36] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016. 6

[37] Yukun Su, Guosheng Lin, and Qingyao Wu. Self-supervised 3d skeleton action representation learning with motion consistency and continuity. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13328–13338, 2021. 3

[38] Fida Mohammad Thoker, Hazel Doughty, and Cees GM Snoek. Skeleton-contrastive 3d action representation learning. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1655–1663, 2021. 1, 3, 4, 6

[39] Haitao Tian and Pierre Payeur. Stitch, contrast, and segment: Learning a human action segmentation model using trimmed skeleton videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7365–7373, 2025. 3

[40] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 7

[41] Wanjiang Weng, Hongsong Wang, Junbo He, Lei He, and Guosen Xie. Usdrl: Unified skeleton-based dense representation learning with multi-grained feature decorrelation. *arXiv preprint arXiv:2412.09220*, 2024. 6

[42] Cong Wu, Xiao-Jun Wu, Josef Kittler, Tianyang Xu, Sara Ahmed, Muhammad Awais, and Zhenhua Feng. Scd-net: Spatiotemporal clues disentanglement network for self-supervised skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5949–5957, 2024. 1, 3

[43] Linhua Xiang and Zengfu Wang. Joint mixing data augmentation for skeleton-based action recognition. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 3

[44] Wangmeng Xiang, Chao Li, Yuxuan Zhou, Biao Wang, and Lei Zhang. Generative action description prompts for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10276–10285, 2023. 2

[45] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10334–10343, 2019. 5

[46] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 1, 2, 4, 6, 7

[47] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and François Brémond. Unik: A unified framework for real-world skeleton-based action recognition. *arXiv preprint arXiv:2107.08580*, 2021. 1, 2, 7

[48] Di Yang, Yaohui Wang, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Lac-latent action composition for skeleton-based action segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13679–13690, 2023. 1, 2, 3, 5, 6, 7, 8

[49] Jiahang Zhang, Lilang Lin, and Jiaying Liu. Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3427–3435, 2023. 1, 2, 3

[50] Jiahang Zhang, Lilang Lin, and Jiaying Liu. Prompted contrast with masked motion modeling: Towards versatile 3d action representation learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7175–7183, 2023. 2, 6

[51] Jiahang Zhang, Lilang Lin, and Jiaying Liu. Shap-mix: shapley value guided mixing for long-tailed skeleton based action recognition. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 1688–1696, 2024. 3

[52] Nenggan Zheng, Jun Wen, Risheng Liu, Liangqu Long, Jianhua Dai, and Zhefeng Gong. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 3

[53] Yujie Zhou, Haodong Duan, Anyi Rao, Bing Su, and Jiaqi Wang. Self-supervised action representation learning from partial spatio-temporal skeleton sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3825–3833, 2023. 3

[54] Yisheng Zhu, Hu Han, Zhengtao Yu, and Guangcan Liu. Modeling the relative visual tempo for self-supervised skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13913–13922, 2023. 1, 3