SCD-Net: Spatiotemporal Clues Disentanglement Network for Self-Supervised Skeleton-Based Action Recognition

Cong Wu[1,2], Xiao-Jun Wu[1*], Josef Kittler[2], Tianyang Xu[1], Sara Ahmed[2], Muhammad Awais[2], Zhenhua Feng[2]

[1]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China
[2]School of Computer Science and Electronic Engineering, and CVSSP, University of Surrey, Guildford GU2 7XH, UK
congwu@stu.jiangnan.edu.cn, wu_xiaojun@jiangnan.edu.cn; tianyang_xu@163.com; {sara.atito, muhammad.awais, j.kittler, z.feng}@surrey.ac.uk

Abstract

Contrastive learning has achieved great success in skeleton-based action recognition. However, most existing approaches encode the skeleton sequences as entangled spatiotemporal representations and confine the contrasts to the same level of representation. Instead, this paper introduces a novel contrastive learning framework, namely Spatiotemporal Clues Disentanglement Network (SCD-Net). Specifically, we integrate the decoupling module with a feature extractor to derive explicit clues from spatial and temporal domains respectively. As for the training of SCD-Net, with a constructed global anchor, we encourage the interaction between the anchor and extracted clues. Further, we propose a new masking strategy with structural constraints to strengthen the contextual associations, leveraging the latest development from masked image modelling into the proposed SCD-Net. We conduct extensive evaluations on the NTU-RGB+D (60&120) and PKU-MMD (I&II) datasets, covering various downstream tasks such as action recognition, action retrieval, transfer learning, and semi-supervised learning. The experimental results demonstrate the effectiveness of our method, which outperforms the existing state-of-the-art (SOTA) approaches significantly. Our code and supplementary material can be found at https://github.com/cong-wu/SCD-Net.

1 Introduction

Skeleton-based action recognition focuses on identifying human actions via skeleton sequences, which has witnessed significant advancements in recent years. On one hand, deep networks, such as Graph Convolutional Network (GCN) (Yan, Xiong, and Lin 2018), have been investigated and successfully applied for the task at hand. On the other hand, several large-scale datasets, *e.g.*, NTU-RGB+D (Shahroudy et al. 2016), have been proposed, providing an experimental foundation for further development of the area. However, like most visual tasks, the training of a high-performance model typically requires a massive amount of high-quality labelled data. This requirement poses a significant challenge in data collection and annotation. Fortunately, self-supervised learning has emerged as a solution to address this challenge by leveraging inherent associations instead of relying on annotations. In particular, recent investigations (Dong et al. 2023) have demonstrated that contrastive learning, owing to its interpretability and transferability, has emerged as a front-runner in self-supervised skeleton-based action recognition.

However, several crucial aspects are disregarded by existing approaches. First, the encoder is responsible for mapping the input into a latent space where the contrast can be conducted. While most previous methods (Zhang et al. 2022a; Franco et al. 2023) concentrate on obtaining unified information through commonly used spatiotemporal modelling networks. Their designs result in the complete entanglement of information, failing to provide clear indications for subsequent contrastive measures. There have been sporadic attempts (Dong et al. 2023) aiming to extract absolutely isolated spatial or temporal information. But repeated evidence has shown that complete isolation of spatiotemporal information is suboptimal for action recognition (Kay et al. 2017; Lin, Gan, and Han 2019). More importantly, most approaches focus on constructing contrast pairs at same level of representation (Guo et al. 2022) during optimisation; Or attempt to force the interaction between information flows, overlooking the gap between domains (Dong et al. 2023). In addition, existing techniques (Thoker et al. 2021) often limit themselves to scale transformation, which results in not fully capitalising on the potential of data augmentation. Here we introduce a novel contrastive learning framework that focuses on disentangling spatiotemporal clues, and exploits masking in data augmentation to provide more discriminative inputs, thereby prompting the model to learn more robust interactions.

To leverage the intricate features present in skeleton sequences, we propose a dual-path decoupling encoder to generate explicit representations from spatial and temporal domains. Our encoder comprises two main subsystems: a feature extractor and a decoupling module. The role of the feature extractor is to extract fundamental spatiotemporal features from skeleton sequences as the intermediate representations. Since lacking an overall grasp of the skeleton sequence, it is difficult to obtain a picture of the features simply by modelling from a certain perspective. Next, we generate token embeddings by projection and refine the sequence features with a transformer-based module. The decoupling modules are instrumental to deriving disentangled

*Corresponding Author.

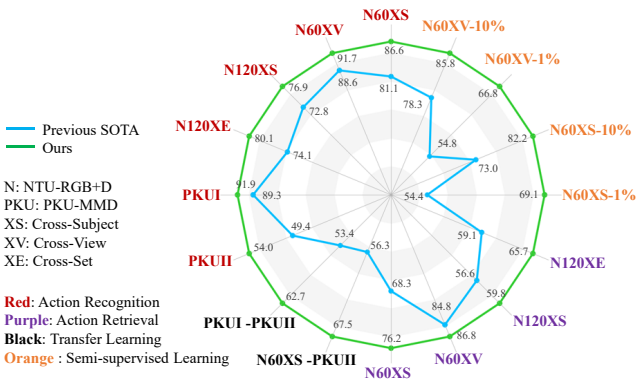


Figure 1: A comparison of the proposed method with HiCo-Transformer (Dong et al. 2023), using multiple evaluation metrics. (Better view in colour.)

joint-based and frame-based representations, leading to enhanced interpretability of the learned representations.

The principle underlying contrastive learning lies in achieving that the encoded query exhibits similarity with its corresponding key while showing dissimilarity with other keys from the backup queue (He et al. 2020). Here we extend contrastive loss to measure discrimination among representations of multiple spatiotemporal granularities. We strategically incorporate a global view of spatiotemporal representation as an anchor and evaluate its correlation with other representations obtained from an alternate encoder. To elaborate further, we fuse and project the cues derived from the encoder into the contrastive space to create a global representation. Our aim is to establish a bridge that facilitates the interaction of information across different domains through the utilisation of this anchor.

Furthermore, to prompt the model to learn more robust interactions, we propose an innovative mask-based data augmentation in a structurally-constrained manner. Specifically, we mask the adjacent area of the randomly selected region in the spatial domain and construct cube-based random masking in the temporal domain. This structured masking strategy serves to significantly increase the variety of training data. Moreover, it enables the model to implicitly capture spatiotemporal contextual relationships within the skeleton sequence.

We perform extensive experiments to demonstrate the effectiveness of the proposed method. As shown in Figure 1, the results indicate that our approach surpasses the mainstream methods in all downstream tasks, demonstrating its superior capabilities in skeleton-based action understanding.

2 Related Work

2.1 Skeleton-based Action Recognition

Skeleton-based action recognition has garnered significant attention by the research community (Ke et al. 2017; Gupta et al. 2021; Wu et al. 2023). In earlier approaches (Du, Wang, and Wang 2015; Chen et al. 2006), customised techniques were employed to classify skeletons via traditional feature extraction methods. Recently, GCN-based approaches (Yan, Xiong, and Lin 2018; Li et al. 2019; Liu et al. 2020) have gained prominence in the field. The general paradigm initially models the skeleton sequence as a spatiotemporal graph and subsequently employs information aggregation and updating techniques. Inspired by the notable achievements of transformer (Dosovitskiy et al. 2020; Liu et al. 2022), some recent methods (Zhang et al. 2021, 2022b) have explored its powerful sequence modelling capability for skeleton-based tasks.

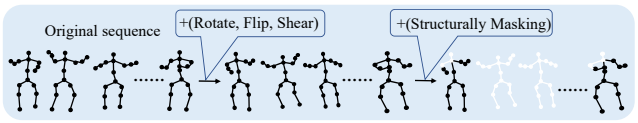
2.2 Contrastive Learning

Contrastive learning is a typical solution for self-supervised learning. Unlike generative learning (Zhu et al. 2020; Huang et al. 2022), contrastive learning does not involve explicit generation or reconstruction of the input. Instead, it focuses on learning discriminative representations through a contrastive loss. Most contrastive learning methods (Chen et al. 2020; Grill et al. 2020) operate on the principle of pulling positive pairs closer to each other, while simultaneously pushing dissimilar pairs farther apart within a projection space. By exploring the internal properties within the data, contrastive learning enables learning more generalised and robust representations, resulting in remarkable performance on downstream tasks (Wang and Liu 2022).

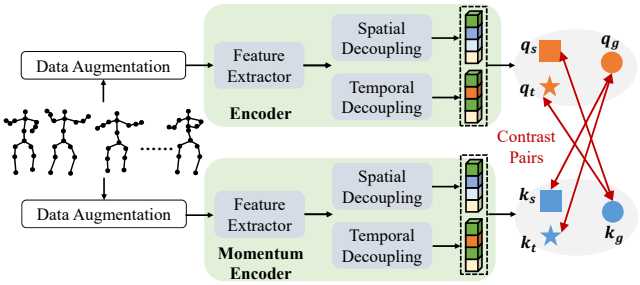
2.3 Contrastive Learning for Skeleton-based Action Recognition

Contrastive learning has also been successfully employed in skeleton-based action recognition. Thoker *et al.* (Thoker et al. 2021) proposed the intra-skeleton and inter-skeleton contrastive loss, achieving promising results in several downstream tasks. Dong *et al.* (Dong et al. 2023) utilised down-sampling operations at different stages of the encoder to obtain multi-scale features for constructing a hierarchical contrastive learning framework. Franco *et al.* (Franco et al. 2023) proposed a novel approach that involves projecting the encoded features into a hyperbolic space, which is a non-Euclidean space that allows more efficient modelling of complex association. Despite these advances, most existing studies overlook the crucial step of extracting and disentangling spatial and temporal clues from skeleton sequences, not to mention the failure of considering the interactions among representations of different domains.

For contrastive learning, data augmentation processes the training sample to obtain positive input pairs with certain differences. Thoker *et al.* (Thoker et al. 2021) used various spatiotemporal augmentation techniques, including pose augmentation, joint jittering, and temporal crop-resize, to generate different inputs for the query and key encoders. While most methods follow similar scale transformation paradigms, Zhou *et al.* (Zhou et al. 2023) proposed a strategy of masking selected nodes and frames, which greatly extends the augmentation to ”destroy” the data structure. However, unlike image data, skeleton sequences have strong physical associations, meaning that even if a certain node or frame is corrupted, it can easily be corrected using the information from adjacent areas (Cheng et al. 2020). Incorporating the structural constraints, we expand the point-based



(a) Data augmentation.



(b) The framework of SCD-Net.

Figure 2: Our model benefits from three innovations: a dual-path encoder for distinct spatiotemporal information decoupling; a bespoke cross-domains contrastive loss promoting the information interaction; a structurally-constrained masking strategy for efficient data augmentation.

masking approach to area-based masking. This extension aims to prevent potential data leakage and enhance the learning capabilities of SCD-Net.

3 The Proposed SCD-Net

In this section, we will initially present the overall framework of SCD-Net, followed by a detailed introduction to each of its components in the subsequent sections.

3.1 The Overall Framework

The overall pipeline of the proposed method consists of two branches, as shown in Figure 2 (b). Each branch has the same components, including data augmentation and encoder. For any input data, we link the outputs obtained by the encoder and momentum encoder to form contrast pairs.

To elaborate further, the input of the network is defined as a sequence of human body key points, denoted as $\mathcal{X} \in \mathbb{R}^{C\times T\times V}$, where T is the length of the sequence, C is the physical coordinate defined in a 2D/3D space, V is the number of key points. In SCD-Net, we first apply data augmentation to generate the augmented views for the encoders. Second, for each encoder, we deploy feature extraction and (spatial/temporal) decoupling operations to generate spatial feature $\mathbf{z}_s \in \mathbb{R}^{C_2}$ and temporal feature $\mathbf{z}_t \in \mathbb{R}^{C_2}$ from the entangled information. Third, we project these clues into the same semantic space to obtain the final representations.

The loss function, $\mathcal{L}_{\theta,\xi}$, is defined as a measure of interactions of these representations. The parameters θ and ξ specify the architecture corresponding to the encoder and the momentum encoder. During the optimisation, the loss is back-propagated only through the encoder, while the parameters of the momentum encoder are updated using a momentum update strategy. So the final optimiser is:

$$\boldsymbol{\theta} \leftarrow \text{optimizer}(\boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\xi}}, r), \boldsymbol{\xi} = \boldsymbol{\xi} * m + \boldsymbol{\theta} * (1-m), \quad (1)$$

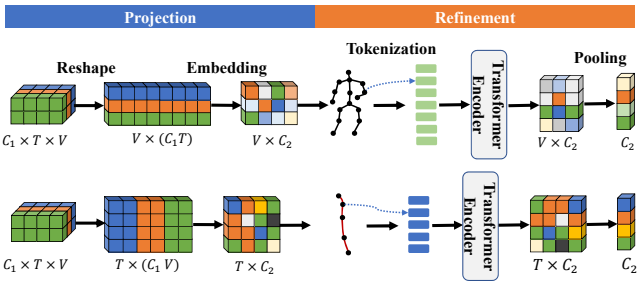


Figure 3: The dual-path decoupling module that provides clean spatial and temporal representations of a skeleton sequence.

where r and m are the learning rate and decay rate.

3.2 The Dual-path Decoupling Encoder

In general, the features extracted from a skeleton sequence are characterised as complex spatiotemporal associations describing an action. However, we argue that this paradigm is not suitable for contrastive learning. As the information is greatly entangled, it is difficult to provide clear guidance for the subsequent comparison. In SCD-Net, we advocate a dual-path decoupling encoder to extricate clear and multiple discriminative cues from the complex sequence information. Such clues provide clear instructions for a subsequent contrast quantification. More importantly, a reliable assessment of the contrast between different domains is likely to provide stronger discrimination.

For brevity, we generally denote the augmented input for the encoder as $\mathcal{X}$. As demonstrated by the existing studies, completely isolating the information flow is sub-optimal (Lin, Gan, and Han 2019; Wang et al. 2021). Given that, we apply a spatiotemporal modelling network to extract the intermediate features. Inspired by the excellent performance in modelling skeleton sequences (Yan, Xiong, and Lin 2018), we use a l_g-layer GCN, consisting of spatial-GCN (S-GCN) and temporal GCN (T-GCN), to obtain unified representations $\mathcal{Y} \in \mathbb{R}^{C_1\times T\times V}$. This can be expressed as a process of aggregation and updating of adjacent features. Specifically, for any $\mathcal{X}_{ti} \in \mathbb{R}^{C}$, where t and i are the frame and joint index, the newly generated features $\mathcal{Y}_{ti} \in \mathbb{R}^{C_1}$ can be expressed as:

$$\mathcal{Y}_{ti} = \sum_{\mathcal{X}_{uj}\in \mathrm{B}(\mathcal{X}_{ti})} \frac{1}{\mathrm{Z}_{ti}(\mathcal{X}_{uj})} \cdot \mathcal{X}_{uj} \cdot \mathrm{w}(\mathrm{l}_{ti}(\mathcal{X}_{uj})), \quad (2)$$

where $\mathrm{B}(\mathcal{X}_{ti})$ denotes the kernel of the graph convolution operation on $\mathcal{X}_{ti}$, $\mathrm{Z}(\cdot)$ represents normalisation, $\mathrm{w}(\cdot)$ is the weight function, and $\mathrm{l}(\cdot)$ maps adjacent nodes to the corresponding subset index.

Given the intermediate spatiotemporal representation $\mathcal{Y}$, the following step is decoupling operation, which involving projection and refinement, as shown in Figure 3. Specifically, we perform a dimension transformation on $\mathcal{Y}$ to derive $\mathcal{Y}_{rs} \in \mathbb{R}^{V\times(C_1T)}$ and $\mathcal{Y}_{rt} \in \mathbb{R}^{T\times(C_1V)}$. These transformed representations are then projected to higher semantic space to obtain the corresponding spatial and temporal

embeddings. For instance, the spatial embedding operation is defined as:

$$\mathcal{Y}_s = \mathcal{W}_{s2} * \mathrm{ReLU}(\mathcal{W}_{s1} * \mathcal{Y}_{rs} + \mathcal{B}_{s1}) + \mathcal{B}_{s2}, \quad (3)$$

where $\mathcal{W}$ and $\mathcal{B}$ are the trainable weights and bias, $\mathcal{Y}_s \in \mathbb{R}^{V \times C_2}$. However, the current embedding is still a rough representation as the current features lack explicit interactions within points or frames. While the feature extraction operation incorporates significant spatiotemporal interactions, these interactions often become intertwined. Hence, it remains crucial to address the interaction of individual spatial and temporal embeddings. Here we use a l_t-layer self-attention network to construct the self-correlation information extraction process that refines the spatial and temporal representations, as shown in Figure 3. The transformer architecture used in this method has two main components: self-attention and feed-forward modules. For instance, we obtain $\boldsymbol{z_s} \in \mathbb{R}^{C_2}$ as follows:

$$\hat{\mathcal{Z}}_{si} = \mathrm{SoftMax}\left[\frac{\mathrm{F}_{qi}(\mathcal{Y}_s) \cdot (\mathrm{F}_{ki}(\mathcal{Y}_s))^T}{\sqrt{d_i}}\right] \cdot (\mathrm{F}_{vi}(\mathcal{Y}_s)),$$

$$\hat{\mathcal{Z}}_s = \mathrm{LN}(\mathrm{F}_c(\mathrm{Concat}[\hat{\mathcal{Z}}_{s1}, .., \hat{\mathcal{Z}}_{si}, ..., \hat{\mathcal{Z}}_{sh}]) + \mathcal{Y}_s),$$

$$\mathbf{z}_s = \mathrm{MaxPooling}(\mathrm{LN}[\mathrm{FFN}(\hat{\mathcal{Z}}_s) + \hat{\mathcal{Z}}_s]), \quad (4)$$

where F represents feature projection, implemented by fully connected layer, with Concat denoting the concatenation operation, LN and FFN means layer normalisation and Feed-Forward Networks, h signifying the number of heads. $\boldsymbol{z_t} \in \mathbb{R}^{C_2}$ can also be obtained by similar operations.

3.3 Cross-domain Contrastive Loss

With the decoupled spatial and temporal representations, as shown in Figure 2, we first obtain the final representations by:

$$\mathbf{q}_s = \mathrm{F}_s(\mathbf{z}_s), \quad \mathbf{q}_t = \mathrm{F}_t(\mathbf{z}_t), \quad (5)$$

where F_s, F_t are the corresponding projection functions, which can be defined by two fully connected layers, similar to Eq. (3). As we discussed earlier, there is an obvious gap between spatial and temporal domains, for which we introduce a global perspective $\mathbf{q}_g$ compatible with both as a intermediary for contrasts.

$$\mathbf{q}_g = \mathrm{F}_g[\mathbf{z}_t, \mathbf{z}_s], \quad (6)$$

where F_g are the corresponding projection function. The outputs ($\mathbf{k}_s$, $\mathbf{k}_t$, $\mathbf{k}_g$) of the corresponding key encoder, can also be obtained by a similar process.

Based on these candidate features, we define a new cross-domain loss. The core of our design lies in anchoring the global representation and building its association with other representations obtained by another encoder. The loss function is defined as,

$$\begin{aligned}\mathcal{L}_{\theta,\xi} \triangleq \lambda_1 \cdot \mathcal{L}(\mathbf{q}_g, \mathbf{k}_s) + \lambda_2 \cdot \mathcal{L}(\mathbf{q}_g, \mathbf{k}_t) + \\ \lambda_3 \cdot \mathcal{L}(\mathbf{q}_s, \mathbf{k}_g) + \lambda_4 \cdot \mathcal{L}(\mathbf{q}_t, \mathbf{k}_g),\end{aligned} \quad (7)$$

where λ is the mixing weight of the sum operation. Specifically, for any given contrast pair $\mathbf{u}$ and $\mathbf{v}$, $\mathcal{L}(\mathbf{u}, \mathbf{v})$ evaluates the correlation between $\mathbf{u}$ and $\mathbf{v}$. The objective is to minimise the distance between positive pairs from the query and key encoders, while maximising the distance from the other features.

To achieve this, we employ the contrastive loss based on InfoNCE (Oord, Li, and Vinyals 2018) as follows:

$$\mathcal{L}(\mathbf{u}, \mathbf{v}) = -\log \frac{\mathrm{h}(\mathbf{u}, \mathbf{v})}{\mathrm{h}(\mathbf{u}, \mathbf{v}) + \sum_{\mathbf{m} \in M} \mathrm{h}(\mathbf{u}, \mathbf{m})}, \quad (8)$$

where $\mathrm{h}(\mathbf{u}, \mathbf{v}) = \exp(\mathbf{u} \cdot \mathbf{v}/\tau)$ is the exponential similarity measurement. We denote the first-in-first-out queue of the previously extracted features, containing l_m negative samples, by M.

3.4 Data Augmentation

By imposing structural constraints, our approach applies the masking operation within a local region around the current randomly selected joints or frames instead of relying only on isolated points or frames. In this way, we substantially eliminate explicit local contextual associations, and force the encoders to model robust contextual relationships through interactive contrastive learning.

Structurally Guided Spatial Masking Considering the physical structure of skeleton, when a certain joint is selected for masking, we simultaneously mask the points in its adjacent area. Let us represent the adjacency relationship using the matrix $\mathbf{P}$. $\mathbf{P}_{ij} = 1$, if joints i and j are connected, otherwise $\mathbf{P}_{ij} = 0$. We denote $\mathbf{D} = \mathbf{P}^n$, where n is the exponent. The element $\mathbf{D}_{ij}$ in $\mathbf{D}$ represents the number of paths that can be taken to reach node j from node i by walking n steps. Note that reversal and looping are allowed. To impose a structural constraint, when node i is selected, we perform the same augmentation operation on all nodes j for which $\mathbf{D}_{ij} \neq 0$. The only undesirable artefact of this operation is that it may give rise to a variable number of candidate joints. To avoid this, for several randomly selected nodes, the actual augmentation is applied only to a fixed number (k) of points exhibiting the highest overall response on $\mathbf{D}$.

Cube-based Temporal Masking The sequence follows a linear relationship in time. To avoid information leakage between adjacent frames (Tong et al. 2022), we construct a cube, defined by a selected segment and its adjacent frames. Specifically, we start by dividing the input sequence into s cubes of equal length. Next, we randomly select r cubes as the candidates for masking.

We denote the data augmentation candidates as $\mathcal{T}$. Given a skeleton sequence $\mathcal{X}$, the augmented view is obtained by:

$$\mathcal{X}^a \triangleq \mathrm{t}_n(\mathcal{X}_n^a, p_n) \cdots \triangleq \mathrm{t}_1(\mathcal{X}_1^a, p_1), \quad (9)$$

where $\mathcal{X}_1^a = \mathcal{X}, t_1, \ldots, t_n \sim \mathcal{T}$, and if $p = \mathrm{False}$, t degenerates into an identity map.

4 Experiments

4.1 Experimental Settings

Datasets We evaluate the proposed method on four benchmarking datasets, NTU-RGB+D (60&120) (Shahroudy et al. 2016) and PKU-MMD (I&II) (Liu et al. 2017).

Method	Encoder	NTU-60		NTU-120		PKU-MMD I	PKU-MMD II
		x-sub	x-view	x-sub	x-setup	x-sub	x-sub
Encoder-decoder							
LongT GAN (AAAI'18)	GRU	52.1	56.4	-	-	67.7	26.5
EnGAN-PoseRNN (WACV'19)	LSTM	68.6	77.8	-	-	-	-
H-Transformer (ICME'21)	Transformer	69.3	72.8	-	-	-	-
SeBiReNet (ECCV'20)	GRU	-	79.7	-	-	-	-
Colorization (ICCV'21)	GCN	75.2	83.1	-	-	-	-
GL-Transformer (ECCV'22)	Transformer	76.3	83.8	66.0	68.7	-	-
Hybrid learning							
MS^{2}L (ACMMM'21)	GRU	52.6	-	-	-	64.9	27.6
PCRP (TMM'21)	GRU	54.9	63.4	43.0	44.6	-	-
Contrastive-learning							
CrosSCLR (CVPR'21)	GCN	72.9	79.9	-	-	84.9	21.2
AimCLR (AAAI'22)	GCN	74.3	79.7	63.4	63.4	87.8	38.5
ISC (ACMMM'21)	GRU&CNN&GCN	76.3	85.2	67.1	67.9	80.9	36.0
HYSP (ICLR'23)	GCN	78.2	82.6	61.8	64.6	83.8	-
SkeAttnCLR (IJCAI'23)	GCN	80.3	86.1	66.3	<u>74.5</u>	87.3	<u>52.9</u>
ActCLR (CVPR'23)	GCN	80.9	86.7	69.0	70.5	-	-
HiCo-Transformer (AAAI'23)	Transformer	<u>81.1</u>	<u>88.6</u>	<u>72.8</u>	74.1	<u>89.3</u>	49.4
SCD-Net (Ours)	GCN&Transformer	**86.6**	**91.7**	**76.9**	**80.1**	**91.9**	**54.0**

Table 1: A comparison of the proposed method with the mainstream methods in action recognition. Bold and underlined fonts indicate the highest and second highest results, respectively.

Implementation Details For the input data, 64 frames are randomly selected for training and evaluation. We perform data augmentation operations, including rotate, flip and shear, as well as the proposed structural spatial masking and temporal masking, on the selected sequence. Each operation has a 50% chance of being executed. For masking, we set $n = 2$, $k = 8$, $s = 16$, $r = 6$. For the encoder, we refer to MoCo (He et al. 2020) and build a query encoder and the corresponding key encoder. The two encoders have exactly the same structure as shown in Figure 3. For feature extractor, we borrow the structure from CTR-GCN (Chen et al. 2021) as the basic operation. For network optimisation, we set the queue length of M to 8192 (except 2048 for PKU-MMD I), moco momentum to 0.999, softmax temperature to 0.2, and λ to 1.

More details are presented in supplementary material.

4.2 Comparison with the SOTA Methods

We compare SCD-Net with several SOTA methods, including: (1) Encoder-decoder based methods: LongT GAN (Zheng et al. 2018), EnGAN-PoseRNN (Kundu et al. 2019), P&C (Su et al. 2020), H-Transformer (Cheng et al. 2021), SeBiReNet (Nie, Liu, and Liu 2020), Colorization (Yang et al. 2021), GL-Transformer (Kim et al. 2022); (2) Hybrid learning based methods: ASSL (Si et al. 2020), MS^{2}L (Lin et al. 2020), PCRP (Xu et al. 2021), HI-TRS (Chen et al. 2022); (3) Contrastive-learning based methods: CrosSCLR (Li et al. 2021), MCC (Su et al. 2021), AimCLR (Guo et al. 2022), ISC (Thoker et al. 2021), HYSP (Franco et al. 2023), SkeAttnCLR (Hua et al. 2023), ActCLR (Lin, Zhang, and Liu 2023), HiCo-Transformer (Dong et al. 2023). To evaluate the merits of the proposed SCD-Net, we construct multiple downstream tasks, including action recognition, action retrieval, transfer learning and semi-supervised learning.

Method	NTU-60		NTU-120	
	x-sub	x-view	x-sub	x-setup
LongT GAN	39.1	48.1	31.5	35.5
P&C	50.7	76.3	39.5	41.8
AimCLR	62.0	-	-	-
ISC	62.5	82.6	50.6	52.3
SkeAttnCLR	<u>69.4</u>	76.8	46.7	58.0
HiCo-Transformer	68.3	<u>84.8</u>	<u>56.6</u>	<u>59.1</u>
SCD-Net (Ours)	**76.2**	**86.8**	**59.8**	**65.7**

Table 2: A comparison with the mainstream methods in action retrieval.

Action Recognition Here we adopt the linear evaluation method, which involves fixing the pre-trained parameters and training only a fully connected layer for label prediction. Table 1 presents a comparison of our approach with other SOTA methods on several popular datasets. The results demonstrate that our method outperforms all the existing approaches by a large margin. Specifically, we achieve 5.5% and 3.1% improvements over the previous best method on NTU-60 x-sub and x-view, respectively. On NTU-120, our approach surpasses the previous SOTA by 4.1% and 5.6% on x-sub and x-set, respectively. Again, SCD-Net achieves 91.9% on PKU-MMD I and 54.0% on PKU-MMD II, which are much higher than the existing SOTA results.

Action Retrieval Referring to (Thoker et al. 2021), we use the KNeighbors classifier (Cover and Hart 1967) for action retrieval while keeping all the pre-trained parameters fixed. As reported in Table 2, our SCD-Net achieves promising results on the NTU-60's x-sub and x-view datasets, with the accuracy of 76.2% and 86.8%, respectively. Additionally, on the NTU-120's x-sub and x-set datasets, our method attains the accuracy of 59.8% and 65.7%, surpassing all the existing

Method	Transfer to PKU-MMD II	
	PKU-MMD I	NTU-60
LongT GAN	43.6	44.8
MS^{2}L	44.1	45.8
IS	45.1	45.9
HiCo-Transformer	53.4	56.3
SCD-Net (Ours)	**62.7**	**67.5**

Table 3: A comparison with the mainstream methods in transfer learning.

Method	x-sub		x-view	
	1%	10%	1%	10%
LongT GAN	35.2	62.0	-	-
MS^{2}L	33.1	65.2	-	-
ASSL	-	64.3	-	69.8
ISC	35.7	65.9	38.1	72.5
MCC	-	60.8	-	65.8
Colorization	48.3	71.7	52.5	78.9
CrosSCLR	-	67.6	-	73.5
HI-TRS	-	70.7	-	74.8
GL-Transformer	-	68.6	-	74.9
HiCo-Transformer	54.4	73.0	54.8	78.3
SCD-Net (Ours)	**69.1**	**82.2**	**66.8**	**85.8**

Table 4: A comparison with the mainstream methods in semi-supervised learning.

methods by a significant margin.

Transfer Learning For transfer learning, follow (Dong et al. 2023), we apply the knowledge representation learned from one domain to another domain. Specifically, we load the pre-trained parameters from the PKU-MMD I and NTU-60 datasets respectively, and fine-tune the model on the PKU-MMD II dataset, following the cross-subject evaluation protocol. The results presented in Table 3 demonstrate that our SCD-Net brings a performance improvement of 9.3% and 11.2%, as compared with the current SOTA results.

Semi-supervised Learning For semi-supervised learning, we first load the pre-trained parameters and then fine-tune the entire network on a partially labelled training set. In our experiment, we randomly select limited of labelled samples from the NTU-60 dataset for further training. The results in Table 4 show that even when only 1% of the labels are available, our method achieves the accuracy of 69.1% and 66.8% on x-sub and x-view, respectively. With 10% of the labelled data available, the performance of our model is further improved to 82.2% and 85.8%.

4.3 Ablation Study

In this part, we verify all the innovative components of the proposed SCD-Net. All the experimental results are focused on the action recognition task using cross-subject evaluation on the NTU-60 dataset.

The Decoupling Encoder The primary role of our novel encoder is to extract crucial spatial and temporal representations. In Table 5, when we discard the feature extractor, the performance drops a lot. This shows that the way of extracting completely isolated information flow is not feasible in the current task, which is also in line with our expectation. It is worth noting that using non-shared feature extractors for the two branches leads to better performance than using a shared one. When we attempt to discard decouple module, compared with the default setting, the accuracy is decreased from 86.6% to 63.7% as the output is impacted by the spatiotemporal entanglement. This situation improves after converting spatiotemporal representations into temporal and spatial domain-specific embeddings, resulting in an accuracy of 84.0%. However, it was still inferior to the design with the refinement model. This is because refinement provides powerful sequence modelling capabilities, thereby refining the current rough representations.

Extractor	Decoupling	Accuracy	
		Top-1	Top-5
None	Projection & Refinement	80.0	94.9
Non-Shared	Projection & Refinement	**86.6**	**97.6**
Shared	Projection & Refinement	85.2	96.0
Non-Shared	None	63.7	89.6
Non-Shared	Projection	84.0	97.1

Table 5: Ablation experiments with the decoupling encoder.

GCN		Trans			Accuracy
Layer	Channel	Layer	Head	Channel	Top-1
2	64	1	8	2048	85.4
3	**64**	**1**	**8**	**2048**	**86.6**
4	64	1	8	2048	86.6
3	32	1	8	2048	86.4
3	128	1	8	2048	85.8
3	64	2	8	2048	86.4
3	64	1	4	2048	86.2
3	64	1	16	2048	85.7
3	64	1	8	1024	85.7
3	64	1	8	4096	85.1

Table 6: The details of the encoder. The bold part in black indicates the best performance.

Encoder Parameters In Table 6, we investigate the impact of the parameter settings on the model performance. Overall, the optimal performance is achieved when we use a 3-layer GCN block, 64 as the number of output channels, and set the transformer with 1 layer, 8 heads, and 2048 output channels. The results also demonstrate that changing the parameters does not significantly affect the model's performance, indicating the stability of our approach. Additionally, we can see that the network size does not necessarily improve the performance, suggesting that it is not dependent on the network size.

Loss Function We report the results of different configurations of the loss function in Table 7. We can see that the interactive loss performs better than the traditional instance loss, leading to 0.7% and 1.6% performance boost. When using all three granularities jointly, the model achieves optimal performance. This is because there is a significant gap between the nature of the video information conveyed by

Granularity	Loss Type	Accuracy	
		Top-1	Top-5
S-T	Instance Loss	84.6	97.1
S-T	Interactive Loss	85.3	97.2
S-T-G	Instance Loss	85.0	97.5
S-T-G	Interactive Loss	**86.6**	**97.6**
S-T-G	Interactive & Instance Loss	86.5	97.6

Table 7: A comparison of different loss functions. 'S', 'T', 'G' represent Spatial, Temporal and Global representations.

Data Augmentation Strategy		Accuracy	
Conventional	Masking	Top-1	Top-5
None	None	70.1	91.1
✓	None	85.4	97.4
✓	Ours	**86.6**	**97.6**
✓	Random	85.6	97.4
✓	Spatial Only	86.2	97.5
✓	Temporal Only	83.7	96.7
None	Ours	76.0	93.6

Table 8: A comparison of different data augmentation strategies.

the spatial and temporal features, although they describe the same action. The global anchor provides more comprehensive representations, which bridge this gap and enhances the discriminative ability. It is worth noting that the use of both loss functions jointly does not improve the performance further. This could be attributed to the fact that the supervisory information across the information flow already provides adequate guidance, and further guidance mechanisms are unnecessary.

Data Augmentation Here we investigate the impact of different data augmentation strategies on the model performance. The results are reported in Table 8. Without any augmentation, the performance drops by more than 16%, compared to the default setting. When using only the conventional augmentation methods, including rotation, flipping, and shearing, the model achieves an accuracy of 85.4%. After introducing the proposed structurally guided spatiotemporal augmentation, the performance of the model increases 1.2% further. Even with random masking, the performance is still lower than the default setting.

It is worth noting that discarding either spatial or temporal masking leads to a performance degradation. Also, when only masking is used, the performance of the model is mediocre, even far worse than using only the conventional data augmentation methods. That is because our method performs a compensation, instead of replacement. A proper masking further improves the diversity of the input data and promotes the model in learning more robust spatiotemporal context associations. When combining all these techniques, the performance of the model is the best.

Visualisation of The Decoupled Clues As shown in Figure 4, we use t-SNE (Van der Maaten and Hinton 2008) to analyse the decoupled clues from SCD-Net. We select three groups of data with different emphases for comparison. The

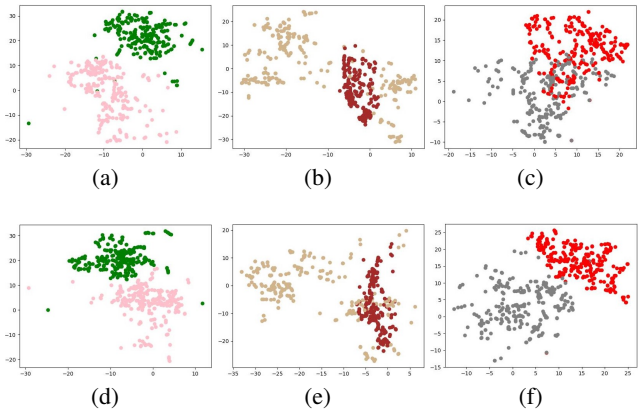


Figure 4: Visualisation of decoupled clues. Spatial clue: (a) 'throw' vs 'clapping'; (b) 'brush teeth' vs 'brush hair'; (c) 'drop' vs 'pick up'. Temporal clue: (d) 'throw' vs 'clapping'; (e) 'brush teeth' vs 'brush hair'; (f) 'drop' vs 'pick up';

first row represents the spatial clue and the second one is the temporal clue. We can notice that from (a) and (d), 'throw' and 'clapping' have great separability on spatially and temporally. From (b) and (e), 'brush teeth' vs 'brush hair' are more separable on spatial domain, because the most significant difference is the object. According to (c) and (f), 'drop' and 'pick up' are more separable on temporal domain, while showing certain entanglement on the spatial domain, as they are in reverse order in temporally. More importantly, the results demonstrate that our encoder successfully decouples the corresponding features, which makes the specificity between different cues correspond to the same samples.

5 Conclusion

In this paper, we presented a new contrastive learning framework for unsupervised skeleton-based action recognition. The key innovation is the design of spatiotemporal clue extraction mechanism. In the proposed method, we first used a spatiotemporal modelling network to encode an action sequence, followed by a decoupling module for obtaining pure spatial and temporal representations. An cross-domain loss was proposed to guide the learning of discriminative representations conveyed by different representations. The training of the system was facilitated by a novel data augmentation method tailored for the proposed unsupervised learning framework. This method imposes structural constraints on action data perturbations to enhance the efficacy of contextual modelling and increase the diversity of the data. Extensive experimental results obtained on widely used benchmarking datasets demonstrated the merits of the proposed method that defines a new SOTA of the area.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (62020106012, 62332008, 62106089, U1836218), and National Key Research and Development Program of China (2023YFF1105102,

2023YFF1105105), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. KYCX22_2299), and the EPSRC Grants (EP/R018456/1, EP/V002856/1).

References

Chen, H.-S.; Chen, H.-T.; Chen, Y.-W.; and Lee, S.-Y. 2006. Human action recognition using star skeleton. In *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*, 171–178.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607.

Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; and Hu, W. 2021. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13359–13368.

Chen, Y.; Zhao, L.; Yuan, J.; Tian, Y.; Xia, Z.; Geng, S.; Han, L.; and Metaxas, D. N. 2022. Hierarchically self-supervised transformer for human skeleton representation learning. In *European Conference on Computer Vision*, 185–202. Springer.

Cheng, K.; Zhang, Y.; Cao, C.; Shi, L.; Cheng, J.; and Lu, H. 2020. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, 536–553. Springer.

Cheng, Y.-B.; Chen, X.; Chen, J.; Wei, P.; Zhang, D.; and Lin, L. 2021. Hierarchical Transformer: Unsupervised Representation Learning for Skeleton-Based Human Action Recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.

Cover, T.; and Hart, P. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1): 21–27.

Dong, J.; Sun, S.; Liu, Z.; Chen, S.; Liu, B.; and Wang, X. 2023. Hierarchical contrast for unsupervised skeleton-based action representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 525–533.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, Y.; Wang, W.; and Wang, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1110–1118.

Franco, L.; Mandica, P.; Munjal, B.; and Galasso, F. 2023. HYperbolic Self-Paced Learning for Self-Supervised Skeleton-based Action Representations. *arXiv preprint arXiv:2303.06242*.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33: 21271–21284.

Guo, T.; Liu, H.; Chen, Z.; Liu, M.; Wang, T.; and Ding, R. 2022. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 762–770.

Gupta, P.; Thatipelli, A.; Aggarwal, A.; Maheshwari, S.; Trivedi, N.; Das, S.; and Sarvadevabhatla, R. K. 2021. Quo vadis, skeleton action recognition? *International Journal of Computer Vision*, 129(7): 2097–2112.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Hua, Y.; Wu, W.; Zheng, C.; Lu, A.; Liu, M.; Chen, C.; and Wu, S. 2023. Part Aware Contrastive Learning for Self-Supervised Action Recognition. *arXiv preprint arXiv:2305.00666*.

Huang, C.; Wen, J.; Xu, Y.; Jiang, Q.; Yang, J.; Wang, Y.; and Zhang, D. 2022. Self-supervised attentive generative adversarial networks for video anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; and Boussaid, F. 2017. A new representation of skeleton sequences for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3288–3297.

Kim, B.; Chang, H. J.; Kim, J.; and Choi, J. Y. 2022. Global-local motion transformer for unsupervised skeleton-based action learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, 209–225. Springer.

Kundu, J. N.; Gor, M.; Uppala, P. K.; and Radhakrishnan, V. B. 2019. Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1459–1467. IEEE.

Li, L.; Wang, M.; Ni, B.; Wang, H.; Yang, J.; and Zhang, W. 2021. 3D Human Action Representation Learning via Cross-View Consistency Pursuit. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4741–4750.

Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; and Tian, Q. 2019. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3595–3603.

Lin, J.; Gan, C.; and Han, S. 2019. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7083–7093.

Lin, L.; Song, S.; Yang, W.; and Liu, J. 2020. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2490–2498.

Lin, L.; Zhang, J.; and Liu, J. 2023. Actionlet-Dependent Contrastive Learning for Unsupervised Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2363–2372.

Liu, C.; Hu, Y.; Li, Y.; Song, S.; and Liu, J. 2017. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3202–3211.

Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; and Ouyang, W. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 143–152.

Nie, Q.; Liu, Z.; and Liu, Y. 2020. Unsupervised 3d human pose representation with viewpoint and pose disentanglement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, 102–118. Springer.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1010–1019.

Si, C.; Nie, X.; Wang, W.; Wang, L.; Tan, T.; and Feng, J. 2020. Adversarial self-supervised learning for semi-supervised 3d action recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, 35–51. Springer.

Su, K.; et al. 2020. Predict & cluster: Unsupervised skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9631–9640.

Su, Y.; Lin, G.; Sun, R.; Hao, Y.; and Wu, Q. 2021. Modeling the uncertainty for self-supervised 3d skeleton action representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, 769–778.

Thoker, F. M.; et al. 2021. Skeleton-contrastive 3D action representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1655–1663.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35: 10078–10093.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Wang, L.; Tong, Z.; Ji, B.; and Wu, G. 2021. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1895–1904.

Wang, Z.; and Liu, W. 2022. Robustness verification for contrastive learning. In *International Conference on Machine Learning*, 22865–22883. PMLR.

Wu, C.; Wu, X.-J.; Xu, T.; Shen, Z.; and Kittler, J. 2023. Motion Complement and Temporal Multifocusing for Skeleton-based Action Recognition. *IEEE transactions on circuits and systems for video technology*.

Xu, S.; Rao, H.; Hu, X.; Cheng, J.; and Hu, B. 2021. Prototypical contrast and reverse prediction: Unsupervised skeleton based action recognition. *IEEE Transactions on Multimedia*.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Yang, S.; Liu, J.; Lu, S.; Er, M. H.; and Kot, A. C. 2021. Skeleton cloud colorization for unsupervised 3d action representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13423–13433.

Zhang, H.; Hou, Y.; Zhang, W.; and Li, W. 2022a. Contrastive positive mining for unsupervised 3d action representation learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, 36–51. Springer.

Zhang, J.; Jia, Y.; Xie, W.; and Tu, Z. 2022b. Zoom transformer for skeleton-based group activity recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12): 8646–8659.

Zhang, Y.; Wu, B.; Li, W.; Duan, L.; and Gan, C. 2021. STST: Spatial-temporal specialized transformer for skeleton-based action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3229–3237.

Zheng, N.; Wen, J.; Liu, R.; Long, L.; Dai, J.; and Gong, Z. 2018. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhou, Y.; Duan, H.; Rao, A.; Su, B.; and Wang, J. 2023. Self-supervised Action Representation Learning from Partial Spatio-Temporal Skeleton Sequences. *arXiv preprint arXiv:2302.09018*.

Zhu, Y.; Min, M. R.; Kadav, A.; and Graf, H. P. 2020. S3vae: Self-supervised sequential vae for representation disentanglement and data generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6538–6547.