Prompted Contrast with Masked Motion Modeling: Towards Versatile 3D Action Representation Learning

Jiahang Zhang
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
zjh2020@pku.edu.cn

Lilang Lin
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
linlilang@pku.edu.cn

Jiaying Liu*
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
liujiaying@pku.edu.cn

ABSTRACT

Self-supervised learning has proved effective for skeleton-based human action understanding, which is an important yet challenging topic. Previous works mainly rely on contrastive learning or masked motion modeling paradigm to model the skeleton relations. However, the sequence-level and joint-level representation learning cannot be effectively and simultaneously handled by these methods. As a result, the learned representations fail to generalize to different downstream tasks. Moreover, combining these two paradigms in a naive manner leaves the synergy between them untapped and can lead to interference in training. To address these problems, we propose **P**rompted **C**ontrast with **M**asked **M**otion **M**odeling, PCM^3, for *versatile* 3D action representation learning. Our method integrates the contrastive learning and masked prediction tasks in a mutually beneficial manner, which substantially boosts the generalization capacity for various downstream tasks. Specifically, masked prediction provides novel training views for contrastive learning, which in turn guides the masked prediction training with high-level semantic information. Moreover, we propose a dual-prompted multi-task pretraining strategy, which further improves model representations by reducing the interference caused by learning the two different pretext tasks. Extensive experiments on *five* downstream tasks under three large-scale datasets are conducted, demonstrating the superior generalization capacity of PCM^3 compared to the state-of-the-art works. Our project is publicly available at: https://jhang2020.github.io/Projects/PCM3/PCM3.html.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**; **Computer vision representations**.

KEYWORDS

Skeleton-based action recognition, contrastive learning, masked modeling, self-supervised learning

*Corresponding author. This work is supported by the National Natural Science Foundation of China under contract No.62172020.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611774

ACM Reference Format:
Jiahang Zhang, Lilang Lin, and Jiaying Liu. 2023. Prompted Contrast with Masked Motion Modeling: Towards Versatile 3D Action Representation Learning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3611774

1 INTRODUCTION

Human activity understanding is a crucial problem in multi-media processing on account of its significant role in real-life applications, such as human-robotics interaction [18], healthcare [25] and autonomous driving [2]. As a highly efficient representation for human activity understanding, 3D skeletons represent the human form by 3D coordinates of key body joints. In comparison to other representations such as RGB videos and depth data, skeletons are lightweight, compact, and privacy-preserving. Owing to these competitive advantages, skeletons have become widely used in human action analysis.

Many efforts have been made on the supervised skeleton-based action learning [6, 9, 35, 47]. However, the performance of these methods heavily relies on huge amount of labeled data, which requires time-consuming and expensive data annotation work. This inherent shortcoming of full supervision limit their applications in the real world. Therefore, more and more attention has been paid to self-supervised 3D action representation learning recently to learn meaningful features from unlabeled data.

Self-supervised 3D action representation learning research has mainly focused on two paradigms: reconstruction-based and contrastive learning-based methods. Reconstruction-based methods leverage an encoder-decoder architecture to learn representations by predicting masked skeletons (*i.e.*, masked modeling) or reconstructing original data. These methods focus on joint-level feature modeling and capture spatial-temporal relationships. In contrast, contrastive learning-based methods use data augmentations to construct positive/negative pairs, and apply an instance discrimination task to learn sequence-level semantic features.

However, it is noticed that most recent representation learning methods focus on the single paradigm to model the joint-level (by masked skeleton modeling) [45, 53] or sequence-level (by contrastive learning) [8, 20, 26, 52] features solely. As a result, it is difficult for these methods to generalize well to different downstream tasks, *e.g.*, recognition task and motion prediction, because they cannot learn representations of different granularity simultaneously and effectively. Although some works [21, 40, 43] make valuable efforts to combine the above two approaches to learn richer

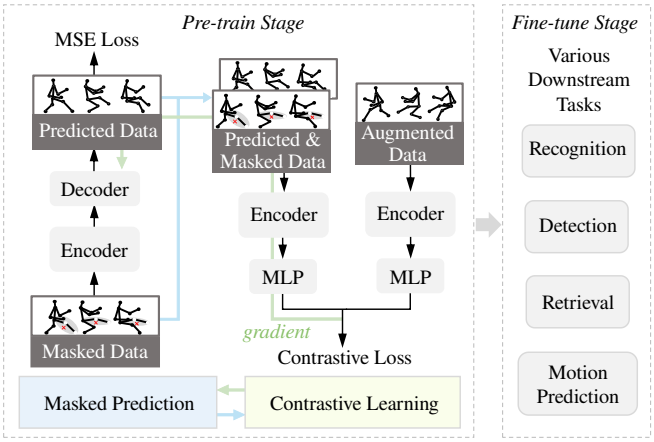

Figure 1: Illustration of the proposed method for versatile action representation learning. We integrate contrastive learning and masked skeleton modeling paradigms in a mutually beneficial manner. The masked prediction provides novel views for contrastive learning (blue arrow), and the generated gradients of contrast (green arrow) serve as high-level semantic guidance for masked prediction in turn, modeling both joint-level and sequence-level features.

representations, only mediocre improvement is observed. It is because simply combining them ignores the interference due to the gaps between feature modeling mechanisms of masked prediction and contrastive learning [30], and fails to utilize the potential synergy. These problems limit the generalization power of model, and versatile 3D action representation learning remains a challenging and under-explored area.

To this end, we propose the *prompted contrast with masked motion modeling*, PCM^3, which explores the mutual collaboration between the above two paradigms for versatile 3D action representation learning as shown in Figure 1. Specifically, the well-designed inter- intra- contrastive learning and topology-based masked skeleton prediction are first proposed as the basic pipelines. Furthermore, we connect the two tasks and explore the synergy between them. The views in masked prediction training are utilized as novel positive samples for contrastive learning. In turn, the masked prediction branch is also updated via the gradients from the contrastive learning branch for higher-level semantic guidance. Meanwhile, to reduce the distraction of learning between different pretext tasks and data views, we propose the dual-prompted multi-task pretraining strategy. Two types of prompts, namely, domain-specific prompts and task-specific prompts are applied to explicitly instruct the model to learn from different data views/tasks. Extensive experiments under *five* downstream tasks are conducted to provide a comprehensive evaluation. The proposed method demonstrates promising generalization capacity compared to state-of-the-art methods. Our contributions can be summarized as follows:

- We propose PCM^3 for multi-granularity representations, which integrates masked skeleton prediction and contrastive learning paradigms in a mutually beneficial manner. We employ the masked prediction network to generate more diverse positive motion views for contrastive learning. Meanwhile, the generated gradients are propagated and guide masked prediction learning in turn with high-level semantic information.
- Considering that different data views and pretraining tasks can cause mutual interference, we introduce domain-specific prompts and task-specific prompts for the multi-task pretraining. These trainable prompts enable the model to achieve more discriminative representations for different skeletons.
- We perform rigorous quantitative experiments to assess the generalization efficacy of state-of-the-art self-supervised 3D action representation learning techniques across five downstream tasks, including recognition, retrieval, detection, and motion prediction, on both uncorrupted and corrupted skeletons. Our study serves as a comprehensive benchmark for the research community, and we believe it can provide valuable insights and aid future investigation in this field.

2 RELATED WORKS

2.1 Skeleton-based Action Recognition

With the huge advances of deep learning, recurrent neural network (RNN)-based, convolutional neural network (CNN)-based, graph convolutional network (GCN)-based, and transformer-based methods are studied for skeleton-based action recognition. RNNs have been widely used to model temporal dependencies and capture the motion features for skeleton-based action recognition. The work in [9] uses RNN to tackle the skeleton as sequence data. Subsequently, Song *et al.* [37, 38] proposed to utilize the attention mechanism and multi-modal information to enhance the feature representations. Some other works [15, 24] transform each skeleton sequence into image-like representations and apply the CNN model to extract spatial-temporal information. Recently, GCN-based methods have attracted more attention due to the natural topology structure of the human body. Many works [6, 35, 47] apply GCN to the spatial and temporal dimension [47] and achieves remarkable results in the supervised skeleton-based action recognition. Meanwhile, transformer models [29, 36] also show promising results, owing to long-range temporal dependency learning by attention.

However, these supervised works rely on the huge labeled data to train the model. In this paper, we explore the self-supervised 3D action representation learning instead.

2.2 Contrastive Learning for Skeleton

Contrastive learning [3, 4, 14] has proven effective for skeleton representation learning. One popular research view is the skeleton augmentations, which are crucial for the learned representation quality. Guo *et al.* [12] explored the use of extreme augmentations in the current contrastive learning pipeline. Zhang *et al.* [52] proposed hierarchical consistent contrastive learning to utilize more strong augmentations. Another perspective is to explore the knowledge of different views in the skeleton. ISC [41] performs a cross-contrastive learning manner using image, graph and sequence representations. Li *et al.* [20] mined the potential positives resort to the different skeleton modalities, *i.e.*, joint, bone, motion, and re-weights training samples according to the similarity. Mao *et al.* [26] performed the mutual-distillation across different views. Different from the above works, we propose to integrate the masked modeling pretext task

with contrastive learning, modeling both joint-level and sequence-level features for more general representation learning.

2.3 Masked Image/Skeleton Modeling

Masked modeling has been explored in stacked denoising autoencoders [42], where the mask operation is regarded as adding noise to the original data. Recently, the masked modeling has achieved remarkable success in self-supervised learning [13, 46] for image representation learning. For skeleton data, LongT GAN [53] directly utilizes an autoencoder-based model optimized by an additional adversarial training strategy. Some works [21, 40] apply the motion prediction pretext task to learn the temporal dependencies in skeleton sequences. Inspired by the masked autoencoder [13], Wu *et al.* [45] proposed a masked skeleton autoencoder to learn the spatial-temporal relationships. In this paper, we explore the synergy between masked modeling and contrastive learning and propose a topology-based masking strategy to further boost the representation learning.

3 THE PROPOSED METHOD: PCM^3

In this part, we firstly describe our designed pipelines for contrastive learning (Section 3.1) as well as our proposed topology-based method for masked modeling (Section 3.2). Then, we further present the synergy exploration between the two pretext tasks in Section 3.3. The prompt-based pretraining strategy and the whole model are given in Section 3.4.

3.1 Skeleton Contrastive Learning

For clarity, the canonical design of contrastive learning for image/skeleton is given following previous works [1, 3, 14], which usually comprises the following components:

- **Data augmentation module** contains a series of manual data transformations to construct different views of original data, which are regarded as positive samples sharing the same semantic.
- **Encoder** $f(\cdot)$ serves as the mapping function from the input space to the latent feature space.
- **Embedding projector** $h(\cdot)$ is successively applied after encoder $f(\cdot)$, mapping the encoded feature into an embedding space where the self-supervised loss is applied.
- **Self-supervised loss** aims to maximize the similarity between positive samples, performing the feature clustering operation to obtain a distinguishable representation space.

Our contrastive learning design is based on MoCo v2 [4]. Specifically, we introduce intra- inter- skeleton transformations and relational knowledge distillation, to assist model to capture diverse motion patterns and boost the representation learning.

1) Intra-skeleton transformation learning. We utilize the following transformations: *Temporal crop-resize, Shear,* and *Joint Jittering* following the previous works [26, 41]. Specifically given a skeleton sequence x, the positive pair (s_{intra}, s') is constructed via the above transformations. Then, we obtain the corresponding feature representations (z_{intra}, z') via the query/key encoder $f_q(\cdot)/f_k(\cdot)$ and embedding projector $h_q(\cdot)/h_k(\cdot)$, respectively. Meanwhile, a memory queue $\mathbf{M}$ is maintained storing numerous negative samples for contrastive learning. We optimize the whole network by InfoNCE objective [28]:

$$\mathcal{L}_{Info}^{Intra} = -\log \frac{\exp(z_{intra} \cdot z'/\tau)}{\exp(z_{intra} \cdot z'/\tau) + \sum_{i=1} \exp(z_{intra} \cdot m_i/\tau)}, \quad (1)$$

where m_i is the feature in $\mathbf{M}$ corresponding to the i-th negative sample and τ is the temperature hyper-parameter. After a training step, the samples in a batch will be updated to $\mathbf{M}$ as negative samples according to a first-in, first-out policy. The key encoder is a momentum-updated version of the query encoder, *i.e.*, $\theta_k \leftarrow \alpha\theta_k + (1-\alpha)\theta_q$, where θ_q and θ_k are the parameters of query encoder and key encoder, and $\alpha \in [0, 1)$ is a momentum coefficient.

2) Inter-skeleton transformation learning. Inspired by the successful application of *Mix* augmentation in self-supervised learning [17, 19, 34, 49], we introduce the *CutMix* [48], *ResizeMix* [31], and *Mixup* [50] to our skeleton contrastive learning. These inter-skeleton transformations utilize two different samples to generate mixed augmented views. Specifically, given two skeleton sequences s_1, s_2, we randomly select a mixing method from the above and obtain the mixed skeleton data s_{inter} as follows:

- *Mixup* [50]: We interpolate the two skeleton sequences according to a sampled mixing ratio λ, *i.e.*, $s_{inter} = (1-\lambda)s_1 + \lambda s_2$.
- *CutMix* [48]: The randomly selected regions of two skeleton sequences are cut and pasted across the spatial-temporal dimension. And λ is defined as the ratio of replaced joint number to the total joint number.
- *ResizeMix* [31]: This is similar to *CutMix*, but downsamples s_2 first in the temporal dimension before mixing.

Details can be found in Supplementary Material. Subsequently, we can obtain the embeddings corresponding to the mixed data by $z_{inter} = h_q \circ f_q(s_{inter})$, and the following loss is optimized:

$$\mathcal{L}_{Info}^{Inter} = -\log \frac{\exp(z_{inter} \cdot z'_{inter}/\tau)}{\exp(z_{inter} \cdot z'_{inter}/\tau) + \sum_{i=1} \exp(z_{inter} \cdot m_i/\tau)}, \quad (2)$$

where $z'_{inter} = (1-\lambda)(h_k \circ f_k(s_1)) + \lambda(h_k \circ f_k(s_2))$.

3) Relational Knowledge Distillation. To further provide fine-grained semantic consistency supervision for contrastive learning, we introduce a relational knowledge self-distillation loss to positive pairs. Inspired by the works [26, 44, 52], the relational knowledge is modeled as the cosine similarity between z'/z'_{inter} and feature anchors in memory queue $\mathbf{M}$. The relational distribution, *i.e.*, the similarity with respect to negative anchor samples, is enforced to be consistent between each positive pair. Taking the embedding pair (z_{intra}, z') corresponding to the aforementioned intra-transformation as an example, the loss can be expressed as:

$$\begin{aligned} \mathcal{L}_{KL}^{Intra} &= -p\left(z', \tau_k\right) \log p\left(z_{intra}, \tau_q\right), \\ p_j\left(z, \tau\right) &= \frac{\exp(z \cdot m_j/\tau)}{\sum_{i=1} \exp(z \cdot m_i/\tau)}, \end{aligned} \quad (3)$$

where m_i is the stored i_{th} feature anchors in $\mathbf{M}$. τ_k and τ_q are temperature hyper-parameters, set as 0.05 and 0.1, respectively. This distillation term introduces more anchors to mine the fine-grained and semantics-aware similarity relations [44], boosting the representation quality.

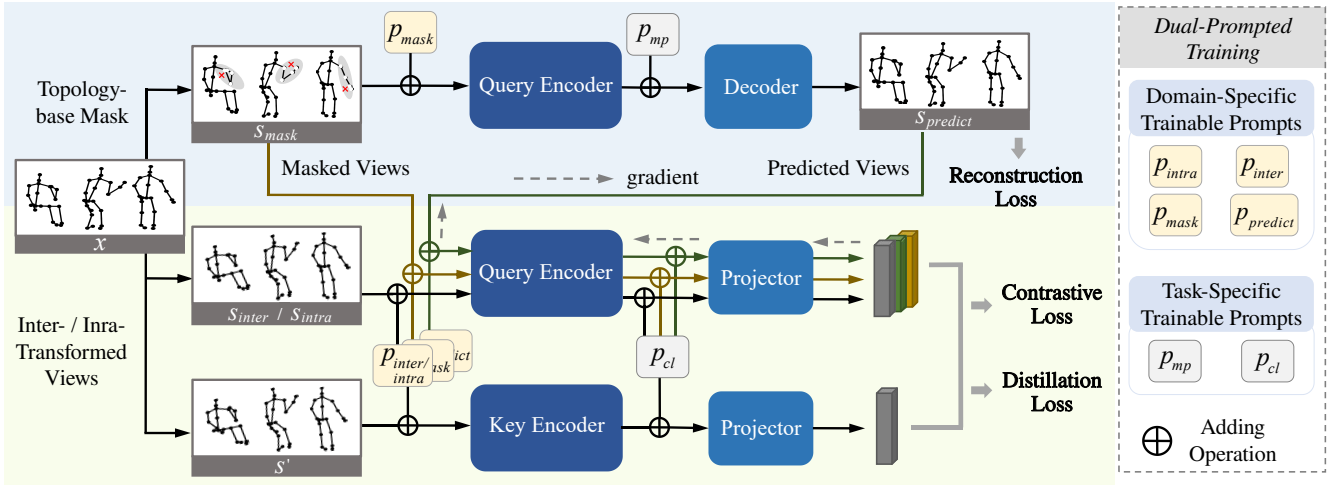

Figure 2: The overview of the proposed method. We integrate the masked skeleton prediction (*blue* part) and the contrastive learning (*yellow-green* part) paradigms in a mutually beneficial manner. For brevity, we represent intra- and inter- transformed views in a single branch in the diagram. The masked and the predicted views are utilized to expose more novel motion patterns for contrastive learning. Meanwhile, the gradients from contrastive learning (*dotted arrows* in figure) are propagated to the masked prediction branch to update the decoder. To further boost the representation learning from different views/tasks, we propose the dual-prompted multi-task pretraining strategy, where domain-specific and task-specific prompts are added in input-wise and feature-wise form, serving as training guidance.

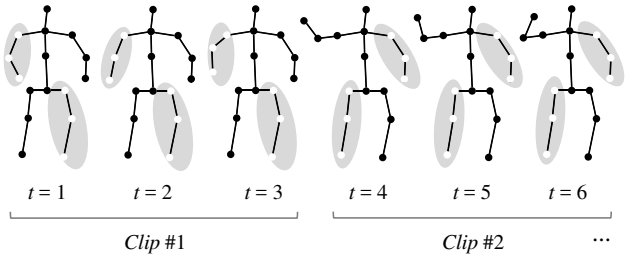

Figure 3: Illustration of the topology-based masking strategy. The gray region is the body parts to be masked.

3.2 Masked Skeleton Prediction

To further enrich the learned representations by model, we integrate masked skeleton modeling and the joint-level feature learning is baked into training process. This further improves generalization ability especially for dense prediction downstream tasks, as compared with using only instance-wise discrimination task, *i.e.*, contrastive learning.

First, in terms of the masking strategy, previous works [45, 53] utilize *Random Mask* to randomly select the masked joints in spatial-temporal dimension. However, given the redundancy of skeleton sequences, the masked joints can be easily inferred by copying adjacent joints in spatial or temporal dimension, which is not conducive to modeling meaningful relationships in skeletons. To this end, we propose *Topology-based* masking strategy, which masks the skeleton in the body-part level instead of the joint level, *i.e. trunk, right-hand, left-hand, right-leg* and *left-leg*. Meanwhile, we divide the sequences into different clips in temporal dimension, and the same parts are all masked in a clip, as shown in Figure 3.

Based on the above masking strategy, we mask the original skeleton x and then feed the masked skeleton s_{mask} into the encoder $f_q(\cdot)$ to obtain the corresponding features. To predict the masked skeleton region, we employ a decoder $dec(\cdot)$, which takes the encoded features as input and outputs the reconstructed skeleton. The MSE loss between original data x and predicted data $s_{predict}$ is optimized in the masked region:

$$\mathcal{L}_{Mask} = \frac{1}{N}\sum^{N} ||(x - dec \circ f_q(s_{mask})) \odot (1 - M)||_2, \quad (4)$$

where N is the number of all masked joints. M is the binary mask in which 1 and 0 correspond to the visible joints and the masked joints, respectively. And 1 is an all-one matrix with the same shape as M.

3.3 Collaboration between Contrastive Learning and Masked Modeling

Despite our proposed novel pipelines for contrastive learning and masked skeleton prediction, we find that simply integrating the two paradigms only yield mediocre performance gains as shown in Table 7. This is due to the inherent gap between feature modeling paradigm of the two tasks [30], and the model cannot directly take advantage of the potential synergy between them. Therefore, in this part, we explore the collaboration and connect the two tasks in a mutually beneficial manner.

1) Novel Positive Pairs as Connection. First, we utilize special data views in masked prediction training to provide more diverse positive samples for contrastive learning. Considering that the

masked skeleton naturally simulates occlusion for skeletons, we take masked skeleton s_{mask} views as challenging positives, to learn the underlying semantic consistency and enhance robustness to occlusion. Meanwhile, we also boost contrastive learning by taking the predicted skeletons output by decoder $dec(\cdot)$ as positive samples. Compared with the masked views, the predicted views contain the inherent noise, uncertainty, and diversity brought by continuous training of the model, which contributes to encoding more diverse movement patterns and thus improves generalization capacity.

In a nutshell, we utilize the masked view s_{mask} and predicted view $s_{predict}$ as positive samples to connect masked modeling with contrastive learning. We present all positive (embedding) pairs as follows:

$$\{(z_{intra}, z'), (z_{inter}, z'_{inter}), (z_{mask}, z'), (z_{predict}, z')\}. \quad (5)$$

They are obtained by the query/key encoder and projector, respectively. Each positive pair is applied to calculate Eq. 1 for contrastive loss and Eq. 3 for distillation loss. Taking (s_{mask}, s') as an example, it replaces the z_{intra} with $z_{mask} = h_q \circ f_q(s_{mask})$ for optimization. **Note** that we use $\mathcal{L}_{Con}$ and $\mathcal{L}_{KL}$ to represent the total contrastive loss and distillation loss, respectively, which comprise component losses in the form of Eq. 1 and Eq. 3 calculated for all four positive pairs defined in Eq. 5.

2) High-Level Semantic Guidance. On the other hand, the gradients of $s_{predict}$ from the contrastive learning branch are propagated to update the reconstructed decoder $dec(\cdot)$ as shown in Figure 2. It provides the high-level semantic guidance for the skeleton prediction together with the MSE loss in Eq. 4 which serves as joint-level supervision, leading to better masked prediction learning and higher quality of $s_{predict}$ as positive samples.

With the above synergetic designs, the masked prediction task provides novel positive samples as meaningful supplements to the contrastive learning. Meanwhile, with the gradients of contrastive learning propagating to the masked modeling branch, the masked prediction task can be conversely assisted via the high-level semantic guidance provided by contrastive learning task. These designs connect the two tasks and yield better representation quality.

3.4 Dual-Prompted Multi-Task Pretraining

For self-supervised pretraining, the whole model is optimized for contrastive learning and masked prediction tasks in a multi-tasking manner. However, the input data are from different distributions (domains), *e.g.*, augmented views and masked views, for different pretext tasks, *i.e.*, contrastive learning and masked prediction. Previous works directly feed them into the encoder to learn respective representations. This can cause ambiguity and interfere with feature modeling in terms of learning from different data/tasks.

To this end, we propose a novel dual-prompted multi-task pretraining strategy to explicitly instruct the model to learn from different domains/tasks. Specifically, two types of prompts named *domain-specific prompt* and *task-specific prompt* are designed, which are implemented as trainable vectors to provide training guidance.

1) Domain-Specific Prompt. To deal with different domains of input, we maintain domain-specific prompts for each input view, *i.e.*, $p_{inter}, p_{intra}, p_{mask}$, and $p_{predict}$, of which the dimension equals to the skeleton spatial size. Then, these domain-specific prompts are added to the corresponding input data (s_* means any view):

$$s_* = s_* + p_*. \quad (6)$$

These decorated skeletons are fed into the encoder for self-supervised pretraining. The trainable prompts enable the model to learn domain-specific knowledge and achieve better representations.

2) Task-Specific Prompt. For task-specific prompts, we apply deep-feature prompt after encoder instead of input-wise prompt to encourage the encoder to extract more general features for various tasks. After obtaining the representations $feat_* = f_q/f_k(s_*) \in \mathbb{R}^{\mathrm{d}}$, we add the task-specific prompts p_{cl} or p_{mp} to the $feat_*$. Specifically, $p_{cl}, p_{mp} \in \mathbb{R}^{\mathrm{r}}$ where the dimension $\mathrm{r} < \mathrm{d}$ for efficiency [11], are added to the randomly selected r-dimensional channels from the original feature $feat_*$. If the feature is to feed into the projector $h_q/h_k(\cdot)$ for contrastive learning, the p_{cl} is added, otherwise, the p_{mp} is added for masked prediction. These prompts can effectively learn task-specific knowledge and reduce interference between different pretext tasks.

Overall, the following objective is applied to the whole model as shown in Figure 2:

$$\mathcal{L} = \mathcal{L}_{Con} + \lambda_m \mathcal{L}_{Mask} + \lambda_{kl} \mathcal{L}_{KL}, \quad (7)$$

where the loss weight λ_m and λ_{kl} are set to 40.0, 1.0 in implementation. Note that the prompts are tuned only in the pretraining stage since they are targeted for self-supervised pretext tasks rather than downstream tasks. Therefore, we simply drop all prompts after the pretraining stage.

4 EXPERIMENTAL RESULTS

4.1 Datasets

1) NTU RGB+D 60 Dataset (NTU 60) [33]. There are 56,578 videos with 25 joints in each frame. 60 action categories are defined in the dataset. We adopt the following two evaluation protocols: a) Cross-Subject (xsub): the data for training and testing are collected from 40 different subjects. b) Cross-View (xview): the data for training and testing are captured in 3 different views: front view, 45 degrees view for left side and right side.

2) NTU RGB+D 120 Dataset (NTU 120) [22]. NTU 120 is an extension to NTU 60 dataset. There are 114,480 videos collected with 120 action categories included. Two recommended protocols are also adopted: a) Cross-Subject (xsub): the data for training and testing are collected from 106 different subjects. b) Cross-Setup (xset): the data for training and testing are collected from 32 different setups with different camera locations.

3) PKU Multi-Modality Dataset (PKUMMD) [23]. PKUMMD is a large-scale dataset towards multi-modality 3D understanding of human actions. In particular, PKUMMD supports the evaluation of action detection. The actions are organized into 51 action categories and almost 20,000 instances are included. The PKUMMD is divided into two subsets, Part I and Part II. We adopt cross-subject protocol following previous works.

4.2 Implementation Details

We follow the experiment settings of the recent works [26, 41]. For data preprocessing, all sequences of skeletons are downsampled to 300 frames and then crop-resized to 64 frames to feed the model.

Table 1: Comparison of unsupervised action recognition results.

Method	Year	Backbone	NTU 60		NTU 120		PKUMMD
			xsub (%)	xview (%)	xsub (%)	xset (%)	Part II (%)
Single-stream:							
Long TGAN [53]	AAAI'2018	GRU	39.1	48.1	-	-	26.0
MS^{2}L [21]	ACM MM'2020		52.6	-	-	-	27.6
CRRL [43]	TIP'2022		67.6	73.8	56.2	57.0	41.8
ISC [41]	ACM MM'2021		76.3	85.2	67.1	67.9	36.0
CMD [26]	ECCV'2022		79.8	86.9	70.3	71.5	43.0
HaLP [32]	CVPR'2023		79.7	86.8	71.1	72.2	43.5
H-Transformer [7]	ICME'2021	Transformer	69.3	72.8	-	-	-
GL-Transformer [16]	ECCV'2022		76.3	83.8	66.0	68.7	-
PCM^3 (Ours)	-	GRU	**83.9**	**90.4**	**76.5**	**77.5**	**51.5**
Three-stream:							
3s-HiCo [8]	AAAI'2023	GRU	82.6	90.8	75.9	77.3	-
3s-CMD[26]	ECCV'2022		84.1	90.9	74.7	76.1	52.6
3s-CrosSCLR [20]	CVPR'2021	GCN	77.8	83.4	67.9	66.7	21.2
3s-AimCLR [12]	AAAI'2022		78.9	83.8	68.2	68.8	38.5
3s-HiCLR [52]	AAAI'2023		80.4	85.5	70.0	70.4	53.8
3s-HYSP [10]	ICLR'2023		79.1	85.2	64.5	67.3	-
3s-SkeleMixCLR [49]	arXiv'2022		82.7	87.1	70.5	70.7	57.1
3s-CPM [51]	ECCV'2022		83.2	87.0	73.0	74.0	51.5
3s-PCM^3 (Ours)	-	GRU	**87.4**	**93.1**	**80.0**	**81.2**	**58.2**

Table 2: Performance comparison on NTU 60 under semi-supervised evaluation protocol.

Method	NTU 60			
	xview		xsub	
	1% data	10% data	1% data	10% data
LongT GAN [53]	-	-	35.2	62.0
MS^{2}L [21]	-	-	33.1	65.1
HiCLR [52]	50.9	79.6	51.1	74.6
ISC [41]	38.1	72.5	35.7	65.9
CMD [26]	53.0	80.2	50.6	75.4
PCM^3	**53.1**	**82.8**	**53.8**	**77.1**

We adopt a 3-layer Bi-GRU as the encoder backbone, of which the hidden dimension is set as $d = 1024$ following previous works. The task-prompt dimension r is 128. The MLPs are used as the projection heads, mapping features into embeddings with 128 dimensions. A 2-layer GRU with 512 dimensions is used as the decoder $dec(\cdot)$. 3s- denotes the fusion results of three streams, *i.e.* joint, bone and motion modalities of skeleton.

During self-supervised pre-training, the model is trained for 450 epochs in total, with a batch size of 128. The initial learning rate is 0.02 and is reduced to 0.002 at 350_{th} epoch. We employ the SGD optimizer with a momentum of 0.9 and the weight decay is 0.0001. The size of the memory bank M is set to 16384 and τ is 0.07.

4.3 Comparison with State-of-the-Art Methods

To give a comprehensive evaluation of the generalization capacity of the proposed method, PCM^3, we conduct experiments on the following *five* downstream tasks under three widely used datasets.

1) Skeleton-based Action Recognition. After pretraining the encoder $f(\cdot)$ on the self-supervised tasks, we utilize the learned representations to solve the skeleton-based action recognition. Specifically, two evaluation approaches are adopted, *i.e.*, unsupervised learning approach and semi-supervised learning approach.

• **Unsupervised Learning Approach** applies a fully-connected layer after the encoder, which is fixed during training. We report the top-1 accuracy results in Table 1. On NTU datasets, PCM^3 can surpass other state-of-the-art methods notably on all protocols. Especially on NTU 120 dataset, our method shows 5+% improvements compared to the latest methods. Remarkably, the single stream of our method can perform on par with the three streams of SOTA methods. Meanwhile, we give the results on PKUMMD II, which is a relatively small dataset but contains more noisy data in real life. PCM^3 can achieve the best results on all benchmarks, indicating the strong generalization capacity and robustness across datasets.

• **Semi-supervised Learning Approach** jointly trains the encoder $f(\cdot)$ and a fully-connected layer for the recognition task. But only a portion of the labeled training data is available, *i.e.*, 1% and 10%. This reflects the representation quality because a good representation can effectively avoid the over-fitting problem in training. The results are shown in Table 2. As we can see, PCM^3 renews the state-of-the-art scores with varying proportions of available training data, indicating the strong generalization capacity.

Table 3: Action retrieval results with joint stream.

Method	NTU 60		NTU 120	
	xsub	xview	xsub	xset
LongT GAN [53]	39.1	48.1	31.5	35.5
ISC [41]	62.5	82.6	50.6	52.3
CRRL [43]	60.7	75.2	-	-
CMD [26]	70.6	85.4	58.3	60.9
HaLP [32]	65.8	83.6	55.8	59.0
PCM^3	**73.7**	**88.8**	**63.1**	**66.8**

Table 4: Action recognition with occlusion results. $\Delta_{\downarrow}$ represents the average performance reduction compared to that without occlusion.

Method	Occluded NTU 60					
	Spatial Occ.(%)			Temporal Occ. (%)		
	xsub	xview	$\Delta_{\downarrow}$	xsub	xview	$\Delta_{\downarrow}$
MoCo-GRU [14]	64.8	72.6	12.4	68.8	74.8	9.3
ISC [41]	62.8	70.6	14.1	68.9	76.8	7.9
CRRL [43]	56.8	61.4	11.6	61.0	66.2	7.1
AimCLR [12]	54.9	58.5	20.3	54.1	58.6	20.7
CMD [26]	67.1	72.7	13.3	72.7	79.5	7.1
PCM^3	**80.8**	**87.0**	**3.3**	**77.6**	**86.1**	**5.4**

2) Skeleton-based Action Retrieval. We follow the settings introduced by previous work [40]. Specifically, the K-nearest neighbors (KNN) classifier (k=1) is employed to the learned representations to assign action labels for the training set. The results on NTU 60 and NTU 120 datasets are shown in Table 3. The proposed method achieves the best results, surpassing other state-of-the-art methods by a large margin. This indicates a highly distinguishable representation space is obtained through our method.

3) Action Recognition with Occlusion. Occlusions are universal disruptions that constantly occurred in the real world, which can seriously affect the performance of action recognition. We transfer the learned representations from clean dataset to the action recognition task with body occlusion. A linear evaluation protocol is adopted. Following the work [39], we use a synthetic dataset with both spatial and temporal occlusion. For spatial occlusion, we randomly masked the body parts, *e.g. trunk* and *right-hand*. For temporal occlusion, we randomly mask a block of frames to zeros. All masks are generated randomly with the masking ratio sampled from [0.3, 0.7].

The experiments are conducted under the same settings across compared methods, and the results are shown in Table 4. Due to the proposed topology-based masked contrastive learning pretraining, our method can capture the discriminative structures in the distorted data and well handle the spatial occlusion. Meanwhile, this ability also extends well to temporal occlusion. As we can see, our approach showed significant improvements in both occlusion scenarios, as well as minimal performance degradation compared to that under clean data.

Table 5: Action detection results on PKUMMD Part I xsub benchmark with overlap ratio of 0.5.

Method	mAP_a (%)	mAP_v (%)
Randomly Initializaed	29.6	28.2
MS^{2}L [21]	50.9	49.1
CRRL [43]	52.8	50.5
ISC [41]	55.1	54.2
CMD [26]	59.4	59.2
PCM^3	**61.8**	**61.3**

Table 6: The results on motion prediction task.

Method	Random	CRRL [43]	ISC [41]	CMD [26]	PCM^3
MPJPE (mm)	108.5	104.9	144.2	145.0	**101.7**

4) Skeleton-based Action Detection. Following the settings in [5, 23], we evaluate the detection performance under the PKUMMD I dataset, to demonstrate effectiveness for the short-term frame-level discrimination task. We attach a linear classifier (fully-connected layer) to the encoder and finetune the whole model to predict the label of each frame. The encoder is pretrained on NTU 60 xsub dataset and then transfers to the PKUMMD Part I xsub dataset. We adopt the mean average precision of different actions (mAP_a) and different videos (mAP_v) with the overlapping ratio of 0.5 as the evaluation metrics. The results are shown in Table 5. First, we can see the existing contrastive learning methods can largely boost the detection performance compared with the randomly initialized model. It indicates the learned sequence-level representations can benefit the frame-level task to some extent. Besides, our method further improves the state-of-the-art scores owing to the synergetic modeling of the joint-level and sequence-level features.

5) Motion Prediction. Following the previous work [5], we give the results of the motion prediction task, which is a dense prediction downstream task. We use the decoder in [27] after $f(\cdot)$ and follow the short-term motion prediction protocol [27]. As shown in Table 6, our method achieves the best results in terms of the MPJPE metric. Although previous contrastive learning-based methods show good performance in a high-level recognition task, most of them show adverse effects on motion prediction compared with the randomly initialized method. It is because they only focus on the high-level information and ignore the joint-level feature modeling. In contrast, our method extracts the joint-level and sequence-level semantic features and effectively boosts the dense-prediction downstream task performance.

4.4 Ablation Study

In this part, we give a more detailed analysis of the proposed method. The results are reported on the action recognition task under linear evaluation, using NTU 60 dataset xview protocol by default.

1) Effect of masking strategy. We utilize the *topology-based* masking strategy in our framework to construct more challenging corrupted data views. In Table 10 we report the linear evaluation results

Table 7: Ablation study on the synergy of the contrastive learning and masked skeleton prediction on different downstream tasks. *Multi-Task* stands for simply combining the two tasks in a multi-tasking manner.

Method	Recognition Accuracy (%)	Retrieval Accuracy (%)	Recognition *w* Occlusion Accuracy (%)	Detection mAP_a (%)	Motion Prediction MPJPE
Masked Prediction	14.7	64.4	11.7	47.0	102.8
Contrastive Learning	87.3	84.5	76.6	58.4	147.3
Multi-Task	87.5	85.1	77.4	59.7	103.9
Ours	**90.4**	**88.8**	**87.0**	**61.8**	**101.7**

Table 8: Ablation study on the different positive samples and the distillation loss (without any prompt applied).

s_{intra}	s_{inter}	s_{mask}	$s_{predict}$	$\mathcal{L}_{KL}$	Acc. (%)
✓					84.7
✓	✓				87.5
✓	✓			✓	88.2
✓	✓		✓	✓	89.7
✓		✓		✓	89.3
✓	✓	✓	✓	✓	**90.0**

Table 9: Ablation study on the different types of prompts.

Domain-	Task-	Recognition (%)	Detection (%)
		90.0	60.8
✓		90.3	61.0
✓	✓	**90.4**	**61.3**

Table 10: Ablation study on the masking strategy.

Mask Ratio	Topology-based	Random
20%	89.8	89.0
40%	90.2	89.5
60%	**90.5**	**89.7**
80%	90.1	89.5

for action recognition task because it is widely used for evaluation of other representation learning methods. Compared with *Random Mask* strategy, the model shows better results, indicating a more distinguishable feature space. Meanwhile, more realistic occluded data are simulated by this masking strategy, which are often continuous in spatial-temporal dimension. We set mask ratio to 0.6.

2) Analysis of the synergetic design between the two tasks. We first analyze the performance under different downstream tasks when adopting single paradigm, *i.e.*, masked prediction and contrastive learning, or combinations of them, in Table 7. Only employing masked prediction task cannot generate highly distinguishable feature space, resulting in poor performance in linear recognition task. Meanwhile, the contrastive learning mainly modeling high-level features and the learned prior representations are not beneficial for motion prediction. Therefore, to combine the merits of both paradigms, we can directly perform the multi-task learning. Owing to the well-designed contrastive learning and masked prediction pipelines, naive multi-task method can achieve a decent performance. However, it ignores the connection between the two tasks and only shows mediocre improvements on the recognition task. In contrast, our design utilizes the synergy between the two tasks, and further improve the performance and generalization capacity, proving the effectiveness of the proposed method.

Next, we elaborate on the separate effect of our synergetic designs. As shown in Table 8, the proposed inter- and intra- skeleton transformations can well boost the contrastive learning performance. Meanwhile, the masked views and predicted views in the masked prediction training as the positive pairs yield 1.1% and 1.5% improvement, respectively. The relational distillation loss is found effective and further improve representation quality.

Finally, we show the effect of propagating the gradients of contrastive branch to masked prediction here, *80.5* with gradient vs. *80.0* without gradient, that is stopping and separating the gradients of two branches. The gradient information from contrastive branch can serve as high-level semantic guidance, which boosts the masked prediction training and yields more informative positive samples.

3) Effect of the dual-prompted pretraining strategy. We show the effect of different types of prompts in Table 9. As we can see, the proposed domain-specific and task-specific prompts are effective for the representation learning, which can learn the domain- and task-specific knowledge and serve as explicit guidance in-training. This assists the model to discriminate the domain and task identity, reducing the interference and ambiguity when the model learns from multi-view data and multiple tasks. When the two types of prompt are utilized, the model achieves the best results.

5 CONCLUSION

We propose a novel framework called prompted contrast with masked motion modeling, PCM^3, which can effectively learn meaningful representations by exploring the mutual collaboration between contrastive learning and masked prediction tasks. Specifically, the novel views in masked prediction training are utilized as the positive samples for contrastive learning. Meanwhile, contrastive learning provides semantic guidance for masked prediction in turn by propagating the gradients to the prediction decoder. Furthermore, we introduce dual-prompt multi-task pretraining strategy to provide explicit guidance. Extensive experiments are conducted, demonstrating the superior performance and promising generalization capacity of our method.

REFERENCES

[1] Yalong Bai, Yifan Yang, Wei Zhang, and Tao Mei. 2022. Directional Self-Supervised Learning for Heavy Image Augmentations. In *IEEE CVPR*.

[2] Fanta Camara, Nicola Bellotto, Serhan Cosar, Dimitris Nathanael, Matthias Althoff, Jingyuan Wu, Johannes Ruenz, André Dietrich, and Charles W Fox. 2020. Pedestrian models for autonomous driving Part I: low-level models, from sensing to tracking. *IEEE Transactions on Intelligent Transportation Systems* 22, 10 (2020), 6131–6151.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*.

[4] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved baselines with momentum contrastive learning. arXiv:2003.04297

[5] Yuxiao Chen, Long Zhao, Jianbo Yuan, Yu Tian, Zhaoyang Xia, Shijie Geng, Ligong Han, and Dimitris N Metaxas. 2022. Hierarchically Self-supervised Transformer for Human Skeleton Representation Learning. In *ECCV*.

[6] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. 2020. Skeleton-based action recognition with shift graph convolutional network. In *IEEE CVPR*.

[7] Yi-Bin Cheng, Xipeng Chen, Junhong Chen, Pengxu Wei, Dongyu Zhang, and Liang Lin. 2021. Hierarchical transformer: Unsupervised representation learning for skeleton-based human action recognition. In *IEEE ICME*.

[8] Jianfeng Dong, Shengkai Sun, Zhonglin Liu, Shujie Chen, Baolong Liu, and Xun Wang. 2023. Hierarchical Contrast for Unsupervised Skeleton-based Action Representation Learning. In *AAAI*.

[9] Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *IEEE CVPR*.

[10] Luca Franco, Paolo Mandica, Bharti Munjal, and Fabio Galasso. 2023. Hyperbolic Self-paced Learning for Self-supervised Skeleton-based Action Representations. In *ICLR*.

[11] Yulu Gan, Xianzheng Ma, Yihang Lou, Yan Bai, Renrui Zhang, Nian Shi, and Lin Luo. 2023. Decorate the Newcomers: Visual Domain Prompt for Continual Test Time Adaptation. (2023).

[12] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. 2022. Contrastive Learning from Extremely Augmented Skeleton Sequences for Self-supervised Action Recognition. In *AAAI*.

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *IEEE CVPR*.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *IEEE CVPR*.

[15] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. 2017. A new representation of skeleton sequences for 3d action recognition. In *IEEE CVPR*.

[16] Boeun Kim, Hyung Jin Chang, Jungho Kim, and Jin Young Choi. 2022. Global-local Motion Transformer for Unsupervised Skeleton-based Action Learning. *ECCV* (2022).

[17] Sungnyun Kim, Gihun Lee, Sangmin Bae, and Se-Young Yun. 2020. Mixco: Mix-up contrastive learning for visual representation. *arXiv:2010.06300* (2020).

[18] Junwoo Lee and Bummo Ahn. 2020. Real-time human action recognition with a low-cost RGB camera and mobile robot platform. *Sensors* 20, 10 (2020), 2886.

[19] Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. 2020. i-Mix: A domain-agnostic strategy for contrastive representation learning. *arXiv:2010.08887* (2020).

[20] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 2021. 3d human action representation learning via cross-view consistency pursuit. In *IEEE CVPR*.

[21] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. 2020. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *ACM MM*.

[22] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. 2019. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE TPAMI* 42, 10 (2019), 2684–2701.

[23] Jiaying Liu, Sijie Song, Chunhui Liu, Yanghao Li, and Yueyu Hu. 2020. A Benchmark Dataset and Comparison Study for Multi-modal Human Action Analytics. *ACM TOMM* 16, 2 (2020), 41:1–41:24.

[24] Mengyuan Liu, Hong Liu, and Chen Chen. 2017. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition* 68 (2017), 346–362.

[25] Irvin Hussein Lopez-Nava and Angélica Muñoz-Meléndez. 2019. Human action recognition based on low-and high-level data from wearable inertial sensors. *International Journal of Distributed Sensor Networks* 15, 12 (2019), 1550147719894532.

[26] Yunyao Mao, Wengang Zhou, Zhenbo Lu, Jiajun Deng, and Houqiang Li. 2022. CMD: Self-supervised 3D Action Representation Learning with Cross-Modal Mutual Distillation. In *ECCV*.

[27] Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2891–2900.

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. arXiv:1807.03748

[29] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. 2021. Skeleton-based action recognition via spatial and temporal transformer networks. *CVIU* 208 (2021), 103219.

[30] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. 2023. Contrast with Reconstruct: Contrastive 3D Representation Learning Guided by Generative Pretraining. *arXiv:2302.02318* (2023).

[31] Sucheng Ren, Huiyu Wang, Zhengqi Gao, Shengfeng He, Alan Yuille, Yuyin Zhou, and Cihang Xie. 2022. A Simple Data Mixing Prior for Improving Self-Supervised Learning. In *IEEE CVPR*.

[32] Anshul Shah, Aniket Roy, Ketul Shah, Shlok Kumar Mishra, David Jacobs, Anoop Cherian, and Rama Chellappa. 2023. HaLP: Hallucinating Latent Positives for Skeleton-based Self-Supervised Learning of Actions. *arXiv:2304.00387* (2023).

[33] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *IEEE CVPR*.

[34] Zhiqiang Shen, Zechun Liu, Zhuang Liu, Marios Savvides, Trevor Darrell, and Eric Xing. 2022. Un-mix: Rethinking image mixtures for unsupervised visual representation learning. In *AAAI*. 2216–2224.

[35] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *IEEE CVPR*.

[36] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2020. Decoupled spatial-temporal attention network for skeleton-based action recognition. arXiv:1709.04875

[37] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*.

[38] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2018. Skeleton-indexed deep multi-modal feature learning for high performance human action recognition. In *IEEE ICME*.

[39] Yi-Fan Song, Zhang Zhang, and Liang Wang. 2019. Richly activated graph convolutional network for action recognition with incomplete skeletons. In *IEEE ICIP*.

[40] Kun Su, Xiulong Liu, and Eli Shlizerman. 2020. Predict & cluster: Unsupervised skeleton based action recognition. In *IEEE CVPR*.

[41] Fida Mohammad Thoker, Hazel Doughty, and Cees GM Snoek. 2021. Skeleton-contrastive 3D action representation learning. In *ACM MM*.

[42] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research* 11, 12 (2010).

[43] Peng Wang, Jun Wen, Chenyang Si, Yuntao Qian, and Liang Wang. 2022. Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition. *IEEE TIP* 31 (2022), 6224–6238.

[44] Chen Wei, Huiyu Wang, Wei Shen, and Alan Yuille. 2020. Co2: Consistent contrast for unsupervised visual representation learning. arXiv:2010.02217

[45] Wenhan Wu, Yilei Hua, Shiqian Wu, Chen Chen, Aidong Lu, et al. 2022. SkeletonMAE: Spatial-Temporal Masked Autoencoders for Self-supervised Skeleton Action Recognition. *arXiv:2209.02399* (2022).

[46] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. Simmim: A simple framework for masked image modeling. In *IEEE CVPR*.

[47] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*. 7444–7452.

[48] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*.

[49] Chen Zhan, Liu Hong, Guo Tianyu, Chen Zhengyan, Song Pinhao, and Tang Hao. 2022. Contrastive Learning from Spatio-Temporal Mixed Skeleton Sequences for Self-Supervised Skeleton-Based Action Recognition. In *arXiv: 2207.03065*.

[50] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. arXiv:1710.09412

[51] Haoyuan Zhang, Yonghong Hou, Wenjing Zhang, and Wanqing Li. 2022. Contrastive positive mining for unsupervised 3d action representation learning. In *ECCV*. Springer.

[52] Jiahang Zhang, Lilang Lin, and Jiaying Liu. 2023. Hierarchical Consistent Contrastive Learning for Skeleton-Based Action Recognition with Growing Augmentations. In *AAAI*.

[53] Nenggan Zheng, Jun Wen, Risheng Liu, Liangqu Long, Jianhua Dai, and Zhefeng Gong. 2018. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *AAAI*.