



Self-Supervised Skeleton-Based Action Representation Learning: A Benchmark and Beyond

Jiahang Zhang¹ · Lilang Lin¹ · Shuai Yang¹ · Jiaying Liu¹

Received: 3 August 2024 / Accepted: 14 October 2025 / Published online: 7 January 2026
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

Abstract

Self-supervised learning (SSL), which aims to learn meaningful prior representations from unlabeled data, has been proven effective for skeleton-based action understanding. Different from the image domain, skeleton data possesses sparser spatial structures and diverse representation forms, with the absence of background clues and the additional temporal dimension, presenting new challenges for spatial-temporal motion pretext task design. Recently, many endeavors have been made for skeleton-based SSL, achieving remarkable progress. However, a systematic and thorough review is still lacking. In this paper, we conduct, for the first time, a comprehensive survey on *self-supervised skeleton-based action representation learning*. Following the taxonomy of context-based, generative learning, and contrastive learning approaches, we make a thorough review and benchmark of existing works and shed light on the future possible directions. Remarkably, our investigation demonstrates that most SSL works rely on the single paradigm, learning representations of a single level, and are evaluated on the action recognition task solely, which leaves the generalization power of skeleton SSL models under-explored. To this end, a novel and effective SSL method for skeleton is further proposed, which integrates versatile representation learning objectives of different granularity, substantially boosting the generalization capacity for multiple skeleton downstream tasks. Extensive experiments under three large-scale datasets demonstrate our method achieves superior generalization performance on various downstream tasks, including *recognition, retrieval, detection, and few-shot learning*.

Keywords Self-supervised learning · Skeleton-based action understanding · Contrastive learning · Masked skeleton modeling

1 Introduction

Human activity understanding is an essential topic in the research of computer vision due to its wide applications in real life, such as human-robotics interaction (Lee & Ahn, 2020), autonomous driving (Camara et al., 2020), and healthcare (Hussein et al., 2019). Among the different data modalities for actions, skeletons represent the human body

by 3D coordinates of key body joints, which are lightweight, compact, and more robust to changes of view and background. Owing to these desirable advantages, skeleton has attracted much attention in human action analysis.

In the early works, many endeavors have been put into the supervised skeleton-based human activity understanding, *e.g.*, recognition and detection (Yan et al., 2018; Shi et al., 2019; Cheng et al., 2020; Rao et al., 2024). However, these supervised methods heavily rely on huge amounts of labeled data, which requires time-consuming and expensive data annotation work, limiting the wide applications in the real world. As a remedy to this problem, self-supervised learning (SSL) attracts much attention and has been proven successful for representation learning. It exploits supervisory signals from unlabeled data, learning meaningful prior features and boosting generalization capacity of model for downstream tasks. Motivated by recent success in the image domain, great interest has arisen in adopting SSL for skeleton. However, it is not trivial to transfer these approaches into the skeleton data directly, which are with a more com-

Communicated by Shaogang Gong.

✉ Jiaying Liu
liujiaying@pku.edu.cn
Jiahang Zhang
zjh2020@pku.edu.cn
Lilang Lin
linlilang@pku.edu.cn
Shuai Yang
williamyang@pku.edu.cn

¹ Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China

pact spatial structure, additional temporal dimension, and the absence of the background clues. To this end, researchers have made valuable exploration for skeleton-based SSL.

Generally, existing skeleton-based SSL works can be categorized into three types according to the pre-training pretext tasks, *i.e.*, *context-based*, *generative learning*, and *contrastive learning* methods. The context-based methods construct the pseudo-label based on the intrinsic property of data, *e.g.*, the joint angle prediction, to learn the spatial and temporal relations. Generative learning mainly focuses on reconstructing and predicting the skeleton data or the corresponding features. Contrastive learning methods model the high-level representations with an instance discrimination task. The various positive and negative skeleton views are generated by the well-designed spatial-temporal augmentations, boosting the consistency learning of the model.

Despite the huge progress recently, there is still a lack of a thorough literature review and analysis. Therefore, in this paper, we contribute a comprehensive survey of the self-supervised skeleton-based action representation learning. In contrast to other SSL surveys towards image, video, or text data, we focus on skeleton-based representation learning, which is the first literature to the best of our knowledge. This survey conducts a thorough review of mainstream SSL literature for skeleton and also involves the skeleton data collection, benchmark of performance, and the discussion of future possible directions. We believe our extensive work can benefit the research community and bring rich insights for future work.

Based on our literature review, it is noticed that most SSL methods for skeleton focus on the single paradigm, learning representations of single granularity, *e.g.*, joint-level features (by masked skeleton modeling, MSM) (Zheng et al., 2018; Wu et al., 2022) or sequence-level features (by contrastive learning) (Li et al., 2021; Mao et al., 2022; Dong et al., 2023; Zhang et al., 2023). This limits the generalization capacity of the model to more downstream tasks. Although some works (Lin et al., 2020; Su et al., 2020; Wang et al., 2022) make valuable efforts to combine different paradigms, they only achieve mediocre improvement due to the inherent gap of feature modeling mechanisms between the contrastive learning and masked modeling (Qi et al., 2023; Gui et al., 2023). To this end, a novel SSL approach for skeleton is proposed to fully boost the generalization capacity of SSL model, which integrates the contrastive learning and MSM to learn the joint, clip, and sequence level representations jointly. Specifically, we fully utilize the novel motion pattern exposed by the manual designed augmentations and model training for sequence-level contrastive representation learning, while adopting MSM for the joint-level feature modeling. Besides, we further propose a novel clip-level contrastive learning method, which significantly boosts the short-term modeling capacity, along with an effective post-

distillation strategy to achieve a more compact representation space. Finally, extensive experiments under *five* downstream tasks, not limited to the single action recognition task used in most previous works, demonstrate the promising generalization capacity of the proposed method.

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to provide a thorough survey that comprehensively reviews the self-supervised skeleton action representation learning literature. Based on the taxonomy of context-based, generative learning, and contrastive learning, we give a detailed analysis of the pretext task design and highlight the special consideration for skeleton data along with the corresponding challenges.
- Motivated by the limitations revealed by our survey, we explore skeleton-based versatile action representation learning to fully mine the generalization power of SSL models. An effective SSL schema is proposed, which integrates contrastive learning and MSM to jointly model the representations of different granularity, remarkably benefiting different downstream tasks.
- We present the first multi-task benchmark of existing skeleton SSL works with the insightful analysis from the perspective of model backbones, pre-training paradigms, and future possible directions. Finally, we demonstrate promising performance of our proposed method on *five* downstream tasks for versatile action representation learning.

The remaining sections are organized as follows: We first present a thorough review in Sec. 2, for skeleton-based SSL representation learning. Subsequently, based on our investigation, we propose a new method exploring the combination of contrastive learning and masked modeling tasks in Sec. 3. Then, we comprehensively benchmark existing methods in Sec. 4 on multiple downstream tasks, and verify the effectiveness of our proposed approach. Finally, we conclude and summarize with possible future directions in Sec. 5.

2 Review on Skeleton-Based Action Representation SSL

Generally, a two-stage paradigm is utilized in skeleton-based SSL, *i.e.*, pre-training on pretext tasks first and then fine-tuning on downstream tasks. In the pre-training stage, different pretext tasks are designed for deep neural networks, capturing training signals derived from the data itself, called the process of *self-supervision* (Shurrah & Duwairi, 2022). After that, the learned knowledge as feature representations is transferred to downstream tasks as shown in Fig. 1 (c). In principle, this part is not only the goal of SSL, *i.e.*, to improve

downstream task performance with learned representations, but also the way to assess the quality of representation learning methods. Note that some downstream tasks, *e.g.*, motion prediction and 2D-to-3D lifting, are not considered in this survey, because they essentially do not rely on the supervisory signal of human annotation and can serve as pre-training pretext tasks themselves, leading to possible unfair comparison. For reviews on these topics, we direct readers to Lyu et al. (2022); Liu et al. (2022).

Next, we first introduce skeleton data as well as its collection in Sec. 2.1. Then, a review of the skeleton SSL methods is presented in Sec. 2.2 based on the taxonomy of pretext tasks methodologies as shown in Fig. 1 (b). A summary and discussion are finally provided in Sec. 2.3.

2.1 Human Skeleton Representation

The collection methods of skeleton can be divided into two main categories, *i.e.*, marker-based and markerless methods, shown in Fig. 1 (a).

Marker-based motion capture (Mocap) systems often rely on inertial measurement units (IMU) or reflective markers, placed on the body to track the movement of humans. It can provide reliable skeleton data with even sub-millimeter accuracy (Buckley et al., 2019). However, this method is too costly in many application scenarios and requires highly trained personnel to operate. Meanwhile, it suffers from great inconvenience, *e.g.*, the time-consuming placement process and a requirement for a controlled environment. The datasets (Ionescu et al., 2013; Von Marcard et al., 2018) for generation tasks, *e.g.*, pose estimation, are usually collected in this way to guarantee the accuracy.

Markerless Mocap often depends on deep learning algorithms for pose estimation from RGB and depth data. Video RGB data can be easily obtained from the Internet while we can also utilize the camera hardware, *e.g.*, the depth camera including Microsoft Azure Kinect, and single or multiple RGB video cameras, to collect the RGB, infrared, and depth images in deployment. Then to get the 3D motion, computer vision algorithms, on multi-view geometry and pose estimation, *i.e.*, *OpenPose* (Cao et al., 2017) are then employed to detect and extract joint center locations. The whole pipeline is presented in Fig. 1 (a). However, due to the limitations of hardware and estimation algorithms, such method can have large errors compared to marker-based methods (Dolatabadi et al., 2016). Nevertheless, it is still chosen for most skeleton-based datasets (Shahroudy et al., 2016; Liu et al., 2019, 2020) for its simpleness and convenience.

2.2 SSL Methods for Skeleton

SSL has been widely explored in the previous literature for the image domain. Some early works adopt the context-

driven pretext tasks to model the inherent data structure, *e.g.*, predicting the rotation label (Gidaris et al., 2018), correlation map (Li et al., 2023), and clusters based on the pseudo-labels (Caron et al., 2018; Zhan et al., 2020). Besides, contrastive-style methods push the performance of SSL to new heights, including invariance decorrelation (Zbontar et al., 2021; Bardes et al., 2021), distillation-based (Grill et al., 2020; Chen & He, 2021; Caron et al., 2021; Oquab et al., 2023; Darcet et al., 2023), clustering-based (Caron et al., 2020; Li et al., 2020). Meanwhile, the generative learning, especially the recent masked image modeling (Xie et al., 2022b; He et al., 2022), has drawn a surge of interest and been proven effective for downstream vision downstream tasks. These works have played an important role and motivate the new efforts for skeleton-based SSL. However, due to the huge difference between human skeleton and image data as discussed in Sec. 1, it is suboptimal to transfer these methods directly for skeleton representation learning. Therefore, we present this literature review with a high-level methodological taxonomy, and also shed light on the challenges and skeleton-specific designs to motivate further work.

With respect to the pretext tasks, most existing skeleton-based representation learning methods encompass three categories: (1) *context-based*, (2) *generative learning*, and (3) *contrastive-learning* methods. Based on the taxonomy in Fig. 2, we provide a comprehensive survey on the skeleton-based SSL works and highlight the special design for the skeleton as well as the corresponding challenges, to distinguish them from other data modalities.

2.2.1 Context-Based Methods

Context-based methods generate the supervisory training signals according to the inherent contextual information of provided data. The model is encouraged to learn the spatial-temporal relationships by training on the pre-defined task. Emerging from the image domain, the pretext tasks rely on the context understanding, *e.g.*, rotation prediction (Gidaris et al., 2018). In contrast, skeleton data introduces an additional temporal dimension, and possesses more compact spatial information, which presents a new challenge on *how to mine the meaningful spatial-temporal context of skeleton by pretext tasks*. Typically, there are three common types of context-based pretext tasks, view-invariance-based, temporal-order-based, and motion-prior-based method as shown in Fig. 3.

1) View Invariance. Due to the variance of the observation viewpoint, the skeleton estimation can suffer from occlusion and noise. Therefore, learning view-invariant representations should be beneficial for action recognition task, and has been widely studied in the supervised skeleton-based action recognition task (Gao et al., 2022). Li et al. (2018) proposed a view

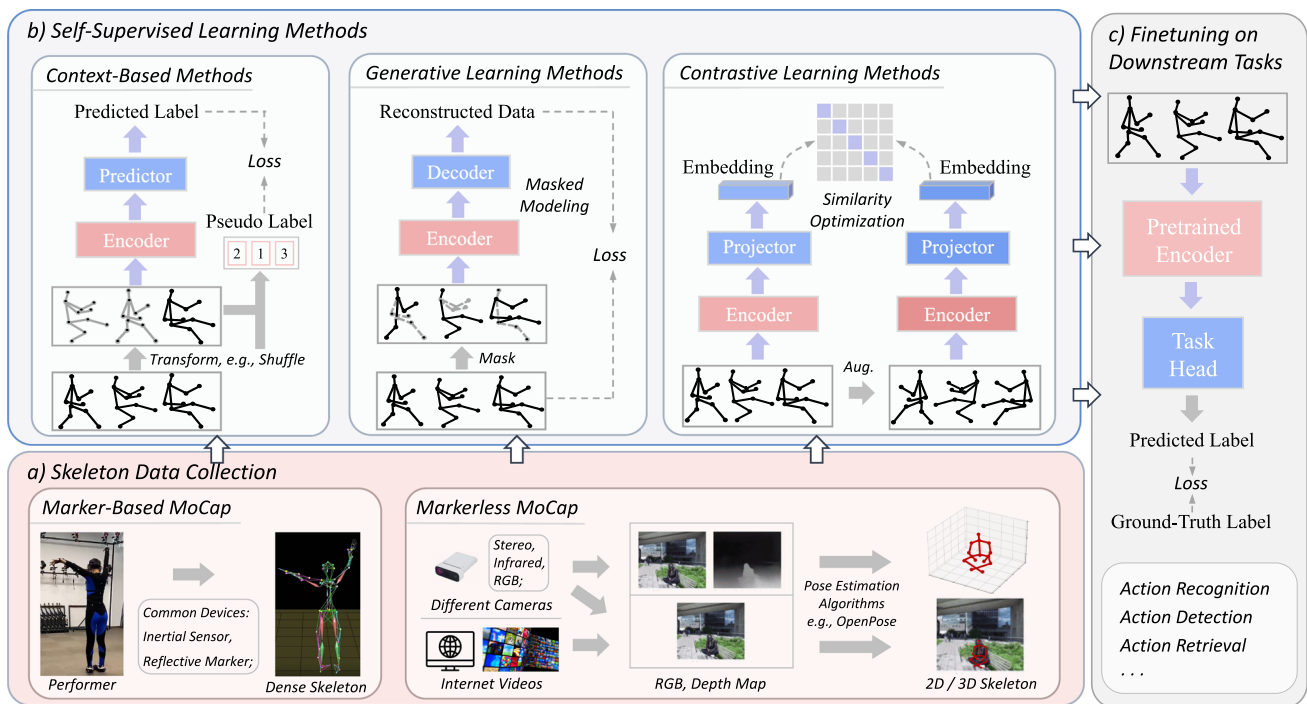
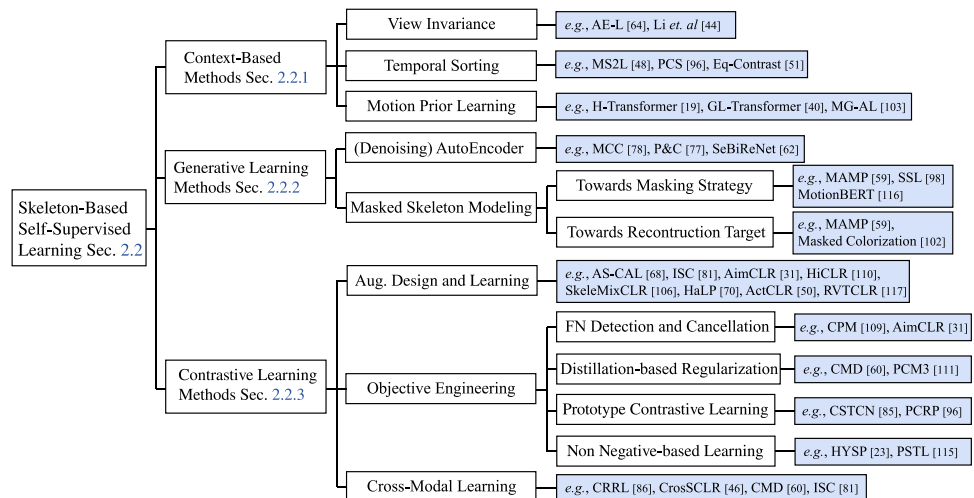


Fig. 1 The taxonomy framework for self-supervised skeleton-based representation learning in our survey. The survey is structured around three dimensions: skeleton data collection, SSL pretext design, and SSL downstream task evaluation, providing a comprehensive review.

Fig. 2 The taxonomy of the skeleton-based self-supervised learning methods in our review.



classification pretext task for unsupervised action representation learning. Specifically, as shown in Fig. 3 (a), skeleton sequences of different views are fed into the encoder, subsequent to which a view classifier predicts the view labels. To learn the view-invariant features, a Gradient Reversal Layer (Ganin et al., 2015) is added to reverse the optimization direction of the encoder. Resort to this, the encoder can learn the features insensitive to view in an adversarial manner. Likewise, Paoletti et al. (2022) adopted the rotation prediction, *i.e.*, pitch angle, yaw angle, and roll angle, as well as a gradient reverse operation to achieve the viewpoint-invariance learning.

There are also some subsequent SSL works (Yang et al., 2024; Men et al., 2023; Bian et al., 2022) focusing on learning view-invariant representations. However, they adopted a contrastive learning paradigm to explicitly learn the alignment between the two views and showed a significant performance improvement.

2) Temporal Sorting. In this pretext task, skeleton data is treated as temporal sequences, and shuffled randomly. The model takes the shuffled skeleton as input and outputs the corresponding shuffling pseudo label to restore the temporal order. Specifically, the skeleton sequence is usually divided

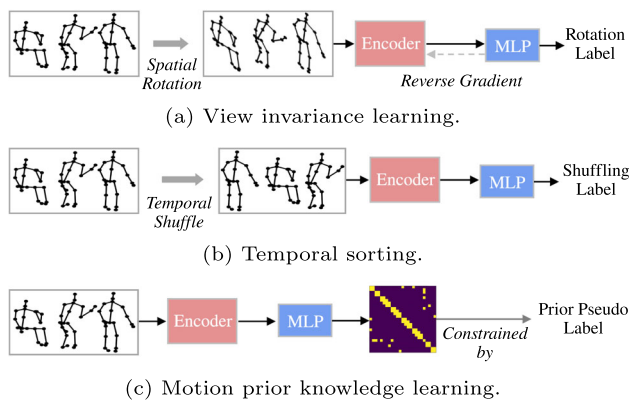


Fig. 3 Three types of context-based SSL methods for skeleton.

into multiple clips first, each of which contains several consecutive frames. Then the shuffling operation takes place at the clip level rather than the frame level because it is difficult to capture motion patterns by the difference between two adjacent frames. The model is trained to predict the shuffling label as shown in Fig. 3 (b), *e.g.*, with a *Cross-Entropy* loss.

Some skeleton-based SSL works (Lin et al., 2020; Xu et al., 2021a) are equipped with temporal sorting to model the temporal dependencies, similar to SSL works in video (Xu et al., 2019; Lee et al., 2017). However, such pretext task is often integrated with other pretext tasks, because it only models temporal features explicitly, leaving the crucial spatial representations of skeleton under-explored. Therefore, new pretext tasks are studied to jointly learn the spatial-temporal relationships of skeleton.

3) Motion Prior. The motion dynamics of skeleton joints contain rich semantic information beneficial for action understanding. Therefore, researchers propose to leverage the inherent motion prior knowledge to generate pseudo labels for pre-training as shown in Fig. 3 (c). Cheng et al. (2021) designed a movement direction prediction task. The model is constrained to estimate the direction of the instantaneous joint velocity, *i.e.*, whether the targeted joints are moving in a positive direction. Based on this, Kim (2022) proposed a multi-interval pose displacement strategy to encourage both local and global attention learning. The key idea is to predict the motion direction and magnitude of the central joint and other joints during a random interval. On the other hand, Yang et al. (2022) introduced more motion prior information as the pseudo label, *i.e.*, intra-joint motion variance, inter-joint motion covariance, intra-frame joint angle, and inter-frame motion deviation. The model is pre-trained by a regression task to predict this prior knowledge.

In summary, compared with image and video modalities, the context-based pretext tasks for skeleton usually involve more spatial-temporal design and motion dynamics modeling. However, these tasks are often designed manually and

hard to capture the high-level underlying semantic distribution. Therefore, more efforts have been paid to the generative and contrastive learning SSL methods, which will be discussed in the following.

2.2.2 Generative Learning Methods

Generative methods utilize the generative power of the neural network to capture the spatial-temporal co-occurrence relationships among skeleton joints, modeling the underlying data distributions. In this case, the meaningful representations are learned by reconstructing or predicting the related signals of input skeletons. Here we do not strictly distinguish between “reconstruction” and “prediction” for generation, but use “reconstruct” to refer uniformly. Generally, the goal of the generative learning tasks can be formulated as:

$$\operatorname{argmin}_{\theta} \mathcal{L}(\mathcal{D}(\mathcal{E}(\mathcal{T}_{in}(x))), \mathcal{T}_{tar}(x); \theta), \quad (1)$$

where x is a skeleton sequence, \mathcal{D} and \mathcal{E} are the decoder and encoder, respectively. \mathcal{T}_{in} is the transformation applied to the original data, *e.g.*, random masking, while \mathcal{T}_{tar} maps the input into the target space where the loss objective \mathcal{L} is applied to optimize the model parameters θ .

When \mathcal{T}_{in} and \mathcal{T}_{tar} are both the identity function, the model is constrained to reconstruct the input data simply, known as AutoEncoder (AE). It naturally creates an information bottleneck to achieve dimension reduction, mapping data from input space onto low-dimensional feature space. Due to its simpleness, AE is used to learn prior representations in earlier skeleton-based works (Su et al., 2020; Xu et al., 2021b). However, since the skeleton data is still redundant in spatial and temporal dimensions, the encoder tends to memorize the input data at a low level instead of modeling the high-level semantic knowledge. Then, denoising AE (DAE) is studied for alleviating this shortcut (Su et al., 2021; Nie et al., 2020). For example, Nie et al. (2020) applied a series of view corruptions and constrained the model to reconstruct the clean data to disentangle the view and pose features. On the other hand, inspired by the remarkable success of masked language/image modeling (MLM, MIM), which can be regarded as a variant of DAE, researchers have aroused a surge of interest in exploring masked skeleton modeling (MSM) for human action representation learning. The earlier works are mainly based on recurrent neural network (RNN) or graph convolutional network (GCN) to predict the masked skeletons. However, lacking good scalability, these models only show mediocre performance improvement. Recently, inspired by the success of Vision Transformer (ViT) as masked autoencoders (MAE) He et al. (2022), Transformer-based models have been explored for the masked skeleton modeling (Chen et al., 2022; Mao et al., 2023). However, due to the relative redundancy in skeleton and the lack of a large-scale dataset as

ImageNet, directly applying Transformer for masked skeleton reconstruction can suffer from the over-fitting problem. To this end, researchers have made endeavors on two crucial designs, masking strategy and reconstruction target, which are discussed as follows.

1) Masking Strategy has been proved crucial as the design of \mathcal{T}_{in} in Eq. (1) for MIM. In the field of RGB images, MAE (He et al., 2022) adopts a patch-based masking strategy with a large masking ratio of 75%. For the skeleton data, researchers have fully explored the masking strategy in the spatial-temporal dimension. Although the optimal masking strategy and ratio can differ with the encoder backbone and MSM task setting, most works have found that spatially body-part-based and temporally segment-based strategy with large masking ratio, *e.g.*, 90% in Mao et al. (2022), can produce decent results. Concretely, body-part level masking refers to performing masking regarding the different joints in a body part as a whole, *e.g.*, *torso* and *left leg*, while the temporal segment-based strategy masks the same joints across consecutive frames. These designs aim to reduce the shortcuts caused by spatial-temporal redundancy in skeleton, which is in line with the observation in images (He et al., 2020) and videos (Tong et al., 2022).

Recently, more elaborate masking strategies have been studied to achieve more effective exploitation of valuable semantic information. The work (Yan et al., 2023) found that masking the limbs is always better than masking the torso and head, especially the right hand and leg. MAMP (Mao et al., 2023) proposes a motion-aware masking strategy. The moving parts are located and masked by calculating the motion displacement between adjacent frames, which achieves better results than the random masking strategy. These results demonstrate that masking motion regions, which are often semantic-rich, promotes the model to learn more meaningful features in masked motion modeling. It can be also explained by the theory in Zhang and Wang (2022), *i.e.*, because samples from the same action often contain similar motion patterns, this masking strategy on the motion regions implicitly leads to better alignment of mask-induced positive pairs, achieving more discriminative feature spaces.

Meanwhile, some literature adopts special mask designs, yielding new MSM task forms. MS²L Lin et al. (2020) directly utilizes temporal frame-level masks to train the model on the motion prediction task. MotionBERT (Zhu et al., 2023a) adopts a 2D-to-3D lifting pretext task, in which it masks all the values in depth z channel and partially in x and y channels, encouraging the model to predict the original 3D skeletons. These works can be divided into masked skeleton modeling in a broad sense.

2) Reconstruction Target, known as the \mathcal{T}_{tar} , also varies from the input space (He et al., 2022; Xie et al., 2022b) to the feature space (Assran et al., 2023; Gao et al., 2024; Xie et

al., 2022; Haghghat et al., 2023; Bardes et al., 2024) as the image domain. Most existing works reconstruct the skeleton data in the input coordinate space using MSE loss. Some works (Wang et al., 2022) propose to perform temporally reverse reconstruction to learn the skeleton dynamics instead of trivial representations that just remember the input.

Instead of directly reconstructing the input, Mao et al. (2023) proposed to reconstruct the motion in MSM, *i.e.*, the difference of the corresponding joints between adjacent frames. The work (Yang et al., 2023) formulates the skeleton as an unordered 3D point cloud and maps the 3D data onto color space. This mapping function is artificially defined according to the spatial-temporal relationship of skeleton joints. Therefore, the model can learn the spatial relation and temporal dependency by reconstructing the color of the skeleton cloud.

2.2.3 Contrastive-Learning Methods

Contrastive learning has been proven effective for different data modalities, *e.g.*, images, point cloud, as well as skeleton data. Generally, contrastive learning pursues the consistency of the *positive samples*, which are usually the augmented counterparts of the original data. MoCo (He et al., 2020) utilizes the negative samples to establish an instance discrimination pretext task. Meanwhile, self-distillation (Grill et al., 2020) and feature decoupling (Zbontar et al., 2021; Bardes et al., 2021) concepts have also been explored in contrastive learning. These early pioneering works (He et al., 2020; Chen et al., 2020; Grill et al., 2020; Zbontar et al., 2021; Robinson et al., 2020), have made a huge impact and encouraged unique designs for skeleton contrastive learning.

For skeleton contrastive representation learning, most existing methods are based on MoCo v2 (He et al., 2020). It adopts an asynchronously momentum-updated key encoder and an online query encoder, along with a memory queue to store a large number of consistent negative samples. Specifically, different augmentations are applied to the skeleton x to generate the positive pair (x_q, x_k) , while the negative sample features m_i are stored in a memory queue. The model is constrained to retrieve the positives among the negative samples, optimizing the following InfoNCE objective $\mathcal{L}_{Info}(z_q, z_k)$:

$$\mathcal{L} = -\log \frac{\exp(z_q \cdot z_k / \tau)}{\exp(z_q \cdot z_k / \tau) + \sum_{i=1} \exp(z_q \cdot m_i / \tau)}, \quad (2)$$

where the z_q/z_k is the query/key embedding encoded by the query/key model and τ is the temperature hyper-parameter. m_i is the i -th feature anchor as the negative sample. Based on this, we review current skeleton contrastive learning methods from three aspects of design, *i.e.*, *augmentation design and learning strategy*, *objective engineering*, and *cross-modal learning*.

1) Augmentation Design and Learning Strategy. Data augmentation exposes novel motion patterns and generates diverse positive views, which have been found crucial to the success of contrastive learning (Tian et al., 2020; Guo et al., 2022). Unlike images, data augmentation for skeleton is relatively less developed. Therefore, the earlier works mainly focus on exploring practical spatial-temporal augmentations for skeleton contrastive learning (Rao et al., 2021; Thoker et al., 2021; Gao et al., 2021). For example, Rao et al. (2021) proposed a series of augmentations including *Rotation*, *Shear*, *Reverse*, *Gaussian Noise*, *Gaussian Blur*, *Joint Mask* and *Channel Mask*, some of which are used as the default basic augmentations in the future research. Notably, the optimal augmentations for different backbones are usually different. For example, the *Joint Mask* is found detrimental for GCNs (Zhang et al., 2023), while beneficial for GRU model (Thoker et al., 2021), which implies that GCNs are more sensitive to the spatial corruption. Generally, this difference arises on account of different modeling mechanisms and model capacities, which further increases the difficulty of the augmentation design and selection.

Inspired by the success of mixing-based augmentations for images, Zhan et al. (2022) proposed SkeleMixCLR equipped with *SkeleMix* augmentation which combines the topological information of different skeleton sequences. To learn from this mixed skeleton, SkeleMixCLR obtains the corresponding part level and the whole-body level features and pursues the local-global invariance. Based on this, SkeAttnCLR (Hua et al., 2023) further designs an attention mechanism to perform local contrastive learning on salient and non-salient features. Due to the generated novel input views and regularization effect on the feature space, mixing augmentation often leads to consistent improvement for representations.

To further improve consistency learning, researchers have made endeavors to introduce more and stronger data augmentations. However, the ensuing problem is the over-distortion (Bai et al., 2022) of the augmented data, leading to the model performance degradation. In other words, some strong augmentations would seriously corrupt the semantic information, change the data distribution, and result in the difficulty of model consistency learning. To address this, AimCLR (Guo et al., 2022) utilizes two branches to encode the weakly and strongly augmented views, respectively. Then the model optimizes the similarity of the distributions output by two branches as a soft consistency learning target:

$$\begin{aligned} \mathcal{L}_{Soft}(z) &= KL(p(z|z_{weak})||p(z|z_{strong})), \\ p(z|z_*) &= \frac{\exp(z \cdot z_*/\tau)}{\exp(z_k \cdot z_*/\tau) + \sum_{i=1}^M \exp(m_i \cdot z_*/\tau)}, \end{aligned} \quad (3)$$

where z_{weak} and z_{strong} are the corresponding embeddings of weakly and strongly augmented views. Instead of the one-hot target, this objective utilizes the similarity distribution

as the soft target to guide consistency learning, which can be viewed as a self-distillation process. Further, (Zhang et al., 2023) introduced more strong augmentations, *e.g.*, randomly dropping the skeleton edges and joints, and proposed a hierarchical contrastive learning framework. It performs a decoupled progressive augmentation invariance learning by optimizing the consistency only between the augmented samples with adjacent strength, where the weakly augmented branch serves as the mimic target of the strongly augmented branch. These works show that the weakly augmented views can effectively guide the learning of the corresponding strongly augmented samples, leading to more stable representation learning and improvement.

In addition, (Lin et al., 2023) explicitly distinguished the static regions and motion regions, namely, *actionlet*, in human skeleton, and introduced a motion-aware augmentation strategy. By mining the actionlet in an unsupervised manner, the semantic-reserving augmentations are employed for actionlet regions, while the noise perturbations for non-actionlet regions, avoiding the over-distortion problem. On the other hand, instead of augmenting at the input level, HaLP (Shah et al., 2023) proposes a latent positive hallucinating method by exploring the latent space around the corresponding prototype.

In summary, the development of data augmentations along with the corresponding learning strategy significantly boosts the performance of skeleton contrastive learning, which has always been an important and popular topic for 3D skeleton contrastive learning.

2) Objective Engineering. In addition to the widely used InfoNCE loss, some SSL works have also made efforts to explore new loss functions for extra regularization or new objectives. Here we introduce them from the perspective of *False Negative* (FN) problem, which widely exists in skeleton contrastive learning based on negative examples, *e.g.*, MoCo v2. Concretely, false negatives refer to the negative samples but from the same semantic category. Traditional contrastive learning relies on the one-hot label and directly pushes them away in the feature space, which forces the model to discard the shared semantic information and leads to slow convergence (Huynh et al., 2022). We point out that, due to the lack of description of objects and backgrounds, there are fewer action categories represented by skeleton data, which leads to a more serious FN problem. To this end, the following aspects are considered to tackle this issue.

The first straightforward solution is based on the false negative detection and cancellation (Chen et al., 2021; Huynh et al., 2022; Li et al., 2021; Guo et al., 2022; Zhang et al., 2022). These methods first calculate the similarity between the sample and negatives, selecting the top k negatives that are most similar as the potential false negatives, which are then involved as extended positives in contrastive learning.

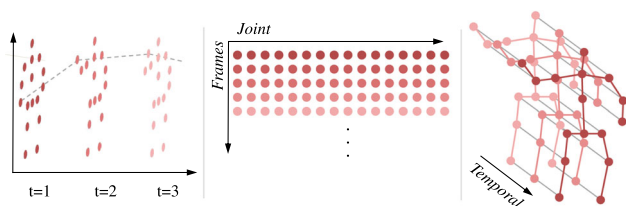


Fig. 4 Different representations of skeleton data. From left to right are *time series*, *2D pseudo-image*, *spatial-temporal graph*.

Some other methods (Xie et al., 2022a; Mao et al., 2022; Zhang et al., 2023) suggest employing the distillation objective as an adaptive re-weighting regularization for the one-hot instance discrimination pretext task, which can be formulated as follows:

$$\begin{aligned} \mathcal{L}_{KD}(z_q, z_k) &= -p(z_k, \tau_k) \log p(z_q, \tau_q), \\ p_j(z, \tau) &= \frac{\exp(z \cdot m_j / \tau)}{\sum_{i=1} \exp(z \cdot m_i / \tau)}. \end{aligned} \quad (4)$$

They assign the attraction weights to negatives based on the calculated similarity. Specifically, if a sample possesses high similarity with the positive, a larger attraction weight would be assigned to involve it in similarity optimization.

Besides, prototype-based contrastive learning is also studied for high-level semantic consistency learning (Li et al., 2020; Caron et al., 2020; Xu et al., 2021a). It usually performs clustering in feature space to assign instances to different cluster prototypes as a pseudo-semantic label. Then the model learns more high-level semantics by contrasting different prototypes. By virtue of such a way, the model pays more attention to the cluster-level discrimination task rather than instance-level to alleviate the false negative problem.

In addition to the above designs, as introduced above, some negative-sample-free contrastive frameworks have been developed recently, which avoid the difficulty of explicitly specifying negative examples. Based on BYOL (Grill et al., 2020), the works (Franco et al., 2023; Chen et al., 2023) performs the positive-only consistency learning in the hyperbolic space. Some other works (Zhang & Hou, 2022; Sun et al., 2023), inspired by the feature decorrelation concept, propose to learn the decorrelated representations based on Barlow twins (Zbontar et al., 2021) and Variance-Invariance-Covariance Regularization (VICReg) (Bardes et al., 2021) as the framework.

3) Cross-Modal Learning. Skeleton naturally provides different data modalities to represent human movement, *e.g.*, *motion* and *bone* modality defined in the previous works (Shi et al., 2019; Chen et al., 2021). Meanwhile, many works (Shi et al., 2019; Chen et al., 2021) have proved that the fusion of different modal knowledge can be beneficial to achieve richer and more representative action depictions. For example, *joint* contains essential information about human movements

while lacking the explicit encoding of some high-order information. *Bone* view can capture the structural patterns well. However, it is subject to the predefined human model and still lacks temporal information. *Motion* explicitly encodes the temporal dynamics while losing the spatial information. It is promising to explore these different modalities to achieve better action representation.

CRRL (Wang et al., 2022) performs contrastive learning between the joint and motion modalities, where the encoders for motion and joint data are homologous, *i.e.*, obtained by momentum-updated strategy. However, this design can be unreasonable because the model can have difficulty dealing with two modalities simultaneously. In contrast, Cross-CLR (Li et al., 2021) adopts three separate encoders for joint, bone, and motion data, and aims to align the distributions among the neighborhoods in latent space. Intuitively, it encourages the sample embedding together with its neighbors of modality v should also be close in the latent space of modality u . However, this leads to a two-stage training process, as the model needs to be pre-trained on a single view first to obtain reliable unique modal knowledge. To this end, CMD (Mao et al., 2022) proposes a general bidirectional distillation objective, where the modal knowledge is directly modeled as the whole similarity distribution in the customized latent space. The proposed cross-modal distillation objective and the instance discrimination task jointly optimize the model, yielding a concise single training phase. On the other hand, to get rid of the limitation of using separate encoders for different modalities, UmURL (Sun et al., 2023) develops a unified modality-agnostic encoder, which can handle different modal inputs relying on the modality-specific input embedding layer and feature projection layer.

Meanwhile, we can also organize skeleton data into different representations as shown in Fig. 4, *e.g.*, the temporal series, 2D pseudo-images with frame and joint dimensions, and spatial-temporal graphs with nodes and edges. Following this concept, ISC (Thoker et al., 2021) encodes these different skeleton representations with different model backbones, which are then projected into a shared latent space. A cross-modal contrastive learning loss is applied for pre-training.

4) Others. In addition to the aforementioned literature, some works also employ contrastive learning. Considering the hierarchical structure of skeleton data, (Dong et al., 2023) performs contrastive learning in a hierarchical manner to model features of different levels. Some works (Lin et al., 2020; Xu et al., 2021a; Tanfous et al., 2022; Wang et al., 2022) combine the contrastive learning paradigm with others in a multi-tasking manner. Notably, in addition to simply combining reconstruction and contrastive learning pretext tasks, PCM³ Zhang et al. (2023) proposes a collaborative design to further improve the representation learning. Eq-Contrast (Lin et al., 2024) formulates the temporal sorting task into an equivariant contrastive learning objective for multi-task pre-

Fig. 5 Summary of the advantages/limitations of different skeleton-based SSL methodologies.

Taxonomy	Context-based	Contrastive Learning	Masked Skeleton Modeling
Pros	<ul style="list-style-type: none"> - Simple and flexible - Training-friendly 	<ul style="list-style-type: none"> - Strong and discriminative representations - Effective for different backbones 	<ul style="list-style-type: none"> - Strong generalization performance (fine-tuning) - Compatibility with Transformer architectures
Cons	<ul style="list-style-type: none"> - Limited generalization - Weak semantic understanding - Lack of general design guidance 	<ul style="list-style-type: none"> - Augmentation sensitive - False negative problem - Dimensional collapse - Expensive training cost 	<ul style="list-style-type: none"> - Critical of model backbone - Masking strategy sensitive - Low training efficiency

training. Meanwhile, not limited to SSL, contrastive learning is also studied in supervised (Zhou et al., 2023a) and semi-supervised (Binqian et al., 2022; Shu et al., 2022) learning for skeleton.

2.3 Summary and Discussion

The skeleton-based SSL literature is categorized into context-based, generative learning, and contrastive learning types. All these methods aims to capture joint features as well as their relationships from unlabeled skeleton data, to obtain a meaningful representation of the motion. Due to the lack of ground-truth labels, these methodologies introduce prior knowledge in different means to enable representation learning, *e.g.*, context task design, augmentation choices, and masking strategy. To summary the different skeleton-based SSL methods based on our taxonomy, we present the advantages and limitations of each methodology category in Fig. 5.

Context-based methods are often simple to implement, flexible for design, and training-friendly compared with the other two paradigms. However, as evidenced by the previous literature, the main concern is the limited model capacity of generalization and semantic understanding due to the gap between the low-level pretext tasks and the downstream tasks. Contrastive learning is recognized by its highly discriminative representations with strong performance, especially under linear evaluation protocol. More importantly, it has been proven effective for different backbones for skeleton modeling. The other side of the coin is augmentation sensitivity, false negative problem, and high training cost as previously discussed. Meanwhile, some researchers point out the concern of dimensional collapse in contrastive learning (Jing et al., 2022; Lin et al., 2024), where the embedding vectors only span a lower dimensional space, limiting the representation capacity. Finally, the masked skeleton modeling is discussed considering its dominant position in generative learning. It is well compatible with Transformer, with expected strong generalization performance, especially under fine-tuning protocol. However, performance can significantly drop when employing other backbones, which can limit its deployment in some low-latency, resource-constrained scenarios. Meanwhile, it is

also concerned about masking strategy sensitivity and time-consuming training.

To end this section and possibly motivate future work, we further summarize and emphasize the special design of skeleton-based SSL approaches that go beyond the high-level SSL methodology taxonomy.

- *Context Pretext Tasks Based on Human Topology and Kinematics.* Different from images, skeleton pretext tasks focus on the spatial-temporal data structure learning of human body, *e.g.*, temporal sorting, angle, and displacement prediction.
- *Data Augmentation for Novel Motion Patterns.* Skeleton augmentation has always been the focus to boost contrastive learning. These efforts include the spatial and temporal, invariant and equivariant (Lin et al., 2024), weak and strong (Guo et al., 2022; Zhang et al., 2023), intra and inter augmentations (Zhan et al., 2022; Zhang et al., 2023) as well as the corresponding learning strategy.
- *Cross-Modal Knowledge Modeling.* Skeleton naturally provides different data modalities (or views) to represent human movement, *e.g.*, joint, motion, and bone, enabling cross-modal knowledge modeling for better action representations.
- *False-Negative Cancellation.* As discussed, the diversity of human motions is much less than natural images, leading to a more serious false negative problem, which is an important topic in skeleton contrastive learning.
- *Masking Strategy in MSM.* As a crucial component in MSM, spatial-temporal masking strategy has also been widely studied specifically for the highly structured human skeleton data.
- *Multi-Granularity Representation Learning.* Skeleton data naturally presents a hierarchy, *e.g.*, sequence-level, clip-level, and joint-level. It can be beneficial to capture these action clues at different scales, enriching the action representation and alleviating the over-fitting problem.

3 The Proposed Method

3.1 Motivation

By reviewing the previous works, we find that most existing skeleton SSL methods employ a single paradigm or simply combine different methods with action recognition task for evaluation solely, which limits the model capacity of learn-

Table 1 Main notations and the definition.

Notation	Definition
$s'(s_{ach}), z'(z_{ach})$	The global key-branch intra-augmented skeleton and its embedding.
s_{intra}, z_{intra}	The query-branch intra-augmented skeleton and its embedding.
s_{inter}, z_{inter}	The query-branch inter-augmented skeleton and its embedding.
s_{clip}, z_{clip}	The query-branch temporal-augmented skeleton clip and its embedding.
s_{mask}, z_{mask}	The query-branch masked skeleton and its embedding.
$s_{predict}, z_{predict}$	The reconstructed skeleton from s_{mask} and its embedding.
z'_{inter}	The linear interpolation of two skeleton embeddings in mix inter-augmentation.
P_*	The prompts of different views.
$f(\cdot), h(\cdot), dec(\cdot)$	The encoder, projector, and decoder.

ing representations at different granularity and levels and leaves the generalization capacity leashed. To this end, we propose prompted contrast with masked motion modeling (PCM³⁺⁺), which combines contrastive learning and MSM by exploring potential collaboration. It enables versatile representation learning by joint, clip, and sequence level feature modeling, and significantly improves the performance of different downstream tasks.

Specifically, our method is based on our previous work (Zhang et al., 2023), which combines contrastive learning and masked skeleton modeling in a synergetic manner. Based on this, we further extend and strengthen the design with the key idea, *i.e.*, versatile action representation learning. In the following, we first introduce the contrastive learning (Sec. 3.2) and MSM (Sec. 3.3) scheme involved in baseline method as well as their collaboration design (Sec. 3.4). Based on this, we introduce our new improvements in PCM³⁺⁺ compared to Zhang et al. (2023). Concretely, an effective clip-level contrastive learning method is designed in Sec. 3.2.(3), which significantly boosts the capacity of short-term motion modeling as illustrated in Fig. 6. Besides, a post-distillation strategy is developed in Sec. 3.5.(2) to further boost the representation compactness. These advancements further improve the representation generalization for various downstream tasks. For clarity, the main notations are summarized in Table 1.

3.2 Skeleton Contrastive Learning

Our pipeline for contrastive learning follows MoCo V2 He et al. (2020). For a positive pair (z, z') , the model optimizes the InfoNCE objective defined in Eq. (2) along with a distillation

objective in Eq. (4) as regularization:

$$\mathcal{L}_{CL}(z, z') = \mathcal{L}_{Info}(z, z') + \mathcal{L}_{KD}(z, z'). \tag{5}$$

The intuition behind this is to combine the one-hot hard labels in InfoNCE to explicitly cluster features while avoiding collapse, and the soft labels in the distillation objective, *i.e.*, the similarity scores, to achieve fine-grained relational knowledge learning, which also alleviates the false negative problem as discussed in Sec. 2.2.3.(1).

To boost the consistency learning, we propose a series of effective spatial-temporal augmentations to construct diverse positive pairs, which are introduced as follows.

1) Spatial Intra- and Inter- Skeleton Contrastive Learning. For intra-skeleton augmentations, we utilize *Temporal Crop-Resize*, *Shear*, and *Joint Jittering* to construct the positive pair (s_{intra}, s') . Then, we obtain the corresponding 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, *Mixing* augmentations are adopted to construct inter-transformed views. Specifically, given two skeleton sequences s_1, s_2 , we obtain the mixed skeleton data s_{inter} , *e.g.*, by *Mixup*. Subsequently, we can obtain the embeddings corresponding to the mixed data by $z_{inter} = h_q \circ f_q(s_{inter})$. The optimized objective is $\mathcal{L}_{CL}(z_{inter}, z'_{inter})$ in Eq. (5), where $z'_{inter} = (1 - \lambda)(h_k \circ f_k(s_1)) + \lambda(h_k \circ f_k(s_2))$.

2) Temporal Asymmetric Clip Contrastive Learning. The above designs only consider the sequence-level consistency modeling, *i.e.*, the augmented data often contains sufficient temporal motion information. However, it ignores the *short-term* motion modeling, which is necessitated for downstream tasks of dense prediction such as action detection. On the other hand, the consistency learning for short-term motions, *e.g.*, clips, which can be viewed as challenging temporal augmented views, can also benefit the whole representation quality learned by model. To this end, we present the clip-level contrastive learning design.

To sample a motion clip, we can apply the *Temporal Crop* transformation, and the obtained clip is much shorter (10% ~ 40%) than the original sequence. However, the clip only contains partial motion information and directly aligning the semantic embeddings of two clips is difficult for model, leading to the unstable training. Therefore, we introduce an asymmetric design, which only feeds the short clip into query branch, while the key branch takes the pre-crop original sequence as an anchor. Meanwhile, we keep the two augmented data possessing the same spatial views to decouple the temporal variation. Specifically, we first apply the intra-augmentations to a sample s to generate the s_{ach} . Then we further utilize the *Temporal Crop* to sample a continuous clip s_{clip} from s_{ach} . It means that, s_{ach} is subsuming

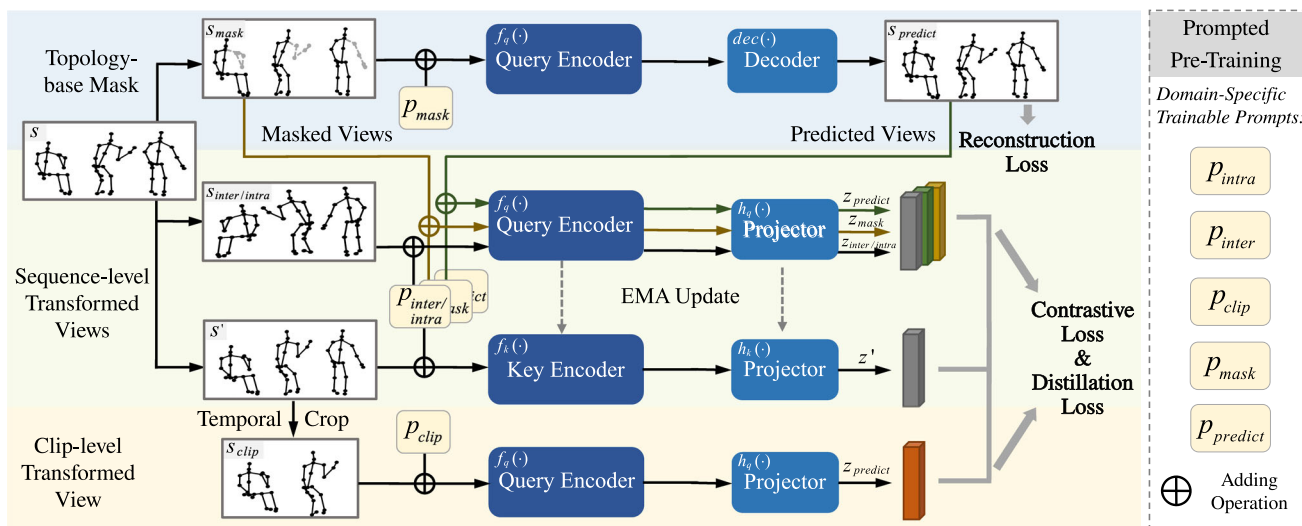


Fig. 6 The overview of the proposed method. We integrate masked skeleton modeling and contrastive learning paradigms to learn the joint- (blue part), clip- (yellow part), and sequence-level (green part) representations. Meanwhile, a series of domain-specific prompts are introduced as training guidance.

s_{clip} in temporal dimension while sharing exactly the same spatial transformations, to provide a precise and feasible target for clip-level consistency learning. Similarly, the model takes the positive pair (z_{clip}, z_{ach}) for \mathcal{L}_{CL} optimization in Eq. (5), where $z_{clip} = h_q \circ f_q(s_{clip})$. In implementation, we just mask $z_{ach} = z'$ for efficiency.

3.3 Masked Skeleton Prediction

To model the joint-level representations, we utilize the masked skeleton modeling task with a segment-wise masking strategy at body part level. To predict the masked regions from masked skeleton s_{mask} , we employ a decoder $dec(\cdot)$ taking the representations from the encoder as input. The MSE loss between original data s and predicted data $s_{predict}$ by decoder is optimized in the masked regions:

$$\mathcal{L}_{Mask} = \mathbb{E} (\| (s - dec \circ f_q(s_{mask})) \odot (\mathbf{1} - M) \|_2), \quad (6)$$

where M is the binary mask and $\mathbf{1}$ is an all-one matrix with the same shape as M .

3.4 On the Connection of Contrastive Learning and Masked Prediction

As discussed in the work (Zhang et al., 2023), simply combining these two tasks can be sub-optimal due to the inherent gap of their feature modeling mechanisms (Qi et al., 2023; Gui et al., 2023). To this end, we explore the potential synergy for exploitation between them.

1) Novel Positive Pairs as Connection. We utilize special data views during the masked prediction training to pro-

vide more diverse positive samples for contrastive learning. First, the masked skeleton view s_{mask} naturally simulates the occlusion for skeletons, serving as challenging positives. Meanwhile, we also take the predicted skeleton view $s_{predict}$ output by decoder $dec(\cdot)$ as positive samples. It contains the inherent uncertainty and diversity brought by continuous training of the model, which contributes to encoding more diverse movement patterns. On the other hand, the semantic consistency of the output skeleton with respect to the model itself is encouraged, i.e., the predicted view can also be perceived well by the encoder, connecting the low-level reconstruction with the high-level semantic modeling.

In a nutshell, we utilize the masked view s_{mask} and the predicted view $s_{predict}$ as positives. Together with the manually constructed positive pairs in Sec. 3.2, we present all positive (embedding) pairs $\{(z_q, z_k)\}$ as follows:

$$\{(z_q, z_k)\} = \{(z_{intra}, z'), (z_{inter}, z'_{inter}), (z_{clip}, z'), (z_{mask}, z'), (z_{predict}, z')\}. \quad (7)$$

They are fed into the encoders to optimize the \mathcal{L}_{Info} and \mathcal{L}_{KD} in Eq. 5.

2) High-Level Semantic Guidance. Note that we feed the predicted view into the contrastive learning pipeline. The gradients of $s_{predict}$ from the contrastive learning branch are propagated to update the reconstructed decoder $dec(\cdot)$. It provides the high-level semantic guidance for the skeleton prediction together with the joint-level MSE loss in Eq. (6), leading to better masked prediction learning and higher quality of $s_{predict}$ as positive samples.

With these synergetic designs, the masked skeleton modeling provides novel positives to the contrastive learning,

boosting the model generalization capacity. On the other hand, the supervision signals as gradients of contrastive learning branch also assists the masked prediction branch with the high-level semantic guidance provided by inter-sample contrast. These designs alleviate the gap of feature modeling mechanism of these two tasks, which limits the performance of previous works that directly combine them, and hence yield better representation quality.

3.5 The Whole Training Strategy

Overall, the model jointly optimizes the contrastive learning and masked skeleton modeling. Here we present the overall objective and the training strategy. For clarity, we utilize \mathcal{L}_{Info}^{all} and \mathcal{L}_{KD}^{all} to represent the sum of the losses for each positive pair defined in Eq. (7). It can be formulated as:

$$\mathcal{L}_{Info}^{all} = \sum_{(z_q, z_k)} \mathcal{L}_{Info}^{query}, \quad \mathcal{L}_{KD}^{all} = \sum_{(z_q, z_k)} \mathcal{L}_{KD}^{query}.$$

$\mathcal{L}_{Info}^{query}$ and \mathcal{L}_{KD}^{query} are the respective terms for the specific query view summarized in Eq. (7). During training, the following objective is applied to the whole model:

$$\mathcal{L} = \mathcal{L}_{Info}^{all} + \lambda_m \mathcal{L}_{Mask} + \lambda_{kd} \mathcal{L}_{KD}^{all}, \quad (8)$$

where λ_m and λ_{kd} are the loss weight. Based on this, we further propose two advanced training strategies to boost the representation learning.

1) Prompted Multi-Task Pre-Training. Considering the difficulty of encoding different data views simultaneously, we employ prompt-based guidance to assist the model to learn from different data views explicitly. Specifically, we attend to a series of *domain-specific prompts* for different augmented views, *i.e.*, p_{intra} , p_{inter} , p_{clip} , p_{mask} , and $p_{predict}$. Their dimension equals to the spatial size of the skeleton data. Then, these domain-specific prompts are added to the corresponding skeleton (s_* means * view of s):

$$s_* = s_* + p_*. \quad (9)$$

These decorated skeletons are fed into the query/key encoder for self-supervised pre-training, providing the training guidance and achieving better representations.

2) Post-Distillation Refinement. To further improve the representation quality, we introduce a post-distillation strategy as an optional refinement process. After obtaining a good prior feature space by Eq. (8), we directly remove the one-hot label constraint in InfoNCE objective \mathcal{L}_{Info}^{all} , *i.e.*,

$$\mathcal{L} = \lambda_m \mathcal{L}_{Mask} + \lambda_{kd} \mathcal{L}_{KD}^{all}, \quad (10)$$

where we only apply distillation term that assigns the attraction weights adaptively according to the similarity. This can be seen as a more explicit feature clustering process to obtain a more compact representation space by alleviating the *false negative* problem. Note this single distillation objective would not cause model collapse due to the well initialized representation space learned in previous pre-training stage.

3.6 Discussion

We summarize our method as a direct response to the challenges mentioned in previous literature review. We target at the versatile action representation learning, adopting our previous work (Zhang et al., 2023) as the baseline for its effectiveness to learn multi-granularity features. Building on this, we further integrate the clip-level contrastive learning with it to present a more comprehensive solution with joint, clip, and sequence level modeling.

As a response to previously discussed challenges, 1) we carefully design the spatial-temporal augmentations involving intra- and inter- data samples to present a strong contrastive learning baseline, which eases the efforts of subsequent works. 2) Further, the proposed multi-granularity representation modeling can effectively enhance the feature diversity, alleviating the dimensional collapse implicitly according to Zhang and Wang (2022). 3) Meanwhile, for the false negative problem, we employ a different perspective, *i.e.*, post-training to deal with it. In the following sections, we will present the quantitative analysis on these designs.

4 Benchmark Skeleton SSL Methods

We provide a comprehensive benchmark of existing methods, along with the skeleton datasets, backbone architectures, and downstream tasks in this section. Meanwhile, the evaluation of proposed PCM³⁺⁺ is reported.

We first introduce the following datasets, which are widely used in skeleton-based SSL evaluation. The summary of more datasets for skeleton action understanding can be found in Supp.

1) NTU RGB+D 60 Dataset (NTU 60) Shahroudy et al. (2016) is the most popular skeleton dataset. There are 56,578 videos with 25 joints for a human, captured by three Microsoft Kinect v2 cameras. The skeleton sequences are divided into 60 action categories, performed by 40 volunteers. Two evaluation protocols are recommended: a) Cross-Subject (xsub): the data for training are collected from 20 subjects, while the other 20 subjects are for testing. b) Cross-View (xview): the training set consists of front and two side views of the action performers, while testing set includes the left and right 45 degree views.

2) NTU RGB+D 120 Dataset (NTU 120) Liu et al. (2019) is an extension to NTU 60 dataset. There are 114,480 videos collected with 120 action categories, performed by 106 subjects. Meanwhile, 32 collection setups with respect to the location and background are used to build the dataset. Two recommended protocols are presented: a) Cross-Subject (xsub): the data for training are collected from 53 subjects, while the testing data are from the other 53 subjects. b) Cross-Setup (xset): the training data uses even setup IDs, while testing data are odd setup IDs.

3) PKU Multi-Modality Dataset (PKUMMD) Liu et al. (2020) is another large-scale benchmark with available skeleton data. Two subsets, Part I and Part II, are provided. PKUMMD Part I contains 1,076 long video sequences, with 20 action labels per video on average, and $\sim 20,000$ instances are included in 51 action categories after trimming. Part II contains 2000 short video sequences with approximately seven instances each, focusing on the short-margin action detection task. It is more challenging due to the data noise and view variation.

4.1 Model Backbones

Different backbones are studied in previous skeleton SSL works, *i.e.*, RNN-based, GCN-based, and Transformer-based models. RNNs treat the skeleton sequences as the temporal series and model temporal dependencies. However, it ignores the spatial structures of skeleton. Inspired by the natural topology structure of the human body, GCNs are widely explored to model spatial-temporal relationships. Recently, Transformer has been utilized to capture the long-temporal dependencies and has demonstrated remarkable results, owing to the attention mechanism.

Besides, some SSL works turn to other model backbones. Convolutional neural networks (CNNs) are utilized to process the skeleton sequence as a pseudo-2D image. Meanwhile, skeletons can also be treated as point clouds, and hence the Dynamic Graph CNN (DGCNN) (Wang et al., 2019) is also explored as the feature extractor.

4.2 Downstream Tasks

1) Action Recognition is the most common SSL evaluation task. Typically, two evaluation protocols are widely adopted. The first is the linear evaluation protocol, where a linear layer is added with the pre-trained model fixed. The other is fine-tuning protocol where the whole model is trained including the subsequent linear layer. Top-k accuracy metric is adopted in this task.

2) Action Retrieval aims to find the skeleton sequences that are similar or near-duplicates of a given query sequence. The metric precision, which is the proportion of retrieved relevant skeleton sequences in all retrieved entries, is reported.

3) Action Detection is a task that detects the start and end time of actions in an untrimmed skeleton sequence as well as its corresponding action label. It can also be referred to as the *Temporal Action Localization* or *Action Segmentation*. Following the previous works (Luo et al., 2018; Liu et al., 2020), the Average Precision (mAP) at different temporal Intersection over Union (tIoU) thresholds between the predicted and the ground truth intervals is utilized as the metric.

4) Occluded Action Recognition focuses on the action understanding with occlusion, which is prevalent in human activity. Typically, the skeleton joints with low confidence or known to be occluded are set to zeros. The model is constrained to predict the correct label from corrupted skeleton sequences with top-k accuracy as the metric.

4.3 Implementations Details

For the proposed method, we follow the implementation and experiment settings of previous works (Zhang et al., 2023; Mao et al., 2022). For the existing methods, we give priority to the reported results in their original paper. Meanwhile, we reproduce some results to keep the fairness of the comparison and implementation. Details can be found in Supp.

4.4 A Multi-Task Benchmark for Skeleton-based SSL

In this section, we present our multi-task benchmark for skeleton-based SSL methods, which is the first time to the best of our knowledge, to benefit the community. Considering most previous SSL works are evaluated only on action recognition, we split the benchmark into two parts for presentation, *i.e.*, a full comparison on action recognition task and the reproduction benchmark on other tasks. First, we curate a comprehensive comparison of the widely adopted action recognition task in Table 2 with the backbone and pretext task category mark.

On the other hand, for other different tasks, an available benchmark is provided covering action retrieval, occluded recognition, action detection, and few-shot learning tasks. We select the recent popular works with official code released, and make huge efforts to reproduce them on the same settings to present a multi-task comparison as shown in Table 3.

4.4.1 Comparison on Action Recognition Task

1) Across Pretext Tasks. As shown in Fig. 7, context-based and generative learning pretext tasks are mainly studied in the earlier works. However, these methods usually fail to achieve a satisfactory performance under linear evaluation due to the modeling of too many low-level features and the lack of effective design. Recently, MAMP (Mao et al., 2023) adopts MSM with large mask ratios and motion-aware reconstruction targets, obtaining a highly discriminative semantic feature space

Table 2 Comparison of skeleton SSL works. P, G and C represents context-based (Pseudo-label), Generative, and Contrastive learning methodologies. We report the best accuracy in the original paper. *s means the fusion results of * streams, and the single joint stream is adopted by default.

Method	Publish	Backbone	Feature Dimension	Pretext Task P G C	NTU 60 (%)		NTU 120 (%)		
					xsub	xview	xsub	xset	
<i>Linear Evaluation Protocol (Arranged by Backbone Model and Publish Year)</i>									
LongT GAN (Zheng et al., 2018)	AAAI 2018	GRU	800	o ● o	39.1	52.1	35.6	39.7	
P&C (Su et al., 2020)	CVPR 2020	GRU	1024×2	o ● o	50.7	76.1	41.1	44.1	
2s-SeBiReNet (Nie et al., 2020)	ECCV 2020	GRU	32	o ● ●	-	79.7	-	69.3	
MS ² L (Lin et al., 2020)	ACM MM 2020	GRU	600	● ● ●	52.6	-	-	-	
PCRP (Xu et al., 2021a)	TMM 2021	GRU	1024	o ● ●	54.9	63.4	43.0	44.6	
AS-CAL (Rao et al., 2021)	Info. Sciences 2021	LSTM	256	o o ●	58.5	64.8	-	-	
ISC (Thoker et al., 2021)	ACM MM 2021	GRU+GCN	1024×2	o o ●	76.3	85.2	67.1	67.9	
CRRL (Wang et al., 2022)	TIP 2022	GRU	300	o ● ●	67.6	73.8	57.0	56.2	
3s-CMD (Mao et al., 2022)	ECCV 2022	GRU	1024×2	o o ●	84.1	90.9	74.7	76.1	
3s-CSTCN (Wang et al., 2023)	TMM 2023	GRU	1024×2	o o ●	85.8	92.0	77.5	78.5	
3s-HiCo (Chen et al., 2022)	AAAI 2023	GRU	512×8	o o ●	82.6	90.8	75.9	77.3	
HaLP (Shah et al., 2023)	CVPR 2023	GRU	1024×2	o o ●	79.7	86.8	71.1	72.2	
3s-PCM ³ (Zhang et al., 2023)	ACM MM 2023	GRU	1024×2	o ● ●	87.4	93.1	80.0	81.2	
3s-Eq-Contrast (Lin et al., 2024)	TIP 2024	GRU	1024×2	● o ●	87.0	92.9	79.4	81.2	
3s-PCM³++	- (This Paper)	GRU	1024×2	o ● ●	88.1	93.5	80.3	81.6	
3s-CrosSCLR (Li et al., 2021)	CVPR 2021	GCN	256	o o ●	77.8	83.4	67.9	66.7	
4s-MG-AL (Yang et al., 2022)	TCSVT 2022	GCN	-	● o o	64.7	68.0	46.2	49.5	
3s-AimCLR (Guo et al., 2022)	AAAI 2022	GCN	256	o o ●	78.9	83.8	68.2	68.8	
3s-CPM (Zhang et al., 2022)	ECCV 2022	GCN	256	o o ●	83.2	87.0	73.0	74.0	
Chen et al (Chen et al., 2023)	TIP 2023	GCN	256	o o ●	78.9	82.3	68.4	67.3	
3s-HiCLR (Zhang et al., 2023)	AAAI 2023	GCN	256	o o ●	80.4	85.5	70.0	70.4	
3s-PSTL (Zhou et al., 2023b)	AAAI 2023	GCN	256	o o ●	79.1	82.6	69.2	70.3	

Table 2 continued

Method	Publish	Backbone	Feature Dimension	Pretext Task		NTU 60 (%)		NTU 120 (%)	
				P	G C	xsub	xview	xsub	xset
3s-SkeAttnCLR (Hua et al., 2023)	IJCAI 2023	GCN	256	o	o o •	82.0	86.5	77.1	80.0
3s-HYSP (Franco et al., 2023)	ICLR 2023	GCN	256	o	o o •	79.1	85.2	64.5	67.3
3s-ActCLR (Lin et al., 2023)	CVPR 2023	GCN	256	o	o o •	84.3	88.8	74.3	75.7
3s-RVTCLR+ (Zhu et al., 2023b)	ICCV 2023	GCN	256	o	o o •	79.7	84.6	68.0	68.9
2s-ViA (Yang et al., 2024)	IJCV 2024	GCN	256	o	o o •	78.1	85.8	69.2	66.9
H-Transformer (Cheng et al., 2021)	ICME 2021	Transformer	2048	•	o o o	69.3	72.8	-	-
GL-Transformer (Kim, 2022)	ECCV 2022	Transformer	48×25	•	o o o	76.3	83.8	66.0	68.7
MAMP (Mao et al., 2023)	ICCV 2023	Transformer	256	o	o • o	84.9	89.1	78.6	79.1
3s-UmURL (Sun et al., 2023)	ACM MM 2023	Transformer	2048	o	o o •	84.4	91.4	75.9	77.2
MacDiff (Wu et al., 2024)	ECCV 2024	Transformer	256×25	o	o • o	86.4	91.0	79.4	80.2
IGM (Lin et al., 2024)	ECCV 2024	Transformer	256×25	o	o • •	86.2	91.2	80.0	81.4
3s-USDRL (Weng et al., 2025)	AAAI 2025	Transformer	4096	o	o o •	87.1	93.2	79.3	80.6
AE-L (Paoletti et al., 2022)	BMVC 2021	CNN	256	•	o • o	69.9	85.4	59.1	62.4
3s-Colorization (Yang et al., 2021)	ICCV 2021	DGCNN	1024	o	o • o	75.2	83.1	64.3	67.5
3s-Masked Colorization (Yang et al., 2023)	TPAMI 2023	DGCNN	1024	o	o • o	79.1	87.2	69.2	70.8
<i>Fully Fine-tuning Protocol (Arranged by Publish Year)</i>									
MCC (Su et al., 2021)	ICCV 2021	GCN	256	o	o • •	83.0	89.7	77.0	77.8
3s-Hi-TRS (Chen et al., 2022)	ECCV 2022	Transformer	512	•	o • •	90.0	95.7	85.3	87.4
3s-Masked Colorization (Yang et al., 2023)	TPAMI 2023	DGCNN	1024	o	o • o	89.1	95.9	81.2	82.4
MAMP (Mao et al., 2023)	ICCV 2023	Transformer	256×25	o	o • o	93.1	97.5	90.0	91.3
MotionBERT (Zhu et al., 2023a)	ICCV 2023	Transformer	512	o	o • o	93.0	97.2	-	-
SSL (Yan et al., 2023)	ICCV 2023	GIN	-	o	o • o	92.8	96.5	84.8	85.7
3s-PCM³++	- (This paper)	GRU	1024×2	o	o • •	88.5	93.9	-	-

Table 3 Benchmark on multiple downstream tasks. *UmURL adopts hard-coded input temporal length and the performance drops significantly when handling the fine-grained temporal understanding task. † indicates more input frames are used.

Method	Retrieval				Occluded		Detection		Few Shot	
	NTU 60		NTU 120		NTU 60 xview		PKUMMD		NTU 120 xset	
	xsub	xview	xsub	xset	Spatial	Temporal	mAP@0.1	mAP@0.5	(5,1)	(5,5)
<i>GRU-based methods</i>										
MoCo-GRU (He et al., 2020)	62.5	82.0	52.2	55.6	72.6	74.8	68.2	63.4	58.5	78.7
ISC (Thoker et al., 2021)	62.5	82.6	50.6	52.3	70.6	76.8	64.6	58.7	62.3	81.3
CRRL (Qi et al., 2023)	60.7	75.2	-	-	61.4	66.2	57.6	52.1	58.3	77.1
CMD (Mao et al., 2022)	70.6	85.4	58.3	60.9	72.7	79.5	73.7	68.4	64.9	82.8
HaLP (Shah et al., 2023)	65.8	83.6	55.8	59.0	72.0	79.8	71.1	64.9	64.2	82.9
HiCo (Dong et al., 2023)	67.9	84.4	55.9	58.7	72.1	76.7	51.8	45.8	64.6	81.7
PCM ³ (Zhang et al., 2023)	73.7	88.8	63.1	66.8	87.0	86.1	73.3	68.2	65.2	82.9
PCM ³ ++	75.4	89.4	64.5	67.1	88.0	85.8	75.5	69.8	66.7	84.8
<i>GCN-based methods</i>										
AimCLR (Guo et al., 2022)	58.8	67.9	43.7	42.4	48.5	54.7	43.9	35.1	66.7	83.8
HiCLR (Zhang et al., 2023)	60.6	73.1	46.0	46.0	46.8	53.3	36.7	29.2	65.0	82.2
SkeAttnCLR (Hua et al., 2023)	55.3	60.2	-	-	47.4	52.9	48.5	41.7	62.1	76.9
ActCLR (Lin et al., 2023)	69.6	76.1	51.4	54.9	48.3	53.3	41.9	35.1	68.0	84.3
<i>Transformer-based methods</i>										
MAMP† (Mao et al., 2023)	62.0	70.0	-	-	56.5	81.0	78.8	72.4	56.2	76.3
UmURL (Sun et al., 2023)	72.0	88.9	59.5	62.2	71.1	77.4	43.4*	38.3*	64.9	77.2

for both linear evaluation and fully fine-tuning protocol. The subsequent works (Lin et al., 2024; Wu et al., 2024) also demonstrate promising results for Transformer-based MSM. However, these methods are only reported based on the joint view, which makes cross-modal knowledge ensemble in this paradigm still an open problem.

Contrastive learning has always been a popular topic triggered by the success in the image field, *e.g.*, SimCLR and MoCo. The model can learn a separable high-level representation space, making it dominant in learning linear representations. These works start from the exploiting the baseline algorithms (Rao et al., 2021; Thoker et al., 2021), and significantly boost the performance by applying stronger augmentations (Guo et al., 2022; Zhang et al., 2023; Lin et al., 2023), effective training strategies (Zhang et al., 2022; Chen et al., 2023), cross-modal knowledge (Li et al., 2021; Mao et al., 2022), achieving rapid improvement in recent three years.

Besides, combining different pre-training paradigms (Zhang et al., 2023; Lin et al., 2024) for skeleton-based representation learning has demonstrated promising results recently. This indicates that the representations learned by different pre-training paradigms can be complementary and beneficial.

2) Across Model Backbones. Different model backbones often employ different pre-training tasks. For example, Transformer often consumes extensive computational resources,

making it difficult for contrastive learning, which often requires encoding multiple data views. In contrast, it naturally fits the MSM schema which reduces computational overhead by masking tokens with large ratios. On the other hand, GRU and GCN models are relatively more efficient in training, often adopting the contrastive learning pretext task. It is also noted that the GRU models are often with larger feature dimensions than GCN models due to the smaller induced computational graph and less GPU memory occupancy.

Generally, different from supervised training, GRU models have achieved state-of-the-art performance under linear evaluation as shown in Fig. 7 in SSL, and the Transformer models are also promising recently. For fully fine-tuning protocol, GRU model is less effective due to the over-fitting caused by high dimension and huge parameters. Instead, Transformer models are more popular due to the powerful attention mechanism. Remarkably, with self-supervised pre-training, a vanilla Transformer (Mao et al., 2023) has surpassed the GCN/Transformer with complex designs in supervised training (Lee et al., 2023; Shi et al., 2020), demonstrating the significant role for alleviating the over-fitting problem and generalization capacity improvement of SSL.

4.4.2 Comparison on Other Downstream Tasks

1) Experiment Setting. We introduce the implementations of each task first and more details can be found in Supp.

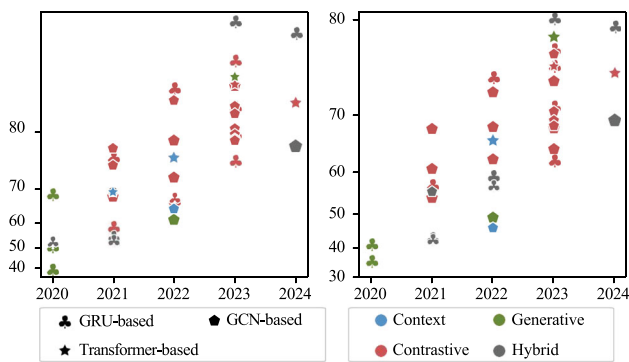


Fig. 7 Action recognition performance of model over time for different SSL methodologies and different backbones. Left: NTU 60, Right: NTU 120, using cross-subject linear evaluation protocol.

Skeleton-based Action Retrieval. Following previous work (Su et al., 2020), a K-nearest neighbors (KNN) classifier ($k=1$) is adopted to retrieve the nearest training sample for each testing data in the representation space. All models are first pre-trained and then evaluated on the target dataset. The precision is reported as accuracy.

Action Recognition with Occlusion. We evaluate the transfer ability of representations learned from the clean dataset to the action recognition with occluded data. We adopt the linear evaluation protocol. Following (Song et al., 2019), we construct a synthetic occluded dataset on both spatial and temporal dimension. For spatial dimension, different body parts are randomly masked. For temporal dimension, we set a random block of frames to zeros. The testing set is constructed by the same masks across different methods, with a masking ratio of [0.3, 0.7].

Skeleton-based Action Detection. We follow the previous works (Luo et al., 2018; Liu et al., 2020), and evaluate the short-term motion modeling capacity by action detection task. We train the attached linear classifier to predict frame-level categories to produce the final proposal. The encoder is pre-trained on NTU 60 dataset and then we fix it, transferring the learned representations to untrimmed PKUMMD Part I dataset. The mAP of different actions is adopted as metric with different tIoUs.

Unsupervised Few-shot Learning. We evaluate the performance of SSL as few-shot learners follow the previous work (Lu et al., 2022). Specifically, the model is first pre-trained on NTU 60 dataset. Then, a simple classifier, *e.g.*, Support Vector Machine (SVM), is fitted on the output features by the pre-trained encoder of the support data set. Finally, the adapted classifier along with the encoder is utilized to infer the query samples. We select 20 new categories

in NTU 120¹ which are not seen in pre-training, as the support/query sets.

2) Comparison and Analytics. The results are summarized in Table 3. We utilize the officially released model weights, keeping the experiments as fair and aligned as possible.

We first present the discussion on the perspective of model backbones. GRU models usually employ higher representation dimensions due to the smaller GPU memory consumption, which makes a more informative representation available, leading to stronger performance in retrieval task than GCNs and Transformers. GCNs model the skeleton more compactly with structured prior knowledge, resulting in a better capacity to handle over-fitting, accounting for the stronger few-shot performance for unseen categories. However, the other side of the coin is the sensitivity to occlusion, where GCN can suffer from serious representation degradation because of the distribution shift caused by occlusion. Besides, it is found the GRU-based often performs better than the GCN model in terms of action detection, which primarily focuses on the spatial structure while ignoring the fine-grained temporal dependency. In contrast, GRUs model the skeletons in a temporal-based auto-regressive manner and possess better scalability for the long-sequences and short-clips. Besides, the Transformer-based model (Mao et al., 2023) also demonstrates promising results when employing the masked skeleton modeling pretext task, and can be expected with a better fine-tuning performance due to the powerful attention mechanism.

On the taxonomy of pretext tasks, due to the lack of promising open-sourced context-based methods, we mainly focus on contrastive and generative learning methods. First, it is found that contrastive learning methods usually perform better than reconstruction-based methods, *e.g.*, MAMP (Mao et al., 2023) and CRRL (Qi et al., 2023), in action retrieval and few-shot learning. These tasks mainly rely on the direct measure in the representation space, *i.e.*, cosine similarity, indicating the dominant position of contrastive learning for achieving highly compact representations. Remarkably, the masked modeling method, *e.g.*, MAMP (Mao et al., 2023), achieves significant results in action detection task. We believe that masked modeling with Transformer architecture is very promising for handling the action detection task. By splitting the long sequences into short clips as patches for masked prediction, the learned representations can well capture fine-grained action cues and achieve notable results. However, it is not observed with as good performance under occluded recognition as expected, although it explicitly involves masked data modeling, which can be explored in future work.

¹ Specifically, the classes (index from 0) 60, 61, 66, 69, 72, 78, 79, 80, 84, 90, 91, 95, 96, 98, 99, 100, 102, 106, 108, 111, 113, 114, 115 are selected.

Table 4 Analytics on the pre-training of different datasets for multiple downstream tasks.

Pre-training	Recognition		Retrieval		Occluded		Detection		Few Shot	
	NTU 60	NTU 120	NTU 60	NTU 120	NTU 60 xview		PKUMMD		NTU 120 xset	
	xsub	xsub	xsub	xsub	Spatial	Temporal	mAP@0.1	mAP@0.5	(5,1)	(5,5)
NTU 60	84.8	73.5	75.4	58.8	88.0	85.8	75.5	69.8	66.7	84.8
NTU 120	85.1	76.7	75.3	64.5	87.7	88.2	75.3	69.4	69.6	87.2
PKUMMD	77.6	67.5	65.1	52.9	78.0	79.3	70.7	65.7	58.9	78.9
NTU 120+PKUMMD	85.8	77.0	76.2	64.5	88.4	88.0	76.2	70.5	69.4	87.0

4.4.3 Cross-Dataset Analytics for Different Downstream Tasks

We study the effect of the dependency relationship between different pre-training datasets on multiple downstream tasks with the proposed method. Specifically, we pre-train the model on NTU 60, NTU 120, PKUMMD, and their hybrid dataset, respectively, which is then evaluated on various downstream tasks under the transfer learning setting. To achieve a better understanding, we first introduce the relationship between different datasets in terms of the data composition. Concretely, NTU 120 extends NTU 60 on the action categories from 60 to 120 without introducing new samples of the original categories. PKUMMD is another skeleton dataset with similar collection settings, sharing a significant overlap of category distribution with NTU 60 and possessing fewer samples, which can be approximated as a smaller version of NTU 60. The quantitative results are presented in Table 4.

- First, as expected, reducing the scale of pre-training data generally results in performance degradation for downstream tasks, especially for the evaluation of the large-scale NTU 120 dataset. This is because the new classes in testing are unseen during pre-training, resulting in limited generalization capacity.
- Second, when the data distribution in testing is highly consistent with pre-training, introducing additional pre-training data of new categories (*e.g.*, from NTU 60 to NTU 120) is not as effective as increasing the sample number within the original categories (*e.g.*, from NTU 120 to NTU 120+PKUMMD, from PKUMMD to NTU 60). Moreover, it is even observed with slight performance degradation, *e.g.*, retrieval, spatial occluded recognition, and detection tasks (75.4→75.3, 88.0→87.7, 75.5→75.3). This can be interpreted as the new action category learning occupies part of the model capacity, impairing the representations of original categories.
- In contrast, for the tasks requiring to explicitly generalize to new data distributions, *e.g.*, temporal occluded recognition with distribution shift by impairing data structure (Note only spatial occlusion is seen in the pre-training) and few shot tasks with new categories, introducing new categories can benefit the performance significantly. This indicates more

Table 5 Ablation study on the clip-level contrastive learning and the post-distillation.

Method	Recognition		Detection
	xsub	xview	Part I
<i>w/o</i> clip contrastive learning	83.9	90.4	73.3
<i>w</i> clip contrastive learning	84.1	90.8	75.5
<i>w/o</i> post-distillation	84.1	90.8	75.5
<i>w</i> post-distillation	84.8	91.0	75.5

The best scores are in bold

essential knowledge of human motion is enabled for model to boost generalization.

To sum up, we analyze the effect of pre-training on different datasets to downstream tasks from two factors, category distribution and the sample number within each class, providing valuable insights for the skeleton SSL dataset choice and construction. Overall, a better performance can be expected with a larger pre-training dataset. However, it is also necessary to consider the actual downstream scenarios and the adopted pre-training methods to adjust data component.

4.5 Analysis of PCM³⁺⁺

1) Evaluating the PCM³⁺⁺. As shown in the previous results in Table 2 and Table 3, the proposed method achieves remarkable results on different downstream tasks compared with the state-of-the-art methods, and demonstrates significant improvement over our baseline method (Zhang et al., 2023). As our new improvement, we integrate an effective clip-level contrastive learning scheme and present a novel post-distillation training strategy. These methodological advancement is studied as follows.

2) Analysis of clip contrastive learning. We add the temporal clips as the asymmetric positive sample of the anchor sequence, which further improves the representation quality as shown in Table 5. Remarkably, the clip-level contrastive learning significantly boost the action detection performance by promoting short-term modeling capacity. Our designs boost versatile representation learning of different granularity and the effect on more downstream tasks can be found in Supp.

Table 6 Ablation study on the novel positive pair and semantic guidance.

Novel Positive Pair		Semantic Guidance	xview
<i>S_{mask}</i>	<i>S_{predict}</i>		
			88.2
✓			89.3
✓	✓		90.0
✓	✓	✓	90.5

The best scores are in bold

3) Effect of the post-distillation design. As shown in Table 5, the post-distillation training strategy can bring further improvement slightly for the recognition task. It alleviates the false negative problem by removing the one-hot pseudo label in the contrastive learning, achieving a more compact representation space.

4) Novel positive pair and semantic guidance. We analyze the effect of synergetic design between masked prediction and contrastive learning in Table 6. As we can see, the masked views and predicted views as positives bring notable improvement, indicating the potential benefits of masked prediction training to contrastive learning. On the other hand, we propagate the gradients of contrastive branch to masked prediction as high-level semantic guidance rather than stop the gradient. This boosts masked prediction training and yields more informative positive samples, leading to further performance improvement.

5) Complexity analysis. We give an analysis of space and computational complexities of our method for pre-training in Table 7. Compared with other GRU-based and Transformer-based methods, our method achieves a significant performance improvement with an acceptable cost of the complexity and training time. For the space complexity, the main extra cost is the reconstruction decoder (4M of space). For the computational complexity, the encoding process of different positive samples contributes most of the overhead.

5 Conclusions and Future Directions

This paper presents a comprehensive survey on skeleton-based action representation SSL, where different literature

is organized following the taxonomy of context-based, generative learning, and contrastive learning approaches. After reviewing existing works, we technically propose a novel and effective framework, for versatile skeleton-based action representation learning which is less explored before as a challenging topic. Finally, a detailed benchmark and insightful discussions are provided, including datasets, model backbones and pretext tasks. Meanwhile, we demonstrate the superior generalization performance of our method under different downstream tasks.

For the future research, the following pending issues deserve more attention:

- **Long-Term Motion Understanding.** Existing methods utilize the trimmed video clip, only containing one motion, as training data. This limits the long-term temporal reasoning capacity containing multiple actions, *e.g.*, *Long-Term Action Anticipation* task.
- **Multi-Modal Learning.** The exploration of multi-modal pre-training including human skeleton data is still insufficient, *e.g.*, texts, and RGB images. For example, as a useful complement to skeleton, RGB data can provide additional background knowledge that can boost action representation learning. Some recent works (Petrovich et al., 2023; Yu et al., 2024) explore the human motion-text representation learning. However, the insufficiency of high-quality motion-text datasets still limits the model generalization capacity.
- **Towards Skeleton in the Wild.** Existing methods are evaluated in simplified and controlled environments and can suffer from serious noise, *e.g.*, occlusion and view variation when deployed on the outdoor scenario. On this point, SSL is promising to boost skeleton representation robustness in the wild.
- **Versatile Representation Learning.** As discussed in previous sections, most works only focus on the recognition task, leaving the generation capacity across different tasks of SSL models under-explored. Meanwhile, new frameworks can also be explored, *e.g.*, Diffusion model (Ho et al., 2020), which is promising to handle both generative and discriminative downstream tasks.

Table 7 FLOPs, params, and training time (all tested on a single NVIDIA A40 GPU) results.

Model	Params	FLOPs	Training Time	Acc.
GL-Trans. (Kim, 2022)	214M	59.4G	53.62h	83.8%
ISC (Thoker et al., 2021)	106M	13.7G	45.49h	85.2%
CMD (Mao et al., 2022)	99M	17.3G	44.13h	86.9%
PCM ³ (Zhang et al., 2023)	103M	15.0G	45.31h	90.4%
PCM ³ ++	103M	17.9G	47.95h	91.0%

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11263-025-02644-8>.

Data Availability The adopted NTU-RGB+D (Shahroudy et al., 2016; Liu et al., 2019) and PKU-MMD (Liu et al., 2020) dataset are popular public benchmarks in skeleton-based action understanding. The code for data processing can be found in <https://github.com/JHAng2020/PCM3>.

Declarations

Competing interests This work was supported in part by the Beijing Major Science and Technology Project under Contract No. Z251100008425023, in part by the National Natural Science Foundation of China under Grant 62471009, and in part by the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

References

- Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabat, M., LeCun, Y., & Ballas, N. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE CVPR*, 15619–15629.
- Bai, Y., Yang, Y., Zhang, W., & Mei, T. (2022). Directional self-supervised learning for heavy image augmentations. In *IEEE CVPR*.
- Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabat, M., LeCun, Y., Assran, M., & Ballas, N. (2024). Revisiting feature prediction for learning visual representations from video. [arXiv:2404.08471](https://arxiv.org/abs/2404.08471).
- Bardes, A., Ponce, J., & LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. [arXiv:2105.04906](https://arxiv.org/abs/2105.04906)
- Bian, C., Feng, W., Meng, F., & Wang, S. (2022). View-invariant skeleton-based action recognition via global-local contrastive learning. [arXiv:2209.11634](https://arxiv.org/abs/2209.11634)
- Buckley, C., Alcock, L., McArdle, R., Rehman, R. Z. U., Del Din, S., Mazzà, C., Yarnall, A. J., & Rochester, L. (2019). The role of movement analysis in diagnosing and monitoring neurodegenerative conditions: Insights from gait and postural control. *Brain sciences*, 9(2), 34.
- Camara, F., Bellotto, N., Cosar, S., Nathanael, D., Althoff, M., Wu, J., Ruenz, J., Dietrich, A., & Fox, C. W. (2020). Pedestrian models for autonomous driving part i: low-level models, from sensing to tracking. *IEEE Transactions on Intelligent Transportation Systems*, 22(10), 6131–6151.
- Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE CVPR*.
- Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 132–149.
- Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, pages 9912–9924.
- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660.
- Chen, J., Jin, Z., Wang, Q., & Meng, H. (2023). Self-supervised 3D behavior representation learning based on homotopic hyperbolic embedding. *IEEE TIP*, 32, 6061–6074.
- Chen T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *ICML*.
- Chen, T.S., Hung, W.-C., Tseng, H.-Y., Chien, S.-Y., & Yang, M.-H. (2021). Incremental false negative detection for contrastive learning. [arXiv:2106.03719](https://arxiv.org/abs/2106.03719)
- Chen, X., & He, K. (2021). Exploring simple siamese representation learning. In *IEEE CVPR*.
- Chen, Y., Zhao, L., Yuan, J., Tian, Y., Xia, Z., Geng, S., Han, L., & Metaxas, D.N. (2022). Hierarchically self-supervised transformer for human skeleton representation learning. In *ECCV*.
- Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., & Hu, W. (2021). Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *ICCV*.
- Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., & Lu, H. (2020). Skeleton-based action recognition with shift graph convolutional network. In *IEEE CVPR*.
- Cheng, Y.-B., Chen, X., Chen, J., Wei, P., & Zhang, D. (2021). *and Liang Lin*. Hierarchical transformer: Unsupervised representation learning for skeleton-based human action recognition. In *IEEE ICME*.
- Darcet, T., Oquab, M., Mairal, J., & Bojanowski, P. (2023). Vision transformers need registers. [arXiv:2309.16588](https://arxiv.org/abs/2309.16588).
- Dolatabadi, E., Taati, B., & Mihailidis, A. (2016). Concurrent validity of the microsoft kinect for windows v2 for measuring spatiotemporal gait parameters. *Medical engineering & physics*, 38(9), 952–958.
- Dong, J., Sun, S., Liu, Z., Chen, S., Liu, B., & Wang, X. (2023). Hierarchical contrast for unsupervised skeleton-based action representation learning. In *AAAI*.
- Franco, L., Mandica, P., Munjal, B., & Galasso, F. (2023). Hyperbolic self-paced learning for self-supervised skeleton-based action representations. In *ICLR*.
- Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189.
- Gao, L., Ji, Y., Yang, Y., & Shen, H. (2022). Global-local cross-view fisher discrimination for view-invariant action recognition. In *ACM MM*.
- Gao, P., Lin, Z., Zhang, R., Fang, R., Li, H., Li, H., & Qiao, Yu. (2024). Mimic before reconstruct: Enhancing masked autoencoders with feature mimicking. *IJCV*, 132(5), 1546–1556.
- Gao, X., Yang, Y., Zhang, Y., Li, M., Yu, J.-G., & Du, S. (2021). Efficient spatio-temporal contrastive learning for skeleton-based 3d action recognition. *IEEE TMM*.
- Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. [arXiv:1803.07728](https://arxiv.org/abs/1803.07728)
- 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. *NeurIPS*, 33, 21271–21284.
- Gui, J., Chen, T., Cao, Q., Sun, Z., Luo, H., & Tao, D. (2023). A survey of self-supervised learning from multiple perspectives: Algorithms, theory, applications and future trends. [arXiv:2301.05712](https://arxiv.org/abs/2301.05712)
- Guo, T., Liu, H., Chen, Z., Liu, M., Wang, T., & Ding, R. (2022). Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *AAAI*.
- Haghighat, M., Moghadam, P., Mohamed, S., & Koniusz, P. Pre-training with random orthogonal projection image modeling. [arXiv:2310.18737](https://arxiv.org/abs/2310.18737)
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *IEEE CVPR*.
- He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *IEEE CVPR*.
- Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *NeurIPS*.

- Hua, Y., Wu, W., Zheng, C., Lu, A., Liu, M., Chen, C., & Wu, S. (2023). Part aware contrastive learning for self-supervised action recognition. [arXiv:2305.00666](#)
- Huynh, T., Kornblith, S., Walter, M.R., Maire, M., & Khademi, M. (2022). Boosting contrastive self-supervised learning with false negative cancellation. In *IEEE WACV*, pages 2785–2795.
- Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2013). Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE TPAMI*, 36(7):1325–1339.
- Jing, L., Vincent, P., LeCun, Y., & Tian, Y. (2022). Understanding dimensional collapse in contrastive self-supervised learning. In *ICLR*.
- Kim, B. (2022). *Hyung Jin Chang, Jungho Kim, and Jin Young Choi*. ECCV: Global-local motion transformer for unsupervised skeleton-based action learning.
- Lee, H.-Y., Huang, J.-B., Singh, M., & Yang, M.-H. (2017). Unsupervised representation learning by sorting sequences. In *ICCV*.
- Lee, J., Lee, M., Lee, D., & Lee, S. (2023). Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *ICCV*.
- Lee, J., & Ahn, B. (2020). Real-time human action recognition with a low-cost rgb camera and mobile robot platform. *Sensors*, 20(10), 2886.
- Li, J., Wong, Y., Zhao, Q., & Kankanhalli, M.S. (2018). Unsupervised learning of view-invariant action representations. *NeurIPS*.
- Li, J., Zhou, P., Xiong, C., & Hoi, S.C.H. (2020). Prototypical contrastive learning of unsupervised representations. [arXiv:2005.04966](#), 2020.
- Li, L., Wang, M., Ni, B., Wang, H., Yang, J., & Zhang, W. (2021). 3D human action representation learning via cross-view consistency pursuit. In *IEEE CVPR*.
- Li, W., Xie, J., & Loy, C.C. (2023). Correlational image modeling for self-supervised visual pre-training. In *IEEE CVPR*, pages 15105–15115.
- Lin, L., Song, S., Yang, W., & Liu, J. (2020). Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *ACM MM*.
- Lin, L., Wu, L., Zhang, J., & Liu, J. (2024). Idempotent unsupervised representation learning for skeleton-based action recognition. In *ECCV*, pages 75–92.
- Lin, L., Zhang, J., & Liu, J. (2023). Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In *IEEE CVPR*.
- Lin, L., Zhang, J., & Liu, J. (2024). Mutual information driven equivariant contrastive learning for 3d action representation learning. *IEEE TIP*.
- Liu, J., Song, S., Liu, C., Li, Y., & Hu, Y. (2020). A benchmark dataset and comparison study for multi-modal human action analytics. *ACM TOMM*, 16(2):41:1–41:24.
- Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., & Kot, A. C. (2019). NTU-RGB D 120: A large-scale benchmark for 3D human activity understanding. *IEEE TPAMI*, 42(10), 2684–2701.
- Liu, W., Qian Bao, Yu, S., & Mei, T. (2022). Recent advances of monocular 2d and 3D human pose estimation: a deep learning perspective. *ACM Computing Surveys*, 55(4), 1–41.
- Lopez-Nava, I. H., & Muñoz-Meléndez, A. (2019). Human action recognition based on low-and high-level data from wearable inertial sensors. *International Journal of Distributed Sensor Networks*, 15(12), 1550147719894532.
- Lu, Y., Wen, L., Liu, J., Liu, Y., & Tian, X. (2022). Self-supervision can be a good few-shot learner. In *ECCV*.
- Luo, Z., Hsieh, J.-T., & Jiang, L. (2018). *Juan Carlos Niebles, and Li Fei-Fei*. In ECCV: Graph distillation for action detection with privileged modalities.
- Lyu, K., Chen, H., Liu, Z., Zhang, B., & Wang, R. (2022). 3D human motion prediction: A survey. *Neurocomputing*, 489, 345–365.
- Mao, Y., Deng, J., Zhou, W., Fang, Y., Ouyang, W., & Li, H. (2023). Masked motion predictors are strong 3d action representation learners. In *ICCV*.
- Mao, Y., Zhou, W., Lu, Z., Deng, J., & Li, H. (2022). CMD: Self-supervised 3D action representation learning with cross-modal mutual distillation. In *ECCV*.
- Men, Q., Ho, E. S. L., Shum, H. P. H., & Leung, H. (2023). Focalized contrastive view-invariant learning for self-supervised skeleton-based action recognition. *Neurocomputing*, 537, 198–209.
- Nie, Q., Liu, Z., & Liu, Y. (2020). Unsupervised 3D human pose representation with viewpoint and pose disentanglement. In *ECCV*.
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khali-dov, V., Fernandez, P., Haziza, D., Massa, F., & El-Nouby, A., et al. (2023). DINOv2: Learning robust visual features without supervision. [arXiv:2304.07193](#).
- Paoletti, G., Cavazza, J., Beyan, C., & Del Bue, A. Unsupervised human action recognition with skeletal graph laplacian and self-supervised viewpoints invariance. [arXiv:2204.10312](#).
- Petrovich, M., Black, M.J., & Varol, G. (2023). Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *ICCV*, pages 9488–9497.
- Qi, Z., Dong, R., Fan, G., Ge, Z., Zhang, X., Ma, K., & Yi, L. (2023). Contrast with reconstruct: Contrastive 3D representation learning guided by generative pretraining. [arXiv:2302.02318](#), 2023.
- Rao, H., Leung, C., & Miao, C. (2024). Hierarchical skeleton meta-prototype contrastive learning with hard skeleton mining for unsupervised person re-identification. *IJCV*, 132(1), 238–260.
- Rao, H., Shihao, X., Xiping, H., Cheng, J., & Bin, H. (2021). Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. *Information Sciences*, 569, 90–109.
- Robinson, J., Chuang, C.-Y., Sra, S., & Jegelka, S. (2020). Contrastive learning with hard negative samples. [arXiv:2010.04592](#)
- Shah, A., Roy, A., Shah, K., Mishra, S.K., Jacobs, D., Cherian, A., & Chellappa, R. (2023). Halp: Hallucinating latent positives for skeleton-based self-supervised learning of actions. [arXiv:2304.00387](#)
- Shahroudy, A., Liu, J., Ng, T.-T., & Wang, G. (2016). NTU RGB-D: A large scale dataset for 3D human activity analysis. In *IEEE CVPR*.
- Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *IEEE CVPR*.
- Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2020). Decoupled spatial-temporal attention network for skeleton-based action recognition.
- Shu, X., Xu, B., Zhang, L., & Tang, J. (2022). Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition. *IEEE TPAMI*.
- Shurrah, S., & Duwairi, R. (2022). Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, 8, Article e1045.
- Song, Y.-F., Zhang, Z., & Wang, L. (2019). Richly activated graph convolutional network for action recognition with incomplete skeletons. In *IEEE ICIP*.
- Su, K., Liu, X., & Shlizerman, E. (2020) Predict & cluster: Unsupervised skeleton based action recognition. In *IEEE CVPR*.
- Su, Y., Lin, G., & Wu, Q. (2021). Self-supervised 3D skeleton action representation learning with motion consistency and continuity. In *ICCV*.
- Sun, S., Liu, D., Dong, J., Qu, X., Gao, J., Yang, X., Wang, X., & Wang, M. (2023). Unified multi-modal unsupervised representation learning for skeleton-based action understanding. In *ACM MM*.
- Tanfous, A.B., Zerroug, A., Linsley, D., & Serre, T. (2022). How and what to learn: Taxonomizing self-supervised learning for 3D action recognition. In *IEEE WACV*.

- Thoker, F.M., Doughty, H., & Snoek, C.G.M. (2021). Skeleton-contrastive 3d action representation learning. In *ACM MM*.
- Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., & Isola, P. (2020). What makes for good views for contrastive learning? *NeurIPS*, 33, 6827–6839.
- Tong, Z., Song, Y., & Wang, J. (2022). *and Limin Wang*. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*.
- Von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., & Pons-Moll, G. (2018). Recovering accurate 3D human pose in the wild using imus and a moving camera. In *ECCV*.
- Wang, M., Li, X., Chen, S., Zhang, X., Ma, L., & Zhang, Y. (2023). Learning representations by contrastive spatio-temporal clustering for skeleton-based action recognition. *IEEE TMM*.
- Wang, P., Wen, J., Si, C., Qian, Y., & Wang, L. (2022). Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition. *IEEE TIP*, 31, 6224–6238.
- Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M. (2019). Dynamic graph cnn for learning on point clouds. *IEEE TOG*, 38(5), 1–12.
- Weng, W., Wang, H., Wang, J., & He, L. (2025). *and Guosen Xie*. Usdrl: Unified skeleton-based dense representation learning with multi-grained feature decorrelation. In *AAAI*.
- Wu, L., Lin, L., Zhang, J., Ma, Y., & Liu, J. (2024). Macdiff: Unified skeleton modeling with masked conditional diffusion. In *ECCV*, pages 110–128.
- Wu, W., Hua, Y., Wu, S., Chen, C., & Lu, A., et al. (2022). Skeleton-mae: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition. [arXiv:2209.02399](https://arxiv.org/abs/2209.02399)
- Xie, J., Li, W., Zhan, X., Liu, Z., Ong, Y.S., & Loy, C.C. (2022). Masked frequency modeling for self-supervised visual pre-training. [arXiv:2206.07706](https://arxiv.org/abs/2206.07706)
- Xie, J., Zhan, X., Liu, Z., Ong, Y.-S., & Loy, C. C. (2022). Delving into inter-image invariance for unsupervised visual representations. *IJCV*, 130(12), 2994–3013.
- Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., & Dai, Q. (2022b). *and Han Hu*. Simmim: A simple framework for masked image modeling. In *IEEE CVPR*.
- Xu, B., Shu, X., & Song, Y. (2022). X-invariant contrastive augmentation and representation learning for semi-supervised skeleton-based action recognition. *IEEE TIP*, 31, 3852–3867.
- Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y. (2019). Self-supervised spatiotemporal learning via video clip order prediction. In *IEEE CVPR*.
- Xu, S., Rao, H., Hu, X., Cheng, J., & Hu, B. (2021a). *and Bin Hu*. Prototypical contrast and reverse prediction: Unsupervised skeleton based action recognition. *IEEE TMM*.
- Xu, Z., Shen, X., Wong, Y., & Kankanhalli, M.S. (2021b). Unsupervised motion representation learning with capsule autoencoders. *NeurIPS*.
- Yan, H., Liu, Y., Wei, Y., Li, Z., Li, G., & Lin, L. (2023). SkeletonMAE: graph-based masked autoencoder for skeleton sequence pre-training. In *ICCV*.
- Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, pages 7444–7452, 2018.
- Yang, D., Wang, Y., Dantcheva, A., Garattoni, L., Francesca, G., & Brémond, F. (2024). View-invariant skeleton action representation learning via motion retargeting. *IJCV*, pages 1–16.
- Yang, S., Liu, J., Lu, S., Er, M.H., & Kot, A.C. (2021). Skeleton cloud colorization for unsupervised 3D action representation learning. In *ICCV*.
- Yang, S., Liu, J., Lu, S., Hwa, E.M., Hu, Y., & Kot, A.C. (20223) Self-supervised 3d action representation learning with skeleton cloud colorization. [arXiv:2304.08799](https://arxiv.org/abs/2304.08799)
- Yang, Y., Liu, G., & Gao, X. (2022). Motion guided attention learning for self-supervised 3D human action recognition. *IEEE TCSVT*, 32(12), 8623–8634.
- Yu, Q., Tanaka, M., & Fujiwara, K. (2024). Exploring vision transformers for 3d human motion-language models with motion patches. In *IEEE CVPR*, pages 937–946.
- Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR.
- Zhan, C., Hong, L., Tianyu, G., Zhengyan, C., Pinhao, S., & Hao, T. (2022). Contrastive learning from spatio-temporal mixed skeleton sequences for self-supervised skeleton-based action recognition. In *arXiv*, 2207, 03065.
- Zhan, X., Xie, J., Liu, Z., Ong, Y.-S., & Loy, C.C. (2020). Online deep clustering for unsupervised representation learning. In *IEEE CVPR*, pages 6688–6697.
- Zhang, H., & Hou, Y. (2022). *and Wenjing Zhang*. Skeletal twins: Unsupervised skeleton-based action representation learning.
- Zhang, H., Hou, Y., Zhang, W., & Li, W. (2022). *Contrastive positive mining for unsupervised 3D action representation learning*. In *ECCV: Springer*.
- Zhang, J., Lin, L., & Liu, J. (2023). Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations. In *AAAI*, 1, 2023.
- Zhang, J., Lin, L., & Liu, J. (2023). Prompted contrast with masked motion modeling: Towards versatile 3d action representation learning. In *ACM MM*.
- Zhang, Q., & Wang, Y. (2022). *and Yisen Wang*. How mask matters: Towards theoretical understandings of masked autoencoders. *NeurIPS*.
- Zheng, N., Wen, J., Liu, R., Long, L., Dai, J., & Gong, Z. (2018). Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *AAAI*.
- Zhou, H., Liu, Q., & Wang, Y. (2023a). Learning discriminative representations for skeleton based action recognition. In *IEEE CVPR*.
- Zhou, Y., Duan, H., Rao, A., Su, B., & Wang, J. (2023b). Self-supervised action representation learning from partial spatio-temporal skeleton sequences. [arXiv:2302.09018](https://arxiv.org/abs/2302.09018).
- Zhu, W., Ma, X., Liu, Z., Liu, L., & Wayne, W. (2023). *and Yizhou Wang*. Motionbert: A unified perspective on learning human motion representations. In *ICCV*.
- Zhu, Y., Han, H., Yu, Z., & Liu, G. (2023). Modeling the relative visual tempo for self-supervised skeleton-based action recognition. In *ICCV*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.