Heterogeneous Skeleton-Based Action Representation Learning

Hongsong Wang[1,2] Xiaoyan Ma[3] Jidong Kuang[3] Jie Gui[3,4,5] *
[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[3]School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China
[4]Engineering Research Center of Blockchain Application, Supervision And Management (Southeast University), Ministry of Education, China
[5]Purple Mountain Laboratories, Nanjing 210000, China
{hongsongwang, 220224977, jidongkuang, guijie}@seu.edu.cn

Abstract

Skeleton-based human action recognition has received widespread attention in recent years due to its diverse range of application scenarios. Due to the different sources of human skeletons, skeleton data naturally exhibit heterogeneity. The previous works, however, overlook the heterogeneity of human skeletons and solely construct models tailored for homogeneous skeletons. This work addresses the challenge of heterogeneous skeleton-based action representation learning, specifically focusing on processing skeleton data that varies in joint dimensions and topological structures. The proposed framework comprises two primary components: heterogeneous skeleton processing and unified representation learning. The former first converts two-dimensional skeleton data into three-dimensional skeleton via an auxiliary network, and then constructs a prompted unified skeleton using skeleton-specific prompts. We also design an additional modality named semantic motion encoding to harness the semantic information within skeletons. The latter module learns a unified action representation using a shared backbone network that processes different heterogeneous skeletons. Extensive experiments on the NTU-60, NTU-120, and PKU-MMD II datasets demonstrate the effectiveness of our method in various tasks of action understanding. Our approach can be applied to action recognition in robots with different humanoid structures.

1. Introduction

With the rapid advancement of sensors and detection algorithms, acquiring accurate human poses has become easier.

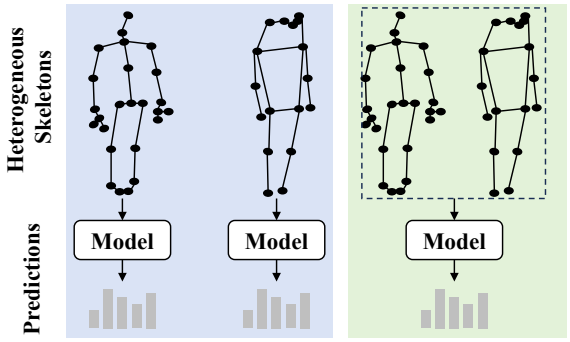


Figure 1. Comparison between our approach (right) and existing works (left). Previous works train specific models for each type of heterogeneous skeletons, whereas we design a unified model for heterogeneous skeletons.

Skeleton-based action recognition has emerged as a prevalent task in the field of computer vision. It holds a significant role in areas such as human-computer interaction, medical rehabilitation, video surveillance and intelligent sports.

Skeletons are a commonly used modality in action recognition tasks. Compared to other modalities, such as videos and depth map sequences, skeletons possess characteristics of high abstraction, low complexity, and good robustness. Furthermore, skeletons naturally align with human actions in a physical sense, allowing for a better representation of human motion. Skeletons can be collected by numerous sensors, thus resulting in heterogeneity in skeleton data. The heterogeneity of skeleton data is mainly manifested in the differences in coordinate dimensions and the number of joints defining the human body structure. For example, the skeleton data recorded by the Kinect V2 depth sensor is a three-dimensional sequence featuring 25 joints, whereas the skeleton data estimated from the RGB video is typically a

*Corresponding Author

two-dimensional sequence with 17 joints. Addressing data heterogeneity in human action recognition will emerge as a significant research topic.

Although deep learning methods have advanced rapidly in the field of skeleton-based human action recognition, with many excellent approaches emerging, these methods primarily focus on proposing different network architectures for single homogeneous data. For instance, these methods can be broadly classified into three categories: Recurrent Neural Networks (RNNs)-based [7, 34], Graph Convolutional Networks (GCNs)-based [40, 42, 50], and Transformers-based [48, 51]. Most of these methods rely on supervised training, requiring the model to be trained from scratch for each specific task. Consequently, these models lack sufficient transferability across different datasets. Recently, self-supervised skeleton-based action recognition [9, 31, 44] has attracted considerable attention owing to its enhanced transferability across various datasets. However, these methods overlook the heterogeneity of data, therefore, the model can only demonstrate transferability across datasets that contain homogeneous skeleton data.

This paper addresses the heterogeneity of human skeleton data, and our goal is to design a unified human action recognition model for heterogeneous data. The diagram of the proposed learning paradigm is illustrated in Figure 1. Compared with existing works that train individual models for different types of heterogeneous data, our method is capable of training a unified model encompassing these data. Our method primarily comprises heterogeneous skeleton processing and unified representation learning. The former converts heterogeneous data into a unified format, while the latter learns a unified action representation for different types of skeleton data.

More specifically, we focus on the two most commonly used types of data: the three-dimensional 25-joint skeleton and the two-dimensional 17-joint skeleton. To unify dimensions, we design a 3D pose estimation module to convert the two-dimensional skeleton into a three-dimensional skeleton, as the three-dimensional data contains richer information about human action. To unify the skeleton topology, we construct a prompted unified skeleton by selecting common joints from different heterogeneous skeletons and designing trainable skeleton-specific prompts to complement the missing joints for each type of skeleton. Since skeletons of different structures all semantically represent the human body but lack semantic information in their coordinate joints, we design semantic motion encoding, which utilizes pretrained language models to encode the semantic information to aid in action representation learning. The unified representation learning comprises an efficient Transformer-based architecture designed to learn unified representations from diverse types of data.

In summary, our main contributions are as follows:

- To the best of our knowledge, this is the first work that studies data heterogeneity in human skeletons and proposes a unified framework for learning heterogeneous skeleton-based action representations.
- We propose a heterogeneous skeleton processing module to unify heterogeneous skeletons from both the perspectives of coordinate dimensions and skeleton topologies.

2. Related Work

Self-Supervised Skeleton-based Action Recognition: Self-supervised skeleton-based action recognition [20, 28, 41, 47] refers to the use of unlabeled skeleton to perform action recognition tasks and has become a research hotspot in recent years. Su et al. [28] train an encoder-decoder network in an unsupervised manner to relate skeleton sequences with actions. Nie et al. [20] propose a Siamese denoising autoencoder for 3D human pose representation learning. Xu et al. [41] introduce a Motion Capsule Autoencoder (MCAE) to address transformation invariance in unsupervised learning. In recent years, more and more studies adopt contrastive learning methods for unsupervised learning [4, 29, 36]. Contrastive methods apply various augmentations to the unlabeled skeleton sequences, generating views that form positive and negative sample pairs. These pairs are then used to train the model to pull positive pairs closer and push negative pairs apart. Su et al. [29] construct fragments with speed variations and motion interruptions to determine positive and negative samples for contrastive learning. Wang et al. [36] perform contrastive learning on representations learned from both skeleton coordinate sequences and velocity sequences. Additionally, some studies introduce other techniques to enhance model performance [3, 6, 14, 19, 24, 39, 52]. Chen et al. [3] employ a hierarchical pre-training approach to enhance action representation capabilities. Dong et al. [6] generate multiple features at different granularities to perform contrastive learning in a hierarchical manner. Lin et al. [14] transform the data into actionlet and non-actionlet regions to enhance contrastive learning ability. Wu et al. [39] extend contrastive loss to measure spatiotemporal representation differences and use a masking strategy to increase the diversity of training data. Weng et al. [38] present a unified dense representation learning framework based on feature decorrelation. Despite the significant progress made by the above methods in advancing self-supervised skeleton action recognition, most of the research focuses on uni-modal data, failing to effectively explore the complementarity among multimodal skeleton data.

Multi-modal Action Representation Learning: Multi-modal information of skeletons, such as joints and bones, has been verified to exhibit strong complementarity for human action understanding [35]. To improve the representation capability of actions, recent works utilize multi-modal

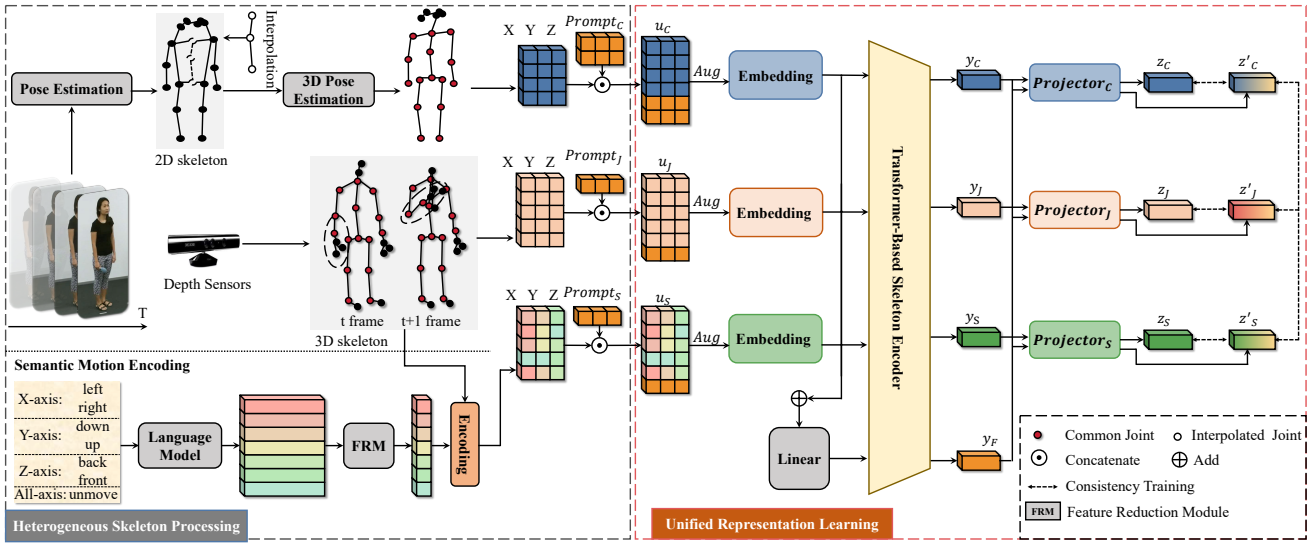


Figure 2. The structure of our framework. This framework comprises heterogeneous skeleton processing and unified representation learning. Initially, we convert the two-dimensional skeleton extracted by the pose estimation model into three dimensions using a 3D pose estimation module. Subsequently, we employ trainable skeleton-specific prompts to construct a prompted, unified skeleton representation. Additionally, we design semantic motion encoding to enrich the representation learning process with more semantic information. The unified representation learning trains a unified action representation network for these processed heterogeneous skeletons.

data to encourage models to learn cross-modal features. For example, Xiang et al. [40] propose a Generative Action Prompt (GAP) method, which leverages action semantic information to enhance the representation learning ability of the skeleton encoder. Chi et al. [4] incorporate the relative positions of joints to augment the multi-modal representation of the skeleton, providing complementary spatial information for the joints. Sun et al. [11] employ an early fusion strategy to input joint, motion, and bone modalities into the same stream, thereby reducing model complexity. In addition, a common approach for utilizing multi-modal data is to extend uni-modal methods to multi-modal ones through late fusion strategies [9, 45, 49]. This approach independently trains multiple single-modal models and then fuses their outputs to enhance representational capacity [12, 18, 44]. While these methods adopt multi-modal skeleton inputs, they fail to effectively handle and exploit heterogeneous data with complementary information, such as fine-grained semantics and skeletons with different topologies. Different from these approaches, we propose is a multi-modal representation learning framework that processes heterogeneous skeletons with different dimensions, topologies, and modalities through a unified network.

3. Method

3.1. Heterogeneous Skeleton Processing

Human skeleton data from different sources exhibit heterogeneity in skeleton structure. This heterogeneity is primarily manifested in two aspects: varying numbers of human body nodes and differing coordinate dimensions. For example, the human skeleton data collected by the Kinect V2 depth sensor consists of 3D data for 25 joints, whereas the human skeleton estimated from RGB video typically consists of 2D data for 17 joints. Each joint represents a distinct part of the human body. The comparison between the 25-joint skeleton and the 17-joint skeleton is in Figure 3. The 25-joint skeleton includes more joints in the hands, whereas the 17-joint skeleton has more joints focused on the face. We use these two skeletons as illustrative examples to process heterogeneous skeletons.

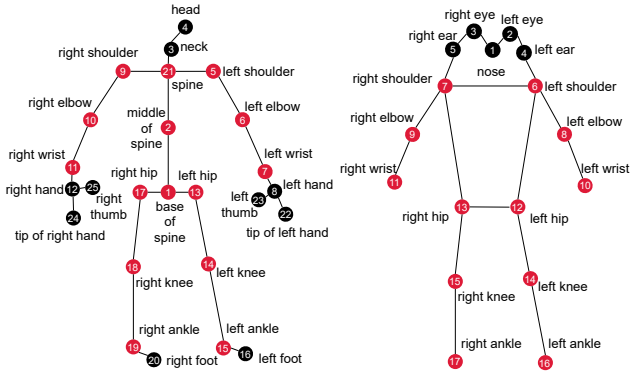
3D Pose Estimation: Three-dimensional skeletons contain more skeleton information compared to two-dimensional skeletons, which can better assist models in extracting features. Therefore, we aim to reconstruct the two-dimensional 17-joint skeleton into a three-dimensional one.

Since the spine lies on the central axis of the body, the missing three spine joints in the 17-joint skeleton can be interpolated using the joints of the left and right shoulders and hips. Thus, we manually add 3 spine-related joints to the 17-joint skeleton using linear interpolation to facilitate 3D pose estimation. The preprocessing step for interpolation is as follows:

$$p_{spine} = (p_{left\_shoulder} + p_{right\_shoulder})/2, \quad (1)$$

$$p_{base\_of\_spine} = (p_{left\_hip} + p_{right\_hip})/2, \quad (2)$$

$$p_{middle\_of\_spine} = (p_{spine} + p_{base\_of\_spine})/2, \quad (3)$$



(a) 25-joint skeleton (b) 17-joint skeleton

Figure 3. Comparison between two popular human skeletons: the 25-joint skeleton and the 17-joint skeleton. The name of each human joint is annotated near the index. Joints that are common to both skeletons are marked in red, while unique joints are marked in black. We consider three joints in the spine of the 25-joint skeleton as common joints because they can be easily interpolated by surrounding joints in the 17-joint skeleton.

where p represents the coordinates of a particular joint. After adding spine-related joints, the original 17-joint skeleton now has a total of 20 joints, but for the sake of continuity and simplicity, we still refer to it as the 17-joint skeleton.

After joint interpolation, a 3D pose estimation module is designed to predict the three-dimensional skeleton given the two-dimensional skeleton. By interpolating additional spine-related joints, more paired two-dimensional and three-dimensional joints are also made available during the training of the 3D pose estimation module.

Prompted Unified Skeleton: As shown in Figure 3, these two skeletons have different numbers of joints. So we has a total of 30 different joints. In order to process skeletons with differing numbers of joints, a unified skeleton format is necessary. Therefore, we use the method of prompt learning to define two trainable skeleton-specific prompts corresponding to the 25-joint skeleton and 17-joint skeleton: $prompt_J \in \mathbb{R}^{5\times3}$ and $prompt_C \in \mathbb{R}^{10\times3}$, which are used to construct prompted unified skeleton $u \in \mathbb{R}^{m\times t\times30\times3}$.

The prompt and the skeleton are concatenated in the dimension of the joint. Moreover, the original joint order differs between the 25-joint skeleton and the 17-joint skeleton. Therefore, we adjust them to a unified order, which is divided into three parts: human facial joints numbered from 1 to 5, common joint points numbered from 6 to 20, and head-hand-foot joint points numbered from 21 to 30.

3.2. Semantic Motion Encoding

The coordinates of the joint only represent physical motion of human action. To learn a unified model for heterogeneous skeletons, it is also crucial to harness the semantic information of the body within these skeletons. However, the semantics of body joints constitute static information regarding the names of the joints, which does not reflect the action being performed. To address this issue, we encode the semantics of joint motion as auxiliary information to assist in action representation learning.

Changes in the direction of motion during actions reflect the trend of motion and contain rich dynamic information. Therefore, they can serve as dynamic information to help process of Semantic Motion. The motion of joints can be divided into three directions: x-axis, y-axis, and z-axis. We use "right" and "left" to represent the direction on the x-axis; use "up" and "down" to represent the direction on the y-axis; use "front" and "back" represent the direction on the z-axis; finally, use "unmove" to represent no motion, with a total of seven directions of motion.

These seven direction word are fed into a pre-trained language model to get high-dimensional semantic features $e \in \mathbb{R}^{7\times l}$, where l represents the length of the feature vector, which depends on the specific language model used. The high-dimensional semantic features make it difficult for the model to learn effective action representation, and the high dimensionality can cause excessive computational pressure. A feature reduction module is designed to map these features to a low-dimensional encoding. We set the embedding to 1 to reduce dimensionality, so that we can construct semantic motion encoding with the same size as the three-dimensional skeleton.

The semantic motion encoding is obtained by:

$$m^x_{t,j} = \begin{cases} \tilde{e}_{left} & s^x_{t,j} - s^x_{t-1,j} < 0, \\ \tilde{e}_{ummove} & s^x_{t,j} - s^x_{t-1,j} = 0, \\ \tilde{e}_{right} & s^x_{t,j} - s^x_{t-1,j} > 0, \end{cases} \tag{4}$$

$$m^y_{t,j} = \begin{cases} \tilde{e}_{down} & s^y_{t,j} - s^y_{t-1,j} < 0, \\ \tilde{e}_{ummove} & s^y_{t,j} - s^y_{t-1,j} = 0, \\ \tilde{e}_{up} & s^y_{t,j} - s^y_{t-1,j} > 0, \end{cases} \tag{5}$$

$$m^z_{t,j} = \begin{cases} \tilde{e}_{back} & s^z_{t,j} - s^z_{t-1,j} < 0, \\ \tilde{e}_{ummove} & s^z_{t,j} - s^z_{t-1,j} = 0, \\ \tilde{e}_{front} & s^z_{t,j} - s^z_{t-1,j} > 0, \end{cases} \tag{6}$$

where $\tilde{e}$ is the semantic feature after dimensionality reduction, $s^x_{t,j}$ and $m^x_{t,j}$ represent the original value and the semantic motion encoding of the x-coordinate of the j-th body joint in the t-th frame, respectively.

Similar to heterogeneous skeleton processing, semantic motion encoding is also converted into a unified format utilizing prompt embeddings. This operation is as follows:

$$u_S = prompt_S \odot s_S \tag{7}$$

where $\odot$ denotes the concatenation operation, s_S and u_S represent the semantic motion encoding before and after this operation, respectively. As this prompted encoding has the same size as the real skeletons, it can perform the same data augmentation operations.

3.3. Unified Representation Learning

We design a unified network to learn action representations from processed heterogeneous skeletons. We adopt the self-supervised learning paradigm due to its capabilities of transferability and generalization. We use the Transformer-based architecture [21] as the feature backbone.

A specific embedding layer is first used to map each type of skeleton to the embeddings. Then, we follow the early fusion strategy [31] to train the network with inputs of heterogeneous skeletons. Let h_J, h_C, and h_S into h_F be the embeddings of the 25-joint skeleton, 17-joint skeleton and semantic motion. The fused embedding h_F is computed as:

$$h_F = linear(\frac{1}{3}(h_J + h_C + h_S)), \tag{8}$$

where $linear(\cdot)$ is learnable linear transformation.

After early fusion of embeddings, a unified feature encoder is used to produce skeleton feature $y \in \mathbb{R}^D$, where D denotes the feature dimension. This process is:

$$y = encoder(h), \tag{9}$$

where $h \in \{h_J, h_C, h_S, h_F\}$, and y is the corresponding feature of h.

Feature Consistency Loss: The unsupervised training losses of feature consistency include two levels: (1) between y_F and $\{y_J, y_C, y_S\}$; (2) among the elements of the set $\{y_J, y_C, y_S\}$. During training, given a batch of features, $Y \in \mathbb{R}^{N\times 4\times D}$, Y is mapped into skeleton-specific spaces by skeleton-specific projectors. Consequently, the consistency learning loss is defined as:

$$\mathcal{L}_{con} = \sum_i MSE(Z_i, Z_i') + \sum_{i\neq j} MSE(Z_i, Z_j), \tag{10}$$

where $i, j \in \{J, C, S\}$, Z and Z' are skeleton-specific representations obtained from specific skeleton feature and fused feature, respectively.

3D Pose Estimation Loss: In heterogeneous skeleton processing, we design a 3D pose estimation module to regress the three-dimensional skeleton from the two-dimensional one. This module is jointly trained with the unified action representation learning. The loss for 3D pose estimation is formulated as:

$$\mathcal{L}_{rec} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \left\| u_i^C - u_i^J \right\|_2^2, \tag{11}$$

where $\mathcal{B}$ is the set of common joints between the 25-joint skeleton and the 17-joint skeleton, u^C and u^J denote the corresponding prompted unified skeletons for the 17-joint and 25-joint skeletons, respectively.

Regularizations: We use VICREG [1] for action representation learning. VICREG includes semantic consistency regularization and variance-covariance (VC) regularization. The variance-covariance (VC) regularization include a variance term and a covariance term. The variance term is:

$$V(Z) = \frac{1}{D}\sum_{j=1}^{D} max(0, \gamma - \sqrt{Var(Z_{:,j}) + \epsilon}), \tag{12}$$

where γ is variance threshold, ϵ is a small scalar preventing numerical instabilities, and $Var(Z_{:,j})$ is the variance of jth embedding dimension vector $Z_{:,j}$. The covariance term is:

$$C(Z) = \frac{1}{D}\sum_{i\neq j}[Cov(Z)]_{i,j}^2, \tag{13}$$

where $Cov(Z)$ is the auto-covariance matrix of Z. The final VC regularization loss is:

$$\mathcal{L}_{reg} = \sum_i VC(Z_i) + VC(Z_i'), \tag{14}$$

where $i \in \{J, C, S\}$, and $VC(Z)$ is formulated as:

$$VC(Z) = \mu V(Z) + C(Z). \tag{15}$$

The total training objective loss is:

$$\mathcal{L} = \lambda\mathcal{L}_{con} + \mathcal{L}_{reg} + \mathcal{L}_{rec}. \tag{16}$$

4. Experiments

4.1. Experiments Settings

Dataset and Evaluation Metric: NTU-60 [16], NTU-120 [25] and PKU-MMD II [17] datasets are used to train and evaluate the model. Consistent with prior works, we use top-1 accuracy as the evaluation metric.

Implementation Details: We employ HRNet [30] for pose estimation to extract two-dimensional skeleton keypoints from the datasets. For semantic motion encoding, we utilize the pre-trained ViT-B/32 text encoder of the CLIP [22]. In the 3D pose estimation module, we apply a 4-layer MLP to convert 2D skeleton coordinates to 3D, standardizing the input using BatchNorm and processing negative values with the LeakyReLU activation function. The encoder backbone incorporates two single-head Transformers to model the spatial and temporal dimensions independently, each with a hidden dimension of 1024. The entire experiment is conducted within the PyTorch framework and accelerated using two NVIDIA GeForce RTX 4090 GPUs.

4.2. Comparison with The SOTA Methods

After representation learning, our method is already capable of extracting expressive features. We compare our method with state-of-the-art methods on two tasks: skeleton-based action recognition and skeleton-based action retrieval.

Method	Publication	Modality	FLOPs/G	NTU-60		NTU-120		PKU-MMD II
				x-sub	x-view	x-sub	x-set	x-sub
AimCLR [9]	AAAI'22	J	1.15	74.3	79.7	63.4	63.4	–
PTSL [49]	AAAI'23	J	**1.15**	77.3	81.8	66.2	67.7	49.3
GL-Transformer [11]	ECCV'22	J	118.62	76.3	83.8	66.0	68.7	–
CPM [44]	ECCV'22	J	2.22	78.7	84.9	68.7	69.6	48.3
CMD [18]	ECCV'22	J	5.76	79.8	86.9	70.3	71.5	43.0
IGM [15]	ECCV'24	J	–	86.2	91.2	80.0	81.4	–
HYSP [8]	ICLR'23	J	–	78.2	82.6	61.8	64.6	–
UmURL [31]	ACM MM'23	J	1.74	82.3	89.8	73.5	74.3	52.1
PCM^3 [46]	ACM MM'23	J	–	83.9	90.4	76.5	77.5	51.5
RVTCLR+ [52]	ICCV'23	J	–	74.7	79.1	—	–	–
HaLP [24]	CVPR'23	J	–	79.7	86.8	71.1	72.2	43.5
ActCLR [14]	CVPR'23	J	–	80.9	86.7	69.0	70.5	–
USDRL [38]	AAAI'25	J	–	84.2	90.8	76.0	76.9	51.8
3s-AimCLR [9]	AAAI'22	J+M+B	3.45	78.9	83.8	68.2	68.8	39.5
3s-CPM [44]	ECCV'22	J+M+B	6.66	83.2	87.0	73.0	74.0	51.5
3s-CMD [18]	ECCV'22	J+M+B	17.28	84.1	90.9	74.7	76.1	52.6
3s-HiCLR [45]	AAAI'23	J+M+B	7.08	78.8	83.1	67.3	69.9	-
3s-PSTL [49]	AAAI'23	J+M+B	3.45	79.1	83.8	69.2	70.3	52.3
3s-HYSP [8]	ICLR'23	J+M+B	–	79.1	85.2	64.5	67.3	–
UmURL [31]	ACM MM'23	J+M+B	**2.54**	84.2	90.9	75.2	76.3	54.0
3s-UmURL [31]	ACM MM'23	J+M+B	5.22	84.4	91.4	75.9	77.2	54.3
3s-RVTCLR+ [52]	ICCV'23	J+M+B	–	79.7	84.6	68.0	68.9	v
3s-ActCLR [14]	CVPR'23	J+M+B	–	84.3	88.8	74.3	75.7	–
USDRL [38]	AAAI'25	J+M+B	–	87.1	93.2	79.3	80.6	59.7
Ours	–	J	1.74	80.2	88.0	70.7	73.5	47.7
Ours	–	C	**1.74**	**84.4**	**90.6**	**73.5**	**78.4**	**54.1**
Ours	–	S	1.74	70.1	75.7	58.3	60.2	33.8
Ours	–	J+C	2.17	86.1	92.7	75.8	80.0	57.3
Ours	–	J+S	2.17	80.7	88.0	71.0	73.2	48.9
Ours	–	C+S	**2.17**	85.0	90.3	73.8	78.3	54.1
Ours	–	J+C+S	**2.54**	**87.8**	**93.7**	**78.9**	**82.2**	**58.2**

Table 1. Comparison with state-of-the-art methods on the skeleton-based action recognition task. Uni-modality and multi-modal methods are compared on the NTU-60, NTU-120, and PKU-MMD II datasets. For simplicity, J and C represent the joint modalities of a three-dimensional 25-joint skeleton and a two-dimensional 17-joint skeleton, respectively, while S signifies the semantic motion encoding.

Skeleton-based Action Recognition: We adopt the same practices as in previous works [12, 31]. Specifically, we freeze the model weights and use them as an encoder. After that only a linear classifier is trained to classify the features extracted by the encoder. We conduct experiments by utilizing 3D Pose Estimation and Semantic Motion Encoding. Moreover, we calculate the computational complexity of the model. Experimental results are presented in Table 1.

The results show that our method significantly outperforms other approaches in heterogeneous skeleton applications, validating the effectiveness of the our method. Compared to the state-of-the-art 3s-UmURL [31], our model performs better when using the same number of skeleton. This advantage is primarily attributed to the finer-grained information of heterogeneous skeleton and richer dynamic semantic information, which allow the model to capture expressive features. In terms of computational efficiency, our model demonstrates lower FLOPs than CrossSCLR [12] and CMD [18], owing to the Unified Representation Learning. Our method also demonstrates excellent flexibility when dealing with a single skeleton or any combination of two sets of skeleton.

Skeleton-based Action Retrieval: In the skeleton-based action retrieval experiment, cosine similarity is employed for action query retrieval. Similarly, the pre-trained model is not fine-tuned. As shown in Table 2, the experimental results demonstrate that our method achieves superior retrieval performance on the NTU-60 and NTU-120 datasets

Method	Modality	NTU-60		NTU-120	
		x-sub	x-view	x-sub	x-set
LongT GAN [47]	J	39.1	48.1	31.5	35.5
P&C [28]	J	50.7	76.3	39.5	41.8
AimCLR [9]	J	62.0	71.5	-	-
ISC [33]	J	62.5	82.6	50.6	52.3
HiCLR [45]	J	67.3	75.3	-	-
HiCo [6]	J	68.3	84.8	56.6	59.1
CMD [18]	J	70.6	85.4	58.3	60.9
HaLP [24]	J	65.8	83.6	55.8	59.0
UmURL [31]	J	71.3	88.3	58.5	60.9
UmURL [31]	J+M+B	72.0	88.9	59.5	62.2
Ours	J	66.3	87.1	55.7	59.8
Ours	C	70.3	85.2	58.5	64.8
Ours	S	53.5	74.8	43.4	46.2
Ours	J+C+S	**72.7**	**90.9**	**61.9**	**66.9**

Table 2. Comparison with state-of-the-art methods on skeleton-based action retrieval task on the NTU-60 and NTU-120 datasets.

when processing heterogeneous skeletons. This further confirms that the proposed heterogeneous skeleton-based action representation learning significantly improves the performance of downstream tasks, clearly indicating that the model effectively extracts more discriminative and expressive features.

Ablation Studies: The ablation experiments in Table 3 include specifically removing the modules of 3D pose estimation, semantic motion encoding, and skeleton-specific prompt embedding during training. To study the impact of the 3D pose estimation module, we conduct a comparative analysis using only the original two-dimensional skeleton data. The results indicate that the reconstructed three-dimensional skeleton achieves superior performance. Additionally, to examine the effectiveness of semantic motion, we replace the semantic features extracted from a language model with a simple numeric representation of joint motion, where positive, negative, and no motion are represented by 1, -1, and 0, respectively. The results confirm that that semantic motion, enriched with detailed action-related information, improves the model's recognition performance. Finally, to evaluate the role of skeleton-specific prompts, we replace traditional zero-padding with trainable prompts for comparison. The results show that trainable prompts enhance the representational capacity of the skeleton topology. Collectively, these enhancements lead to notable improvements in the model's action recognition accuracy.

Discussions: We further discuss the roles of semantic motion encoding and heterogeneous skeleton in action recognition. On the one hand, as can be seen from Table 1 and Figure 4, semantic motion encoding generally improves recognition accuracy across all experimental settings, indicating that data from semantic modalities can effectively assist the model in recognizing actions. On the other hand, as previously analyzed, the 25-joint skeleton includes fine-grained hand and foot joints, which is advantageous for recognizing actions heavily influenced by these joints, such as hand-waving and pointing, as shown in the Figure 4. In contrast, the 17-joint skeleton is better suited for recognizing actions dominated by facial joints, such as head-shaking and face-wiping. By integrating heterogeneous skeletons in a unified representation learning framework, we provide the model with a more comprehensive set of skeleton information, enabling it to capture richer and more detailed features for action recognition.

Method	PKU-MMD II
w/o 3D pose estimation	55.8
w/o semantic motion	57.9
w/o skeleton-specific prompt	57.2
Ours	58.2

Table 3. Ablation studies on 3D pose estimation module, semantic motion encoding and trainable skeleton-specific prompts.

Method	Modality	x-sub		x-view	
		1%	5%	1%	5%
ASSL [27]	J	-	57.3	-	63.6
ISC [33]	J	35.7	59.6	38.1	65.7
MCC [29]	J	-	47.4	-	53.3
Hi-TRS [3]	J	39.1	63.3	42.9	68.3
GL-Transformer [11]	J	-	64.5	-	68.5
Colorization [43]	J	48.3	65.7	52.5	70.3
CrosSCLR [12]	J	48.6	67.7	49.8	70.6
HiCo [6]	J	54.4	-	54.8	-
CPM [44]	J	56.7	-	57.5	-
CMD [18]	J	50.6	71.0	53.0	75.3
UmURL [31]	J+M+B	**59.6**	74.6	**60.3**	78.6
3s-AimCLR [9]	J+M+B	54.8	-	54.3	-
3s-CMD [18]	J+M+B	55.6	74.3	55.5	77.2
Ours	J+C+S	55.0	**76.3**	55.0	**79.1**

Table 4. Comparison with state-of-the-art methods on semi-supervised learning task on the NTU-60 dataset.

4.3. Potential for Semi-Supervised Learning

Following the experimental setup of the previous study [33], we conduct experiments to evaluate the potential of the proposed method in semi-supervised learning. We first pre-train the encoder using our method, then randomly select 1% and 5% of the labeled training data during the fine-tuning phase to jointly optimize the encoder and classifier.

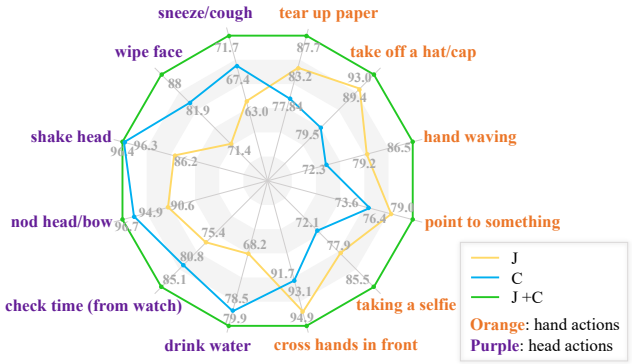


Figure 4. Comparison of skeletons in terms of single-class action accuracy. The experiment is conducted under the NTU-60/x-sub.

Method	Modality	Transfer to PKU-MMD II	
		NTU-60	NTU-120
LongT GAN [47]	J	44.8	-
M2L [13]	J	45.8	-
ISC [33]	J	45.9	-
CrosSCLR [12]	J	54.0	52.8
HiCo [6]	J	56.3	55.4
CMD [18]	J	56.0	57.0
UmURL [31]	J+M+B	59.7	58.5
Ours	J+C+S	**64.3**	**63.1**

Table 5. Comparison with state-of-the-art methods on the transfer learning for action recognition task.

Method	Modality	Accuracy
BEAR [5]	RGB	69.6
MoLo [37]	RGB+Point	73.3
SVT [23]	RGB	62.3
Euclidean [10]	RGB	68.2
Hyperbolic [32]	RGB	73.4
CARL [2]	RGB	41.8
Ours	Skeleton	**75.3**

Table 6. Comparison of action recognition methods transferred from NTU-120 pretraining to the FineGYM dataset.

To our surprise, the results vary greatly between the 1% and 5% settings. Our method outperforms all other methods in the 5% setting, but only achieves relatively good results in the 1% setting. After analysis, we believe that our method requires a certain amount of data for support. Too little data can lead to a decrease in model performance. Because the representation learning of heterogeneous skeletons is relatively complex and challenging for the model. However, as the amount of data increases, the improvement speed of our method is also very significant, allowing it to quickly surpass previous methods, which demonstrates the great potential of our method.

4.4. Transferability for Action Recognition

To validate that the representations learned through heterogeneous skeleton topologies offer improved generalization, we conduct transfer learning experiments. In these experiments, we first conduct pre-training on the model using unsupervised learning with the source dataset. Subsequently, we solely fine-tune the linear classification layer on the target dataset to assess its adaptability across various datasets.

Transferability on 3D Skeleton Dataset: As shown in Table 5, we use the NTU-60 and NTU-120 datasets as the source datasets and the PKU-MMD II dataset as the target dataset. Experiments are conducted under the x-sub setting. The results show that our method significantly outperforms other baseline approaches in transfer learning tasks. This advantage is mainly attributed to the superior representational capacity of the heterogeneous skeleton topology, which enables the learned features to transfer effectively and exhibit robust generalization capabilities.

Transferability on 2D Skeleton Dataset: We further validate the transferability of our method on heterogeneous skeletons by fine-tuning the classifier layer on the realistic skeleton dataset FineGYM [26], which provides two-dimensional skeletons with 17-joint skeleton. As can be seen from Table 6, our method achieves competitive results on FineGYM in transfer learning task. This indicates that our method exhibits outstanding transferability and flexibility when dealing with heterogeneous skeleton.

5. Conclusion

In this work, we study data heterogeneity in human skeletons and introduce a unified framework for learning heterogeneous skeleton-based action representations. This framework consists of two components: heterogeneous skeleton processing and unified representation learning. We also design semantic motion coding to leverage the semantic information within skeletons. We take the three-dimensional 25-joint skeleton and the two-dimensional 17-joint skeleton as examples for self-supervised representation learning. Extensive experiments demonstrate the proposed model's ability to recognize human actions from different heterogeneous data. These heterogeneous skeletons are highly complementary for action recognition, and significantly enhance the performance through efficient early fusion. The proposed model also exhibits excellent transferability across various scenarios of skeleton-based action understanding. However, the shortcoming is that the model is limited to processing skeleton data for up to two people.

Acknowledgments

This work was supported by National Science Foundation of China (62172090, 62302093), Jiangsu Province Natural Science Fund (BK20230833), Start-up Research Fund of Southeast University (RF1028623097), and Big Data Computing Center of Southeast University.

References

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021. 5

[2] Minghao Chen, Fangyun Wei, Chong Li, and Deng Cai. Frame-wise action representations for long videos via sequence contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13801–13810, 2022. 8

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

[4] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20186–20196, 2022. 2, 3

[5] Andong Deng, Taojiannan Yang, and Chen Chen. A large-scale study of spatiotemporal representation learning with a new benchmark on action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20519–20531, 2023. 8

[6] Jianfeng Dong, Shengkai Sun, Zhonglin Liu, Shujie Chen, Baolong Liu, and Xun Wang. Hierarchical contrast for unsupervised skeleton-based action representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 525–533, 2023. 2, 7, 8

[7] Yong Du, Yun Fu, and Liang Wang. Representation learning of temporal dynamics for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 25(7):3010–3022, 2016. 2

[8] Luca Franco, Paolo Mandica, Bharti Munjal, and Fabio Galasso. Hyperbolic self-paced learning for self-supervised skeleton-based action representations. In *International Conference on Learning Representations*, 2023. 6

[9] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 762–770, 2022. 2, 3, 6, 7

[10] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 8

[11] Boeun Kim, Hyung Jin Chang, Jungho Kim, and Jin Young Choi. Global-local motion transformer for unsupervised skeleton-based action learning. In *European Conference on Computer Vision*, pages 209–225. Springer, 2022. 3, 6, 7

[12] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 3d human action representation learning via cross-view consistency pursuit. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4741–4750, 2021. 3, 6, 7, 8

[13] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *Proceedings of the ACM international conference on multimedia*, pages 2490–2498, 2020. 8

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

[15] Lilang Lin, Lehong Wu, Jiahang Zhang, and Jiaying Liu. Idempotent unsupervised representation learning for skeleton-based action recognition. In *European Conference on Computer Vision*, pages 75–92. Springer, 2025. 6

[16] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2019. 5

[17] Jiaying Liu, Sijie Song, Chunhui Liu, Yanghao Li, and Yueyu Hu. A benchmark dataset and comparison study for multi-modal human action analytics. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(2):1–24, 2020. 5

[18] Yunyao Mao, Wengang Zhou, Zhenbo Lu, Jiajun Deng, and Houqiang Li. Cmd: Self-supervised 3d action representation learning with cross-modal mutual distillation. In *European Conference on Computer Vision*, pages 734–752. Springer, 2022. 3, 6, 7, 8

[19] Yunyao Mao, Jiajun Deng, Wengang Zhou, Yao Fang, Wanli Ouyang, and Houqiang Li. Masked motion predictors are strong 3d action representation learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10191, 2023. 2

[20] Qiang Nie, Ziwei Liu, and Yunhui Liu. Unsupervised 3d human pose representation with viewpoint and pose disentanglement. In *European Conference on Computer Vision*, pages 102–118. Springer, 2020. 2

[21] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208:103219, 2021. 5

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5

[23] Kanchana Ranasinghe, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Michael S Ryoo. Self-supervised video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2874–2884, 2022. 8

[24] Anshul Shah, Aniket Roy, Ketul Shah, Shlok Mishra, David Jacobs, Anoop Cherian, and Rama Chellappa. Halp: Hallucinating latent positives for skeleton-based self-supervised learning of actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18846–18856, 2023. 2, 6, 7

[25] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019, 2016. 5

[26] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2020. 8

[27] Chenyang Si, Xuecheng Nie, Wei Wang, Liang Wang, Tieniu Tan, and Jiashi Feng. Adversarial self-supervised learning for semi-supervised 3d action recognition. In *European Conference on Computer Visio*, pages 35–51. Springer, 2020. 7

[28] Kun Su, Xiulong Liu, and Eli Shlizerman. Predict & cluster: Unsupervised skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9631–9640, 2020. 2, 7

[29] Yukun Su, Guosheng Lin, and Qingyao Wu. Self-supervised 3d skeleton action representation learning with motion consistency and continuity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13328–13338, 2021. 2, 7

[30] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 5

[31] Shengkai Sun, Daizong Liu, Jianfeng Dong, Xiaoye Qu, Junyu Gao, Xun Yang, Xun Wang, and Meng Wang. Unified multi-modal unsupervised representation learning for skeleton-based action understanding. In *Proceedings of the ACM International Conference on Multimedia*, pages 2973–2984, 2023. 2, 5, 6, 7, 8

[32] Dídac Surís, Ruoshi Liu, and Carl Vondrick. Learning the predictability of the future. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12607–12617, 2021. 8

[33] Fida Mohammad Thoker, Hazel Doughty, and Cees GM Snoek. Skeleton-contrastive 3d action representation learning. In *Proceedings of the ACM international conference on multimedia*, pages 1655–1663, 2021. 7, 8

[34] Hongsong Wang and Liang Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 499–508, 2017. 2

[35] Hongsong Wang and Liang Wang. Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection. *IEEE Transactions on Image Processing*, 27(9):4382–4394, 2018. 2

[36] Peng Wang, Jun Wen, Chenyang Si, Yuntao Qian, and Liang Wang. Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition. *IEEE Transactions on Image Processing*, 31:6224–6238, 2022. 2

[37] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong Sang. Molo: Motion-augmented long-short contrastive learning for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18011–18021, 2023. 8

[38] Wanjiang Weng, Hongsong Wang, Junbo Wang, Lei He, and Guosen Xie. Usdrl: Unified skeleton-based dense representation learning with multi-grained feature decorrelation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 2, 6

[39] Cong Wu, Xiao-Jun Wu, Josef Kittler, Tianyang Xu, Sara Ahmed, Muhammad Awais, and Zhenhua Feng. Scd-net: Spatiotemporal clues disentanglement network for self-supervised skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5949–5957, 2024. 2

[40] Wangmeng Xiang, Chao Li, Yuxuan Zhou, Biao Wang, and Lei Zhang. Generative action description prompts for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10276–10285, 2023. 2, 3

[41] Ziwei Xu, Xudong Shen, Yongkang Wong, and Mohan S Kankanhalli. Unsupervised motion representation learning with capsule autoencoders. *Advances in Neural Information Processing Systems*, 34:3205–3217, 2021. 2

[42] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[43] Siyuan Yang, Jun Liu, Shijian Lu, Meng Hwa Er, and Alex C Kot. Skeleton cloud colorization for unsupervised 3d action representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13423–13433, 2021. 7

[44] Haoyuan Zhang, Yonghong Hou, Wenjing Zhang, and Wanqing Li. Contrastive positive mining for unsupervised 3d action representation learning. In *European Conference on Computer Vision*, pages 36–51. Springer, 2022. 2, 3, 6, 7

[45] Jiahang Zhang, Lilang Lin, and Jiaying Liu. Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3427–3435, 2023. 3, 6, 7

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

[47] Nenggan Zheng, Jun Wen, Risheng Liu, Liangqu Long, Jianhua Dai, and Zhefeng Gong. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2, 7, 8

[48] Yuxuan Zhou, Zhi-Qi Cheng, Chao Li, Yanwen Fang, Yifeng Geng, Xuansong Xie, and Margret Keuper. Hypergraph transformer for skeleton-based action recognition. *arXiv preprint arXiv:2211.09590*, 2022. 2

[49] Yujie Zhou, Haodong Duan, Anyi Rao, Bing Su, and Jiaqi Wang. Self-supervised action representation learning from partial spatio-temporal skeleton sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3825–3833, 2023. 3, 6

[50] Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xian-Sheng Hua. Blockgcn: Redefine topology awareness for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2049–2058, 2024. 2

[51] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15085–15099, 2023. 2

[52] Yisheng Zhu, Hu Han, Zhengtao Yu, and Guangcan Liu. Modeling the relative visual tempo for self-supervised skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13913–13922, 2023. 2, 6