SkeletonMAE: Graph-based Masked Autoencoder for Skeleton Sequence Pre-training

Hong Yan[1*] Yang Liu[1*†] Yushen Wei[1] Zhen Li[2] Guanbin Li[1] Liang Lin[1]
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]The Chinese University of Hong Kong, Shenzhen, China
{yanh36,weiysh8}@mail2.sysu.edu.cn, {liuy856,liguanbin}@mail.sysu.edu.cn,
lizhen@cuhk.edu.cn, linliang@ieee.org

Abstract

Skeleton sequence representation learning has shown great advantages for action recognition due to its promising ability to model human joints and topology. However, the current methods usually require sufficient labeled data for training computationally expensive models. Moreover, these methods ignore how to utilize the fine-grained dependencies among different skeleton joints to pre-train an efficient skeleton sequence learning model that can generalize well across different datasets. In this paper, we propose an efficient skeleton sequence learning framework, named Skeleton Sequence Learning (SSL). To comprehensively capture the human pose and obtain discriminative skeleton sequence representation, we build an asymmetric graph-based encoder-decoder pre-training architecture named SkeletonMAE, which embeds skeleton joint sequence into graph convolutional network and reconstructs the masked skeleton joints and edges based on the prior human topology knowledge. Then, the pre-trained SkeletonMAE encoder is integrated with the Spatial-Temporal Representation Learning (STRL) module to build the SSL framework. Extensive experimental results show that our SSL generalizes well across different datasets and outperforms the state-of-the-art self-supervised skeleton-based methods on FineGym, Diving48, NTU 60 and NTU 120 datasets. Moreover, we obtain comparable performance to some fully supervised methods. The code is avaliable at https://github.com/HongYan1123/SkeletonMAE.

1. Introduction

Human action recognition has attracted increasing attention in video understanding [83, 7, 45, 60, 39], due to its wide applications [2, 14, 61, 51, 52, 86, 44] in human-computer interaction, intelligent surveillance security, virtual reality, etc. In terms of visual perception [24], even without appearance information, humans can identify action categories by only observing the motion of joints. Unlike RGB videos [3, 14, 13], the skeleton sequences only contain the coordinate information of the key joints of the human body [84], which is high-level, light-weighted, and robust against complex backgrounds and various conditions including viewpoint, scale, and movement speed [11, 69]. Additionally, with the development of human pose estimation algorithms [8, 1, 85], the localization method of human joints (i.e., key points) has made great progress and it is feasible to obtain accurate skeleton sequences. At present, the existing 2D pose estimation method is more accurate and more robust than the 3D pose estimation methods [11]. In Figure 1 (a), we visualize 2D poses estimated with HR-Net [67] for two action classes on FineGym dataset [59]. It

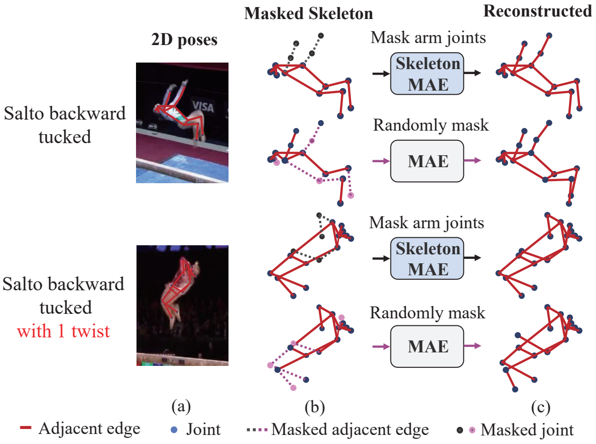


Figure 1: Traditional MAE usually uses random masking strategy to reconstruct skeleton, which tends to ignore action-sensitive skeleton regions. Differently, our proposed SkeletonMAE reconstructs the masked skeleton joints and edges based on the prior human topology knowledge, to obtain a comprehensive perception of the action.

[*]Both authors contributed equally to this work.
[†]Corresponding author.

can be seen that the 2D poses can accurately capture human skeletons and motion details.

Due to the promising ability to model multiple granularities and large variations in human motion, the skeleton sequence is more suitable to distinguish similar actions with subtle differences than the RGB data. To capture discriminative spatial-temporal motion patterns, most of the existing skeleton-based action recognition methods [11, 83, 5] are fully supervised and usually require large amounts of labeled data for training elaborate models, which is time-consuming and labor-intensive. To mitigate the problem of limited labeled training data, self-supervised skeleton-based action recognition methods [31, 18, 65] have attracted increasing attention recently. Some contrastive learning methods [31, 18] adopted data augmentation to generate pairs of positive and negative samples, but they rely heavily on the number of contrastive pairs. With the popularity of the encoder-decoder [64, 48], some methods [91, 65] reconstructed the masked skeleton sequence by link reconstruction to encourage the topological closeness following the paradigm of graph encoder-decoder. However, these methods are usually good at link prediction and node clustering but are unsatisfactory in node and graph classifications. For accurate action recognition, the fine-grained dependencies among different skeleton joints (i.e., graph classifications) are essential. Therefore, previous self-supervised learning-based methods tend to ignore the fine-grained dependencies among different skeleton joints, which restricts the generalization of self-supervised skeleton representation. As shown in Figure 1 (b)-(c), the arm joints and edges are essential to discriminate between these two similar actions. Different from the randomly masking strategy of MAE [20], our masking strategy is action-sensitive and reconstructs specific limbs or body parts that dominate the given action class. Our SkeletonMAE utilizes prior human topology knowledge to guide the reconstruction of the masked skeleton joints and edges to achieve a comprehensive perception of the joints, topology, and action.

To address the aforementioned challenges, we propose an efficient skeleton sequence representation learning framework, named Skeleton Sequence Learning (SSL). To fully discover the fine-grained dependencies among different skeleton joints, we build a novel asymmetric graph-based encoder-decoder pre-training architecture named SkeletonMAE that embeds skeleton joint sequences in Graph Convolutional Network (GCN). The SkeletonMAE aims to reconstruct the masked human skeleton joints and edges based on prior human topology knowledge so that it can infer the underlying topology of the joints and obtain a comprehensive perception of human action. To learn discriminative spatial-temporal skeleton representation, the pre-trained SkeletonMAE encoder is integrated with the Spatial-Temporal Representation Learning (STRL) module to learn spatial-temporal dependencies. Finally, the SSL is fine-tuned on action recognition datasets. Extensive experimental results on FineGym, Diving48, NTU 60 and NTU 120 show that our SSL generalizes well across different datasets and outperforms the state-of-the-art methods significantly. Our contributions are summarized as follows:

- To comprehensively capture human pose and obtain discriminative skeleton sequence representation, we propose a graph-based encoder-decoder pre-training architecture named SkeletonMAE, that embeds skeleton joint sequence into GCN and utilize the prior human topology knowledge to guide the reconstruction of the underlying masked joints and topology.
- To learn comprehensive spatial-temporal dependencies for skeleton sequence, we propose an efficient skeleton sequence learning framework, named Skeleton Sequence Learning (SSL), which integrates the pre-trained SkeletonMAE encoder with the Spatial-Temporal Representation Learning (STRL) module.
- Extensive experimental results on FineGym, Diving48, NTU 60 and NTU 120 datasets show that our SSL outperforms the state-of-the-art self-supervised skeleton-based action recognition methods and achieves comparable performance compared with the state-of-the-art fully supervised methods.

2. Related Work

Action Recognition. One of the most challenging tasks for action recognition is to distinguish similar actions from subtle differences [40, 41, 42]. Recently, some challenging action recognition datasets like FineGym [59], Diving48 [34], NTU RGB+D 60 [58] and NTU RGB+D 120 [37] are proposed. These datasets contain a large number of challenging actions that require discriminative and fine-grained action representation learning. For example, in FineGym [59], an action is divided into action units, sub-actions, or phases, and the model is required to distinguish between "split leap with 1 turn" and "switch leap with 1 turn". The higher inter-class similarity and a new level of granularity in the fine-grained setting make it a challenging task, which makes coarse-grained backbones and methods [14, 2, 71, 79] struggle to overcome. To tackle the more challenging fine-grained action recognition task, most of the existing works [50, 33] are fully supervised and consider fine-grained actions as distinct categories and supervise the model to learn action semantics. However, collecting and labeling these fine-grained actions is time-consuming and labor-intensive, which limits the generalization of a well-trained model to different datasets. To utilize unlabeled data, we propose a graph-based encoder-decoder pre-training architecture named SkeletonMAE.

Skeleton-based Action Recognition. Due to the promising ability to model multiple granularities and large variations in human motion, the skeleton data is more suitable for the aforementioned action recognition task than the RGB data [4]. Early skeleton-based action recognition methods are usually handcrafted, exploiting the geometric relationship of skeleton joints [47, 73, 76, 74], which greatly limits the feature representation of skeletons. Benefiting from the advantages of deep learning, some methods [95, 62, 63] utilized RNNs as the basic model, Du *et al*. [10] presented a pioneering work based on hierarchical RNNs. But RNNs easily suffer from vanishing gradients [21]. Inspired by the booming Graph Convolutional Networks (GCN) [28], Yan *et al*. [83] proposed a spatial-temporal graph convolutional network to learn the spatial and temporal pattern from skeleton data. However, their manually defined topology is arduous to model the relations among joints in underlying topology. Chen *et al*. [5] proposed a channel-wise topology graph convolution, which models channel-wise topology with a refinement method. Duan *et al*. [11] proposed a PoseConv3D model that relies on a 3D heatmap volume instead of a graph sequence as the base representation of human skeletons. Different from previous methods that required large amounts of labeled data for training elaborate models, we utilize unlabeled skeleton sequences to pre-train a graph-based encoder-decoder named SkeletonMAE to comprehensively capture human pose and obtain discriminative skeleton sequence representation.

Self-supervised Learning for Skeleton Sequence. To learn more effective representation for unlabeled skeleton data, self-supervised learning [43] has achieved inspiring progress recently. For contrastive learning approaches, ASCAL [55] and SkeletonCLR [31] applied momentum encoders for contrastive learning with single-stream skeleton sequences. AimCLR [18] used an extreme data augmentation strategy to add additional hard contrastive pairs. Most contrastive learning methods adopt data augmentation to generate positive and negative pairs, but they rely heavily on the number of contrastive pairs. For generative learning approaches, LongT GAN [91] proposed the encoder-decoder to reconstruct masked input sequence skeletons. Based on the LongT GAN, P&C [65] strengthened the encoder and weakened the decoder for feature representation. Wu *et al*. [80] proposed a spatial-temporal masked autoencoder framework for self-supervised 3D skeleton-based action recognition. Colorization [87] used three pairs of encoder-decoder frameworks to learn spatial-temporal features from skeleton point clouds. Due to the limitation of the reconstruction criterion, previous generative methods usually fail to fully discover the fine-grained spatial-temporal dependencies among different skeleton joints. In our SkeletonMAE, we utilize the prior human topology knowledge to infer the skeleton sequence and obtain a comprehensive perception of the action.

3. Methodology

In this section, we introduce the details of Skeleton Sequence Learning (SSL), which contains two parts: 1) pre-training SkeletonMAE and 2) fine-tuning on downstream datasets based on the pre-trained SkeletonMAE.

3.1. Pre-training SkeletonMAE

In this section, we introduce graph-based asymmetric encoder-decoder pre-training architecture SkeletonMAE, to learn human skeleton sequence representations without supervision. Since Graph Isomorphism Network (GIN) [82] provides a better inductive bias, it is more suitable for learning more generalized self-supervised representation [22]. Therefore, we adopt GIN as the backbone of SkeletonMAE. Besides, we evaluate different backbones of SkeletonMAE in Table 4, including GIN [82], GCN [28], and GAT [72].

3.1.1 SkeletonMAE Structure

Inspired by the effective representation learning by masked autoencoder (MAE) in NLP [9], image recognition [20], and video recognition [70], we focus on the human skeleton sequence and build an asymmetric graph-based encoder-decoder pre-training architecture named SkeletonMAE that embeds skeleton sequence and its prior topology knowledge in GIN. The SkeletonMAE is implemented following the paradigm of graph generative self-supervised learning.

We follow the joint label of the Kinetics Skeleton dataset [60]. Specifically, as Figure 2(d) shown, we divide all $N = 17$ joints into $R = 6$ local regions according to the natural parts of the body: $\mathcal{V}_0, ..., \mathcal{V}_5$. Notably, compared to the randomly masking strategy from MAE [20] to select skeleton joints, our masking strategy is action-sensitive and reconstructs specific limbs or body parts that dominate the given action class. Then, we mask these skeleton regions and make the SkeletonMAE reconstruct the masked joint features and their edges based on the adjacent joints. By reconstructing the masked skeleton joints and edges, the SkeletonMAE can infer the underlying topology of joints and obtain a comprehensive perception of the action.

As shown in Figure 2, the SkeletonMAE is an asymmetric encoder-decoder architecture, which includes an encoder and a decoder. The encoder consists of L_D GIN layers that map the input 2D skeleton data to hidden features. The decoder, which consists of only one GIN layer, reconstructs the hidden features under the supervision of the reconstruction criterion. According to the prior human skeleton knowledge that the human skeleton can be represented as a graph with joints as vertices and limbs as edges, we formulate the human skeleton as the following graph structure.

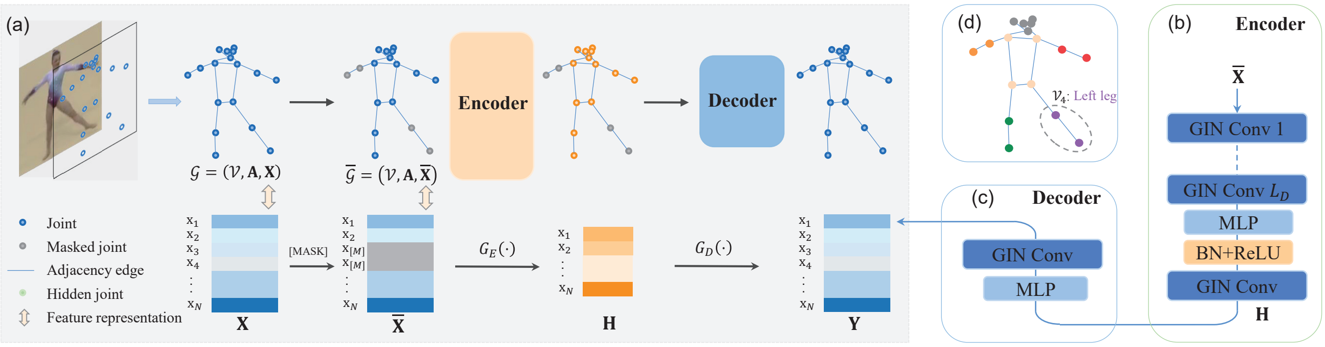


Figure 2: The details of our skeleton sequence pre-training architecture SkeletonMAE. (a) We build a GIN-based asymmetric encoder-decoder structure, to reconstruct joint features to enhance action representation ability. (b) The GIN-based encoder structure contains L_D GIN neural network layers, to learn the joint representation spatially. (c) The decoder consists of one GIN layer, which uses the hidden features from the encoder as the input and reconstructs the original input joint features. (d) Partition the joints in the skeleton sequence according to the natural structure of the human body. 5 joints $\{\mathcal{V}_0$: **Head**$\}$, 4 joints $\{\mathcal{V}_1$: **Torso**$\}$, 2 joints $\{\mathcal{V}_2$: **Left arm**, $\mathcal{V}_3$: **Right arm**, $\mathcal{V}_4$: **Left leg**, $\mathcal{V}_5$: **Right leg**$\}$.

The skeleton sequence of two-dimensional coordinates of N human skeleton joints and T frames is pre-processed in the following way. Specifically, we embed all skeleton joints and their topology into a structure $\mathcal{G}$, the skeleton structure and the joint feature are fused to obtain a joint sequence matrix $\mathbf{S} \in \mathbb{R}^{N\times T\times 2}$. And then the $\mathbf{S}$ is linearly transformed to $\mathbf{S} \in \mathbb{R}^{N\times T\times D}$ with learnable parameters. We empirically set T and D to 64. For each skeleton frame $\mathbf{X} \in \mathbb{R}^{N\times D}$ from $\mathbf{S}$, let $\mathcal{G} = (\mathcal{V}, \mathbf{A}, \mathbf{X})$ denote a skeleton, where $\mathcal{V} = \{v_1, v_2, ..., v_N\}$ is the node set that contains all skeleton joints, $N = |\mathcal{V}|$ is the number of joints. The number of joints is $N = 17$. $\mathbf{A} \in \{0, 1\}^{N\times N}$ is an adjacency matrix, where $\mathbf{A}_{i,j} = 1$ if joints i and j are physically connected, otherwise 0. The feature of v_i is represented as $\mathbf{x}_i \in \mathbb{R}^{1\times D}$. And G_E, G_D denote the GIN encoder and the GIN decoder, respectively.

3.1.2 Skeleton Joints Masking and Reconstruction

Since the prior human skeleton topology $\mathbf{A}$ is embedded (Figure 2) and we specify the aggregation of joints in Section 3.1.1. Inspired by the GraphMAE [22] that randomly reconstructs the masked graph nodes, our SkeletonMAE reconstructs the masked skeleton feature $\mathbf{X}$ based on the prior skeleton topology, rather than reconstructing graph structure $\mathbf{A}$ [68, 17] or reconstructing both structure $\mathbf{A}$ and features $\mathbf{X}$ [57, 53].

To mask skeleton joint features, we randomly select one or more joint sets from $\mathcal{V} = \{\mathcal{V}_0, ..., \mathcal{V}_5\}$, which consists of a subset $\overline{\mathcal{V}} \subseteq \mathcal{V}$ for masking. For the human skeleton sequence, each joint communicates with some of its adjacent joints to represent the specific action class. Therefore, it is not feasible to mask all joint sets for all action classes. Then, each of their features is masked with a learnable mask token vector $[\mathbf{MASK}] = \mathbf{x}_{[\mathbf{M}]} \in \mathbb{R}^D$. Thus, the masked joint feature $\overline{\mathbf{x}}_i$ for $\mathbf{v}_i \in \overline{\mathcal{V}}$ in the masked feature matrix $\overline{\mathbf{X}}$ can be defined as $\overline{\mathbf{x}}_i = \mathbf{x}_{[\mathbf{M}]}$ if $\mathbf{v}_i \in \overline{\mathcal{V}}$, otherwise $\overline{\mathbf{x}}_i = \mathbf{x}_i$. We set $\overline{\mathbf{X}} \in \mathbb{R}^{N\times D}$ as the input joint feature matrix of the SkeletonMAE, and each joint feature in $\overline{\mathbf{X}}$ can be defined as $\overline{\mathbf{x}}_i = \{\mathbf{x}_{[\mathbf{M}]}, \mathbf{x}_i\}$, $i = 1, 2, \cdots, N$. Therefore, the masked skeleton sequence can be formulated as $\overline{\mathcal{G}} = (\mathcal{V}, \mathbf{A}, \overline{\mathbf{X}})$ and the objective of SkeletonMAE is to reconstruct the masked skeleton features in $\overline{\mathcal{V}}$ given the partially observed joint features $\overline{\mathbf{X}}$ with the input adjacency matrix $\mathbf{A}$. The process of SkeletonMAE reconstruction is formalized as:

$$\begin{cases} \mathbf{H} = G_E(\mathbf{A}, \overline{\mathbf{X}}), & \mathbf{H} \in \mathbb{R}^{N\times D_h} \\ \mathbf{Y} = G_D(\mathbf{A}, \mathbf{H}), & \mathbf{Y} \in \mathbb{R}^{N\times D} \end{cases}, \tag{1}$$

where $\mathbf{H}$ and $\mathbf{Y}$ denote the encoder output and the decoder output, respectively. The objective of SkeletonMAE can be formalized as minimizing the divergence between $\mathbf{X}$ and $\mathbf{Y}$.

3.1.3 Reconstruction Criterion

The common reconstruction criterion for masked autoencoders is a mean squared error (MSE) in image and video tasks. However, for skeleton sequence, the multi-dimensional and continuous nature of joint features makes MSE hard to achieve promising feature reconstruction because the MSE is sensitive to dimensionality and vector norms of features [15]. Inspired by the observation [16] that the l_2-normalization in the cosine error maps vectors to a unit hyper-sphere and substantially improves the training stability, we utilize the cosine error as the reconstruction.

To make the reconstruction criterion focus on harder ones among imbalanced easy-and-hard samples [22], we use the Re-weighted Cosine Error (RCE) for SkeletonMAE. The RCE is based on the intuition that we can down-weigh easy samples' contribution in training by scaling the cosine error with a power of $\beta \geq 1$. For predictions with high confidence, their corresponding cosine errors are usually smaller than 1 and decay faster to zero when the scaling factor $\beta > 1$. Formally, given the original feature $\mathbf{X} \in \mathbb{R}^{N \times D}$ and the reconstructed output $\mathbf{Y} \in \mathbb{R}^{N \times D}$, the RCE is defined as:

$$\mathcal{L}_{\mathrm{RCE}} = \sum_{\mathbf{v}_i \in \widetilde{\mathcal{V}}} \left(\frac{1}{|\widetilde{\mathcal{V}}|} - \frac{\mathbf{x}_i^{\mathrm{T}} \cdot \mathbf{z}_i}{|\widetilde{\mathcal{V}}| \times \|\mathbf{x}_i\| \times \|\mathbf{z}_i\|}\right)^{\beta}, \qquad (2)$$

which represents the average of the similarity gap between the reconstructed feature and the input feature over all masked joints. And β is set to 2 in our work.

By training the SkeletonMAE to reconstruct the skeleton sequence, the pre-trained SkeletonMAE can comprehensively perceive the human skeleton structure and obtain discriminative action representation. After pre-training, the SkeletonMAE can be elegantly embedded into the Skeleton Sequence Learning (SSL) framework for fine-tuning.

3.2. Fine-tuning for Skeleton Action Recognition

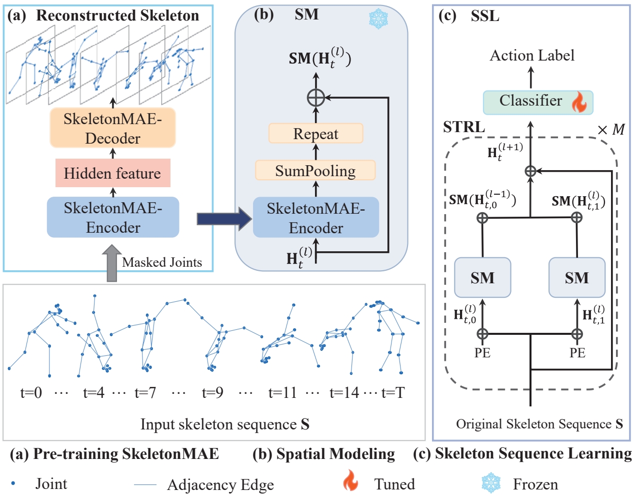
To evaluate the SkeletonMAE's generalization ability for skeleton action recognition, we build a complete skeleton action recognition model named Skeleton Sequence Learning (SSL), based on the pre-trained SkeletonMAE. To capture multiple-person interaction, we integrate two pre-trained SkeletonMAE encoders to build the Spatial-Temporal Representation Learning (STRL) module, as shown in Figure 3(b)-(c). The entire SSL consists of an M-layer STRL model and a classifier. The SSL model is finally fine-tuned on skeleton action recognition datasets with cross-entropy loss to recognize actions.

3.2.1 Spatial-Temporal Representation Learning

The STRL contains two pre-trained SkeletonMAE encoders for Spatial Modeling (SM). The input of SM is skeleton sequence $\mathbf{S}$. The output of SM is connected with the input by 1×1 convolution for residual connection (Figure 3 (b)).

As shown in Figure 3 (c), the input skeleton sequence $\mathbf{S} \in \mathbb{R}^{N \times T \times D}$ is firstly added with the learnable temporal position embedding PE to obtain the skeleton sequence feature $\mathbf{H}_t^{(l)} \in \mathbb{R}^{P \times N \times D^{(l)}}$. To model multiple human skeleton interactions, we obtain two individual features ($P = 2$) for two persons $\mathbf{H}_{t,0}^{(l)} \in \mathbb{R}^{N \times D^{(l)}}$ and $\mathbf{H}_{t,1}^{(l)} \in \mathbb{R}^{N \times D^{(l)}}$ from $\mathbf{H}_t^{(l)}$. Here, we take the joint feature of the 0-th person as an example, the operation of the 1-th person is implemented similarly. We send the joint representation $\mathbf{H}_{t,0}^{(l)}$ and prior knowledge of the joint $\widetilde{\mathbf{A}}$ into the SM module,

$$\mathrm{SM}(\mathbf{H}_{t,0}^{(l)}) = \mathrm{Repeat}(\mathrm{SP}(G_E\left(\widetilde{\mathbf{A}}, \mathbf{H}_{t,0}^{(l)}\right)); N) \oplus \mathbf{H}_{t,0}^{(l)}, \qquad (3)$$



(a) Pre-training SkeletonMAE (b) Spatial Modeling (c) Skeleton Sequence Learning

Figure 3: The pipeline of Skeleton Sequence Learning (SSL). (a) During pre-training, we build an encoder-decoder module named SkeletonMAE that embeds skeleton joints and its prior topology knowledge into GIN and reconstructs the underlying masked joints and topology. (b) The SM consists of the pre-trained SkeletonMAE encoder. (c) We integrate SM structures to build the M-layer Spatial-Temporal Representation Learning (STRL) model and then conduct end-to-end fine-tuning.

where G_E is the SkeletonMAE encoder, $\mathrm{SP}(\cdot)$ denotes the sum-pooling, $\mathrm{Repeat}(\cdot; N)$ means repeating the single joint into N joints representations after sum-pooling and then connect it with the $\mathbf{H}_{t,0}^{(l)}$ residual to get the global joint representation $\mathrm{SM}(\mathbf{H}_{t,0}^{(l)})$. In this way, the SM module can obtain global information through a single joint representation, and constrain some joint features through all joint representations. Similarly, $\mathrm{SM}(\mathbf{H}_{t,1}^{(l)})$ is obtained in the same way. As shown in Figure 3(c), we get the joint features $\mathrm{SM}(\mathbf{H}_t^{(l)})$ that contains the action interaction bewtween the 0-th person and the 1-th person. According to the update rules of graph convolution [28], we can obtain $\mathbf{H}_t^{(l+1)}$ from $\mathbf{H}_t^{(l)}$ in a multi-layer GCN. For more details, please refer to the Supplementary in Section D. The final skeleton sequence representation is defined as follows:

$$\mathbf{H}_t^{(l+1)} = \sigma\left(\mathrm{SM}(\mathbf{H}_t^{(l)})\mathbf{W}^{(l)}\right). \qquad (4)$$

where $\mathbf{W}^{(l)}$ denotes the trainable weight matrix in the l^{th} layer, $\sigma(\cdot)$ denotes the ReLU activation function. Then, we adopt the widely-used multi-scale temporal pooling [5, 38] to get the final output. Finally, a classifier consisting of MLP and softmax predicts the action class.

4. Experiments

All experiments are conducted with a single modality (2D pose) and evaluated on the corresponding train/test sets.

4.1. Datasets

We evaluate our SSL on four benchmark datasets FineGym [59], Diving48 [34], NTU RGB+D 60 [58] and NTU RGB+D 120 [37] in the mainstream skeleton action recognition task. For all datasets except FineGym, we follow the pre-processing protocol provided by [11] to obtain the skeleton sequence from the 2D pose estimator. The pre-processing adopts a Top-Down approach for pose extraction, where the detector is Faster-RCNN [56] with ResNet50 backbone and the pose estimator is HRNet [67] pre-trained on COCO-keypoint [36]. To make a fair comparison, we added pixel noise to the joint during training, making the original joint confidence rate unreliable, thus we do not use the originally fixed threshold.

FineGym is a large-scale fine-grained action recognition dataset for gymnastic videos, which contains 29K videos of 99 fine-grained gymnastics action classes, which requires action recognition methods to distinguish different sub-actions in the same video. In particular, it provides temporal annotations at both action and sub-action levels with a three-level semantic hierarchy. We follow the method [11] to extract the skeleton data from the 2D pose estimator.

NTU RGB+D 60 and 120. NTU RGB+D is a large-scale skeleton-based action recognition dataset, where NTU 60 contains 56,880 skeleton sequences and 60 action classes. NTU 120 has 114,480 skeleton sequences and 120 action categories. The action samples are captured from 155 different camera viewpoints. The subjects in NTU 120 are in a wide range of age distribution (from 10 to 57) and from different cultural backgrounds (15 countries), which brings very realistic variation to the quality of actions. The NTU 60 and 120 datasets have a large amount of variation in subjects, views, and backgrounds.

Diving48 is a challenging fine-grained dataset that focuses on complex and competitive sports content analysis. It is formed of over 18k video clips from competitive dive championships that are distributed over 48 fine-grained classes with minimal biases. The difficulties of the dataset lie in that actions are similar and differ in body parts and their combinations which require the model to capture details and motion in body parts and combine them to perform classification. We report the accuracy on the official train/test split.

4.2. Implementation Details

In this paper, our SkeletonMAE is optimized by the Adaptive Moment Estimation (Adam) with the initial learning rate as $1.5e^{-4}$ and the PReLU is the activation function. The batch size is 1024 and the training epoch is 50. At the fine-tuning stage, we use the Stochastic Gradient Descent (SGD) with momentum (0.9) and adopt the warm-up strategy for the first 5 epochs. The total fine-tuning epochs are 110. The learning rate is initialized to 0.1 and is divided by 10 at the 90 epoch and the 100 epoch. And we employ 0.1 for label smoothing. We use a large batch size of 128 to facilitate training our attention mechanism and enhancing the model's perception for all human action classes. Both our pre-training and fine-tuning models are implemented by PyTorch [54], and our SSL is trained on a single NVIDIA GeForce RTX 2080Ti GPU. For more details of implementation, please refer to the Supplementary in Section A.

Method	Modality	Mean Acc. (%)
Fully Supervised		
I3D [2]	RGB	64.4
ST-S3D [81]	RGB	72.9
TSN [77]	RGB+Flow	79.8
TRNms [93]	RGB+Flow	80.2
TSM [35]	RGB+Flow	81.2
GST-50 [46]	RGB	84.6
MTN [30]	RGB	88.5
LT-S3D [81]	RGB	88.9
TQN [90]	RGB+Text	90.6
PYSKL [11]	Skeleton	93.2
PYSKL [11]	RGB+Skeleton+Limb	**95.6**
Unsupervised Pre-train		
SaL [49]	RGB	42.7
TCC [12]	RGB	45.6
GTA [19]	RGB	49.5
CARL [3]	RGB	60.4
SSL (ours)	Skeleton	**91.8**

Table 1: The comparison with the state-of-the-art unsupervised pre-train and supervised methods on FineGym.

Pre-training and Fine-tuning Setting. For each dataset, the SkeletonMAE encoder is pre-trained with unlabeled data from the training set. Then, we load the learned parameter weights to fine-tune the SSL model.

Evaluation Metrics. To make a fair comparison, we follow previous methods [11, 3, 18] and report the Mean Top-1 accuracy(%) on FineGym dataset and Top-1 accuracy(%) on Diving48, NTU 60, and NTU 120 datasets.

4.3. Downstream Evaluation

For a fair comparison, we compare SSL with other models without pre-training on large-scale action datasets, *e.g.*, Kinetics [27] or Sports1M [26]. The comparison results on FineGym, NTU 60 & 120, and Diving48 datasets are shown in Table 1, Table 2, and Table 3, respectively.

Results on FineGym Dataset. In Table 1, our SSL with skeleton input outperforms most of the fully supervised

Method	Backbone	Supervision	Joint Number	2D Skeleton	NTU 60		NTU 120	
					X-sub (%)	X-view (%)	X-sub (%)	X-set (%)
ST-GCN [83]	GCN	Fully Supervised	25	✗	81.5	88.3	-	-
AS-GCN [32]	GCN	Fully Supervised	25	✗	86.8	94.2	-	-
2s-AGCN [60]	GCN	Fully Supervised	25	✗	88.5	95.1	82.9	84.9
Shift-GCN [6]	GCN	Fully Supervised	25	✗	90.7	96.5	85.9	87.6
MS-G3D [45]	GCN	Fully Supervised	25	✗	91.5	96.2	86.9	88.4
CTR-GCN [5]	GCN	Fully Supervised	25	✗	92.4	**96.8**	**88.9**	**90.6**
PYSKL [11]	CNN	Fully Supervised	17	✓	**93.7**	96.6	86.0	89.6
SkeletonCLR [31]	ST-GCN	Unsupervised Pre-train	25	✗	82.2	88.9	73.6	75.3
CrosSCLR [31]	ST-GCN	Unsupervised Pre-train	25	✗	86.2	92.5	80.5	80.4
Wu *et al.* [80]	STTFormer	Unsupervised Pre-train	25	✗	86.6	92.9	76.8	79.1
AimCLR [18]	ST-GCN	Unsupervised Pre-train	25	✗	86.9	92.8	80.1	80.9
3s-PSTL [94]	ST-GCN	Unsupervised Pre-train	25	✗	87.1	93.9	81.3	82.6
Colorization [87]	DGCNN	Unsupervised Pre-train	25	✗	88.0	94.9	-	-
SSL(ours)	STRL	Unsupervised Pre-train	17	✓	**92.8**(↑ 4.8)	**96.5**(↑ 1.6)	**84.8**(↑ 3.5)	**85.7**(↑ 3.1)

Table 2: The comparison with state-of-the-art unsupervised pre-train and supervised methods on NTU 60 and NTU 120 datasets. '_' means the method with the second-best performance under unsupervised pre-training.

Method	Pre-train	GFLOPs	Acc.(%)
Fully Supervised			
TSN [34]	ImageNet	-	16.8
TRN [25]	ImageNet	-	22.8
P3D [46]	ImageNet	-	32.4
C3D [46]	ImageNet	-	34.5
CorrNet [46]	-	74.8	35.5
CorrNet-R101 [75]	ImageNet	187.3	38.2
MG-TEA-ResNet50 [92]	ImageNet	-	39.5
GSM [66]	ImageNet	107.4	40.3
TSM-R50 [29]	ImageNet	153.8	41.6
TMF [78]	ImageNet	-	42.2
Unsupervised Pre-train			
RESOUND-C3D [34]	K400	-	16.4
Jenni *et al.* [23]	K400	-	29.9
SSL (ours)	Diving48	42.8	**34.1**

Table 3: The comparison with the unsupervised pre-train and supervised methods on the Diving48 dataset.

methods and achieves the best performance among unsupervised pre-train methods with RGB input. For the same input modality, our performance is lower than the fully supervised method PYSKL[11] (with the skeleton as input) by about 1.4%, because the PYSKL adopts stacks of visual heatmaps of skeleton joints as input while we only use human skeleton coordinates. This validates the promising discriminative ability of our skeleton sequence representation.

Results on NTU 60 and NTU 120 Datasets. In Table 2, for NTU 60 X-sub and NTU 60 X-view, compared with unsupervised pre-train methods, our SSL outperforms the current state-of-the-art method Colorization [87] by 4.8% and 1.6%, respectively. Our SSL is also competitive compared with fully supervised methods, outperforming the first six fully supervised methods on NTU 60 X-sub. For NTU 120 X-sub and NTU 120 X-set, our SSL outperforms the previous best-unsupervised pre-train method 3s-PSTL [94] by 3.5% and 3.1%, respectively. Our SSL is superior compared with some fully supervised methods on NTU 120 X-sub and NTU 120 X-set. These results show that our SSL can learn discriminative skeleton representation from large-scale action recognition datasets due to the promising generalization ability of our SkeletonMAE.

Results on Diving48 Dataset. Our SSL with skeleton input outperforms some fully supervised methods. Although our SSL is not pre-trained on additional large-scale pre-training action datasets in Table 3, it still achieves competitive performance among unsupervised pre-train methods. This validates that our pre-training model SkeletonMAE can learn discriminative skeleton sequence representation.

The results on FineGym and Diving48 validate that our SkeletonMAE has a promising ability to enhance the feature representation of skeleton sequence by comprehensively perceiving the underlying topology of actions, and the SSL can learn discriminative action representation.

4.4. Ablation Studies

In this section, we analyze the contributions of essential components and hyper-parameters of our model. Note that unless otherwise specified, all experiments are verified on the FineGym dataset with masking body sub-region 3.

Whether to load pre-trained model or not. To explore the effectiveness of loading the pre-trained SkeletonMAE encoder, we find that the accuracy is 86.3 without load-

L_D	Mean Acc.
1	89.6
2	90.7
3	**91.2**
4	90.9

(a)

Method	Mean Acc.
GraphCL [89]	86.5
JOAO [88]	88.7
Ours(SkeletonMAE)	**91.2**

(b)

# Masked Body Part	0	1	2	3	4	5
GAT [72]	86.8	88.1	88.9	89.5	89.4	90.0
GCN [28]	87.6	88.9	89.3	90.6	89.5	90.5
GIN [82]	**88.6**	**89.5**	**90.2**	**91.2**	**90.3**	**91.2**

(c)

Table 4: (a) Mean accuracy of using the different number of GIN layers in SkeletonMAE encoder. (b) Comparison results with the contrastive learning method as the pre-trained encoder. (c) The results of using different backbones in SkeletonMAE under each masked body part are compared.

ing the pre-trained SkeletonMAE encoder (randomly initialized weights). As Table 5(a) shows, loading the pre-trained model is always better than not loading it. This validates that our SkeletonMAE can learn more comprehensive and generalized representations for unlabeled fine-grained actions by reconstructing the skeleton joint features.

GIN layers in SkeletonMAE. Table 4(a) shows the performance of using different GIN layers in the SkeletonMAE encoder. The performance is the best when $L_D = 3$.

Comparison with contrastive learning methods. To verify the superior ability of our SkeletonMAE when conducting skeleton sequence pre-training, we compare our SkeletonMAE with different contrastive learning methods GraphCL and JOAO. As shown in Table 4(b), our SkeletonMAE achieves the best performance. Besides, we visually compare the action representations of SkeletonMAE and GraphCL by PCA, as shown in Figure 4(a) and Figure 4(b). Compared to GraphCL, the skeleton representation of our SkeletonMAE appears to have a larger inter-class variance and smaller intra-class variance. This validates that our SkeletonMAE can comprehensively capture the human pose and obtain discriminative skeleton sequence representation. We observe similar patterns in all other classes but visualize only five categories for simplicity.

Backbones and masked body parts in SkeletonMAE. As shown in Table 4(c), we show the accuracy of our SSL with different SkeletonMAE backbones and different masked body parts in SkeletonMAE. It can be seen that GIN is always better than both GAT and GCN under the same masked body part. This is because that GIN provides a better inductive bias for graph-level applications. Thus, it is more suitable for learning more generalized skeleton representations. Additionally, we can see that masking body sub-regions 3 and 5 are both optimal across all backbones, which demonstrates the importance of reconstruction of human limbs in action recognition.

M	1	2	3	4
Mean Acc.	89.1(↑2.8)	90.6(↑4.3)	**91.2(↑4.9)**	91.0(↑4.7)

(a)

# Masked Joints Number		5	9	12	15
Ratio of Mask Joints		30%	50%	70%	90%
Accuracy of SSL		89.7	90.3	89.9	90.1
Masked Body Part	**High**	91.8 ($\mathcal{V}_3$,$\mathcal{V}_5$)	91.2 ($\mathcal{V}_0$,$\mathcal{V}_3$,$\mathcal{V}_5$)	91.0 ($\mathcal{V}_1$,$\mathcal{V}_2$,$\mathcal{V}_3$,$\mathcal{V}_4$,$\mathcal{V}_5$)	90.8 ($\mathcal{V}_0$,$\mathcal{V}_1$,$\mathcal{V}_2$,$\mathcal{V}_3$,$\mathcal{V}_5$)
	Low	91.1 ($\mathcal{V}_2$,$\mathcal{V}_4$)	90.1 ($\mathcal{V}_0$,$\mathcal{V}_1$)	91.0 ($\mathcal{V}_1$,$\mathcal{V}_2$,$\mathcal{V}_3$,$\mathcal{V}_4$,$\mathcal{V}_5$)	90.2 ($\mathcal{V}_0$,$\mathcal{V}_1$,$\mathcal{V}_3$,$\mathcal{V}_4$,$\mathcal{V}_5$)

(b)

Table 5: (a) Results of four SSL variants. ↑ represents the accuracy improvement relative to the random initialization of SkeletonMAE in SSL. (b) The comparison of our body part based masked and the random masked strategies. $\mathcal{V}_0$-$\mathcal{V}_5$: Head, Torso, Left arm, Right arm, Left leg, Right leg.

Variants of SSL. To evaluate whether our pre-trained SkeletonMAE is effective across different skeleton action recognition models, we set the different number of STRL layers ($M = 1, 2, 3, 4$) to obtain four variants of the SSL. As shown in Table 5(a), all SSL variants outperform the random initialization of SkeletonMAE in SSL, which validates our body part masking strategy indeed improves the discriminative ability of skeleton feature by learning action-sensitive visual concepts. Additionally, three-layer STRL is the best due to the good compromise between efficiency and computational cost. Moreover, it also validates that our SkeletonMAE generalizes well across different models.

Skeleton Masked Strategy. In Table 5(b), our masked body part strategy is fairly compared with the random masked strategy under the same masked joint conditions. Our method is better than the random mask method across all settings. As mentioned in Section 3.1.1, our masking strategy is action-sensitive and reconstructs specific limbs or body parts that dominates the given action class and is suitable for real-world skeleton-based action recognition. From Table 5, we can see that SSL effectively learns meaningful representations that encode the spatial relationships between joints, enabling the capture of crucial information about the structure and configuration of human movement. This notable achievement can be attributed to the integration of prior knowledge about the human body through pre-training the SkeletonMAE. We report results on the ntu 60 dataset in Supplementary Section E.

Transferability of the SkeletonMAE across datasets.

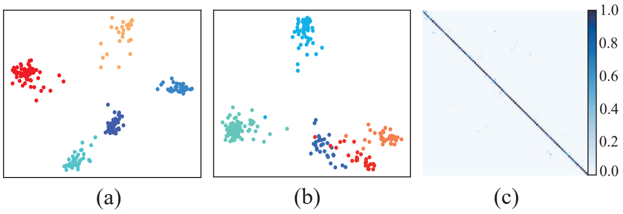


Figure 4: (a) and (b) are 2d-PCA of SkeletonMAE and GraphCL as pre-trained encoder representations. We randomly select five action classes for 2d-PCA visualization, (c) Confusion matrix for fine-grained action recognition.

As shown in Figure 5(a), we pre-train SkeletonMAE on the FineGym dataset and then fine-tune it on NTU 60 X-sub and NTU 120 X-sub datasets. Compared with the method that uses the same dataset for pre-training and fine-tuning, our SkeletonMAE achieves better performance across all masked strategies when conducting dataset transfer. This shows that the SkeletonMAE can learn generalized skeleton representation and effectively transfer the strong representation ability to other datasets.

4.5. Visualization Analysis

Figure 6 shows the reconstruction process of the skeleton sequence by SkeletonMAE. From the same frame, the difference between the reconstructed skeleton sequence and the original skeleton sequence is slight, but overall the human body structure is reserved. This shows that the SkeletonMAE has good spatial representation learning ability. Moreover, SSL effectively captures the temporal evolution and distinguishing characteristics of actions by capturing the relationships between consecutive joint positions and poses. This highlights the positive impact of the multi-scale temporal dependence incorporated within the STRL module. Figure 4(c) shows that our SSL works well for fine-grained action recognition tasks on the FineGym dataset. More visualization results are in Supplementary Section C.

5. Conclusion

In this paper, we propose an efficient skeleton sequence learning framework SSL, to learn discriminative skeleton-based action representation. To comprehensively capture the human pose and obtain skeleton sequence representation, we propose a graph-based encoder-decoder pre-training architecture, SkeletonMAE, that embeds skeleton joint sequence into GCN and utilize the prior human topology knowledge to guide the reconstruction of the underlying masked joints and topology. Extensive experimental results show that our SSL achieves SOTA performance on four benchmark skeleton-based action recognition datasets. In our future work, we will build a multi-level feature refinement module to identify ambiguous skeleton actions.

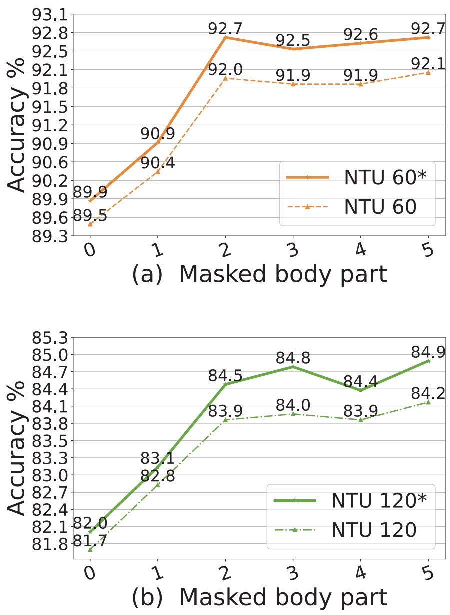


Figure 5: (a) The accuracies with mask body part of 0-5 on the NTU 60 X-sub dataset, and (b) the accuracies on the NTU 120 X-sub dataset. '*' means SkeletonMAE encoder pre-trained on the FineGym dataset.

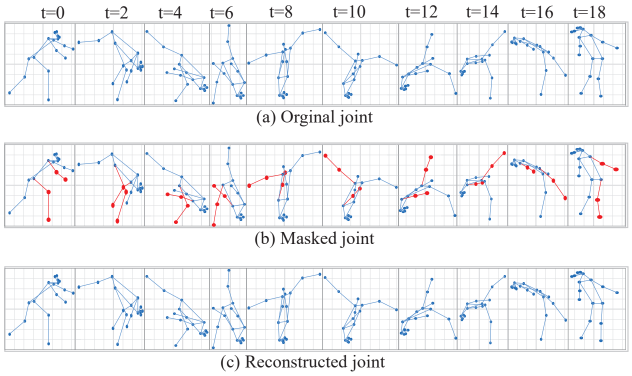


Figure 6: Visualization results for skeleton sequence of "aerial walkover forward" on FineGym. (a) The input skeleton sequence. (b) Masked skeleton sequence (masked parts are 3 and 5). (c) Reconstructed skeleton sequence.

Acknowledgement

This work is supported in part by the National Key R&D Program of China under Grant No.2021ZD0111601, in part by the National Natural Science Foundation of China under Grant No.62002395 and No.61976250, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant No.2023A1515011530, No.2021A1515012311 and No.2020B1515020048, and in part by the Guangzhou Science and Technology Planning Project under Grant No. 2023A04J2030.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 2, 6

[3] 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. 1, 6

[4] Tailin Chen, Desen Zhou, Jian Wang, Shidong Wang, Yu Guan, Xuming He, and Errui Ding. Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4334–4342, 2021. 3

[5] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021. 2, 3, 5, 7

[6] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 183–192, 2020. 7

[7] Sangwoo Cho, Muhammad Maqbool, Fei Liu, and Hassan Foroosh. Self-attention network for skeleton-based human action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 635–644, 2020. 1

[8] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1831–1840, 2017. 1

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[10] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015. 3

[11] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022. 1, 2, 3, 6, 7

[12] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1801–1810, 2019. 6

[13] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 445–450, 2016. 1

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1, 2

[15] Jerome H Friedman. On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, 1(1):55–77, 1997. 4

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 4

[17] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016. 4

[18] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 762–770, 2022. 2, 3, 6, 7

[19] Isma Hadji, Konstantinos G Derpanis, and Allan D Jepson. Representation learning via global temporal alignment and cycle-consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11068–11077, 2021. 6

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 3

[21] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001. 3

[22] Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 594–604, 2022. 3, 4

[23] Simon Jenni and Hailin Jin. Time-equivariant contrastive video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9970–9980, 2021. 7

[24] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14(2):201–211, 1973. 1

[25] Gagan Kanojia, Sudhakar Kumawat, and Shanmuganathan Raman. Attentive spatio-temporal representation learning for diving classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 7

[26] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 6

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

[28] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 3, 5, 8

[29] Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho. Learning self-similarity in space and time as generalized motion for video action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13065–13075, 2021. 7

[30] Mei Chee Leong, Hui Li Tan, Haosong Zhang, Liyuan Li, Feng Lin, and Joo Hwee Lim. Joint learning on the hierarchy representation for fine-grained human action recognition. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1059–1063. IEEE, 2021. 6

[31] 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. 2, 3, 7

[32] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3595–3603, 2019. 7

[33] Tianjiao Li, Lin Geng Foo, Qiuhong Ke, Hossein Rahmani, Anran Wang, Jinghua Wang, and Jun Liu. Dynamic spatio-temporal specialization learning for fine-grained action recognition. In *European Conference on Computer Vision*, pages 386–403. Springer, 2022. 2

[34] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018. 2, 6, 7

[35] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 6

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[37] 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. 2, 6

[38] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017. 5

[39] Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[40] Yang Liu, Zhaoyang Lu, Jing Li, and Tao Yang. Hierarchically learned view-invariant representations for cross-view action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2416–2430, 2018. 2

[41] Yang Liu, Zhaoyang Lu, Jing Li, Tao Yang, and Chao Yao. Deep image-to-video adaptation and fusion networks for action recognition. *IEEE Transactions on Image Processing*, 29:3168–3182, 2019. 2

[42] Yang Liu, Keze Wang, Guanbin Li, and Liang Lin. Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition. *IEEE Transactions on Image Processing*, 30:5573–5588, 2021. 2

[43] Yang Liu, Keze Wang, Lingbo Liu, Haoyuan Lan, and Liang Lin. Tcgl: Temporal contrastive graph for self-supervised video representation learning. *IEEE Transactions on Image Processing*, 31:1978–1993, 2022. 3

[44] Yang Liu, Yu-Shen Wei, Hong Yan, Guan-Bin Li, and Liang Lin. Causal reasoning meets visual representation learning: A prospective study. *Machine Intelligence Research*, 19(6):485–511, 2022. 1

[45] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020. 1, 7

[46] Chenxu Luo and Alan L Yuille. Grouped spatial-temporal aggregation for efficient action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5512–5521, 2019. 6, 7

[47] Fengjun Lv and Ramakant Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *European conference on computer vision*, pages 359–372. Springer, 2006. 3

[48] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017. 2

[49] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European conference on computer vision*, pages 527–544. Springer, 2016. 6

[50] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 122–132, 2020. 2

[51] Ashish S Nikam and Aarti G Ambekar. Sign language recognition using image based hand gesture recognition techniques. In *2016 online international conference on green engineering and technologies (IC-GET)*, pages 1–5. IEEE, 2016. 1

[52] Cosmas Ifeanyi Nwakanma, Fabliha Bushra Islam, Mareska Pratiwi Maharani, Dong-Seong Kim, and Jae-Min Lee. Iot-based vibration sensor data collection and emergency detection classification using long short term memory (lstm). In *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 273–278. IEEE, 2021. 1

[53] Jiwoong Park, Minsik Lee, Hyung Jin Chang, Kyuewang Lee, and Jin Young Choi. Symmetric graph convolutional autoencoder for unsupervised graph representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6519–6528, 2019. 4

[54] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[55] Haocong Rao, Shihao Xu, Xiping Hu, Jun Cheng, and Bin Hu. Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. *Information Sciences*, 569:90–109, 2021. 3

[56] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 6

[57] Amin Salehi and Hasan Davulcu. Graph attention autoencoders. *arXiv preprint arXiv:1905.10715*, 2019. 4

[58] 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. 2, 6

[59] 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. 1, 2, 6

[60] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019. 1, 3, 7

[61] Lei Shi, Yifan Zhang, Jing Hu, Jian Cheng, and Hanqing Lu. Gesture recognition using spatiotemporal deformable convolutional representation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1900–1904. IEEE, 2019. 1

[62] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. 3

[63] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. Spatio-temporal attention-based lstm networks for 3d action recognition and detection. *IEEE Transactions on image processing*, 27(7):3459–3471, 2018. 3

[64] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852. PMLR, 2015. 2

[65] 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, 3

[66] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-shift networks for video action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1102–1111, 2020. 7

[67] 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. 1, 6

[68] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077, 2015. 4

[69] Fida Mohammad Thoker, Hazel Doughty, and Cees GM Snoek. Skeleton-contrastive 3d action representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1655–1663, 2021. 1

[70] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 3

[71] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 2

[72] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017. 3, 8

[73] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 588–595, 2014. 3

[74] Raviteja Vemulapalli and Rama Chellapa. Rolling rotations for recognizing human actions from 3d skeletal data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4471–4479, 2016. 3

[75] Heng Wang, Du Tran, Lorenzo Torresani, and Matt Feiszli. Video modeling with correlation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 352–361, 2020. 7

[76] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1290–1297. IEEE, 2012. 3

[77] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 6

[78] Yanze Wang and Junyong Ye. Tmf: Temporal motion and fusion for action recognition. *Computer Vision and Image Understanding*, 213:103304, 2021. 7

[79] Zhengwei Wang, Qi She, and Aljosa Smolic. Action-net: Multipath excitation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13214–13223, 2021. 2

[80] Wenhan Wu, Yilei Hua, Shiqian Wu, Chen Chen, Aidong Lu, et al. Skeletonmae: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition. *arXiv preprint arXiv:2209.02399*, 2022. 3, 7

[81] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 6

[82] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018. 3, 8

[83] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 1, 2, 3, 7

[84] Jie Yang, Chaoqun Wang, Zhen Li, Junle Wang, and Ruimao Zhang. Semantic human parsing via scalable semantic transfer over multiple label domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19424–19433, 2023. 1

[85] Jie Yang, Ailing Zeng, Shilong Liu, Feng Li, Ruimao Zhang, and Lei Zhang. Explicit box detection unifies end-to-end multi-person pose estimation. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[86] LI Yang, Jin Huang, TIAN Feng, WANG Hong-An, and DAI Guo-Zhong. Gesture interaction in virtual reality. *Virtual Reality & Intelligent Hardware*, 1(1):84–112, 2019. 1

[87] 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. 3, 7

[88] Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. Graph contrastive learning automated. In *International Conference on Machine Learning*, pages 12121–12132. PMLR, 2021. 8

[89] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823, 2020. 8

[90] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4486–4496, 2021. 6

[91] 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*, volume 32, 2018. 2, 3

[92] Yuan Zhi, Zhan Tong, Limin Wang, and Gangshan Wu. Mgsampler: An explainable sampling strategy for video action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1513–1522, 2021. 7

[93] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 803–818, 2018. 6

[94] Yujie Zhou, Haodong Duan, Anyi Rao, Bing Su, and Jiaqi Wang. Self-supervised action representation learning from partial spatio-temporal skeleton sequences. *arXiv preprint arXiv:2302.09018*, 2023. 7

[95] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016. 3