Global-local Motion Transformer for Unsupervised Skeleton-based Action Learning

Boeun Kim[1,2], Hyung Jin Chang[3], Jungho Kim[2], and Jin Young Choi[1]

[1] ASRI, Dept. of ECE., Seoul National University
[2] AIRC, Korea Electronics Technology Institute
[3] School of Computer Science, University of Birmingham

Abstract. We propose a new transformer model for the task of unsupervised learning of skeleton motion sequences. The existing transformer model utilized for unsupervised skeleton-based action learning is learned the instantaneous velocity of each joint from adjacent frames without global motion information. Thus, the model has difficulties in learning the attention globally over whole-body motions and temporally distant joints. In addition, person-to-person interactions have not been considered in the model. To tackle the learning of whole-body motion, long-range temporal dynamics, and person-to-person interactions, we design a global and local attention mechanism, where, global body motions and local joint motions pay attention to each other. In addition, we propose a novel pretraining strategy, multi-interval pose displacement prediction, to learn both global and local attention in diverse time ranges. The proposed model successfully learns local dynamics of the joints and captures global context from the motion sequences. Our model outperforms state-of-the-art models by notable margins in the representative benchmarks. Codes are available at *https://github.com/Boeun-Kim/GL-Transformer*.

Keywords: Unsupervised pretraining, action recognition, Transformer

1 Introduction

In skeleton-based action recognition, to avoid expensive and time-consuming annotation for supervised learning, recent studies have focused on unsupervised learning techniques for pretraining [40, 30, 18, 37, 16, 25, 27, 35, 31, 8]. For unsupervised pretraining suitable for action recognition, learning the global context of the entire motion sequence is essential along with learning local joint dynamics and topology. However, existing methods have limitations in effectively capturing both global context and local joint dynamics.

Several existing unsupervised pretraining methods exploit RNN-based encoder-decoder models [40, 30, 18, 37, 16, 25]. However, RNN-based methods have difficulties in extracting global contexts because of the long-range dependency problem [10, 6]. Other approaches utilize contrastive learning schemes [27, 35, 17]. However, the performance of these methods has been reported to be highly dependent on the selection of the encoder model because the contrastive loss does not induce detailed learning in the local dynamics of the joints [35, 17].

Recently, the transformer, widely used for natural language processing and image recognition, has been applied to the unsupervised pretraining of skeleton-based action recognition. The first and the only model is H-transformer [8], which learns to predict the direction of the instantaneous velocity of joints in each frame. H-transformer still has limitations in learning global attention because predicting only the instantaneous velocity induces the model to learn the local attention rather than the global context in whole-body motions. In addition, H-transformer does not consider person-to-person interactions which are important for classifying actions performed by two or more persons.

In this paper, to tackle the learning of global context, long-range temporal dynamics, and person-to-person interactions, we propose a novel transformer-based pretraining model, which is called GL-Transformer. To this end, we design the GL-Transformer architecture that contains global and local attention (GLA) mechanism. The GLA mechanism comprises spatial multi-head attention (spatial-MHA) and temporal multi-head attention (temporal-MHA) modules. Using the input body motions disentangled into global body motions and local joint motions, the spatial-MHA module performs three types of attention: local(inter-joint), global(body)-from/to-local(joint), and global(person)-to-global(person) attentions. The temporal-MHA module performs global and local attention between any two frames for sequences of every person.

In addition, a novel pretraining strategy is proposed to induce GL-Transformer to learn global attention across the long-range sequence. For the pretraining, we design a multi-task learning strategy referred to as multi-interval pose displacement prediction (MPDP). For MPDP, GL-Transformer is trained with multiple tasks to predict multiple pose displacements (angle and movement distance of every joint) over different intervals at the same time. GL-Transformer learns local attention from a small interval, as well as global attention from a large interval. To enhance performance, we add two factors to GL-Transformer. First, to learn natural joint dynamics across frames, we impose natural-speed motion sequences instead of sequences sampled to a fixed length. Next, we introduce a trainable spatial-temporal positional embedding and inject it to each GL-Transformer block repeatedly to use the order information in every block, which is the valuable information of the motion sequence.

We demonstrate the effectiveness of our method through extensive experimental evaluations on widely used datasets: NTU-60 [28], NTU-120 [19], and NW-UCLA [34]. In the linear evaluation protocol [40], the performance of GL-Transformer exceeds that of H-transformer [8] and other state-of-the-art (SOTA) methods by notable margins. Furthermore, our method even outperforms SOTA methods in semi-supervised settings. The main contributions of this study are summarized as follows:

1. We design a novel transformer architecture including global and local attention (GLA) mechanism to model local joint dynamics and capture the global context from skeleton motion sequences with multiple persons (Sec. 3.2).
2. We introduce a novel pretraining strategy, multi-interval displacement prediction (MPDP), to learn attention in diverse temporal ranges (Sec. 3.3).

3. GL-Transformer renews the state-of-the-art score in extensive experiments on three representative benchmarks: NTU-60, NTU-120, and NW-UCLA.

2 Related Works

Unsupervised Skeleton-based Action Recognition. Earlier unsupervised learning methods for skeleton-based action recognition can be divided into two categories: using RNN-based encoder-decoders and contrastive learning schemes. Several existing methods utilize RNN-based encoder-decoder networks [40, 30, 18, 37, 16, 25]. The decoder of these networks performs a pretraining task to induce the encoder to extract an appropriate representation for action recognition. The decoder of LongT GAN [40] reconstructs the randomly corrupted input sequence conditioned on the representation. MS^{2}L [18] learns to generate more general representations through multi-task learning, which performs tasks such as motion prediction and jigsaw puzzle recognition. Recently, Colorization [38] adopted a GCN to pretrain which regresses the temporal and spatial orders of a skeleton sequence. RNN-based models suffer from long-range dependencies, and the GCN-based models have a similar challenge because they deliver information sequentially along a fixed path [26, 10, 6]. Therefore, the RNN and GCN-based methods have limitations in extracting global representations from the motion sequence, especially from long motions.

Other methods exploit the contrastive learning scheme [27, 35, 31]. These methods augment the original motion sequence and regard it as a positive sample while considering other motion sequences as negative samples. The model is then trained to generate similar representations between the positive samples using contrastive loss. AS-CAL [27] leverages various augmentation schemes such as rotation, shear, reverse, and masking. Contrastive learning schemes have a limitation, in that all sequences other than themselves are regarded as negative samples, even sequences belonging to the same class. CrosSCLR [17] alleviates this issue by increasing the number of positive samples using representations learned from other views, such as velocity or bone sequences. Because the contrastive learning loss adjusts the distances between the final representations extracted by an encoder, it is difficult to train the encoder to reflect the local joint dynamics explicitly by the loss. To address the limitations in both categories of unsupervised action recognition, we introduce the transformer [33] architecture for modeling the local dynamics of joints and capturing the global context from motion sequences.

Transformer-based Supervised Learning. Transformer-based models have achieved remarkable success in various supervised learning tasks using motion sequences, owing to their attention mechanism, which is suitable for handling long-range sequences. In supervised action recognition tasks, recent transformer-based methods [26, 10, 3, 24] outperform GCN-based methods, which have limitations in yielding rich representations because of the fixed graph topology of the human body. In the motion prediction task, the method in [6, 1] employs a

transformer encoder to capture the spatial-temporal dependency of a given motion sequence and a transformer decoder to generate future motion sequences. In the 3D pose estimation task, the method in [39] imposes a 2D pose sequence on the spatial-temporal transformer to model joint relations and estimate the 3D pose of the center frame accurately.

Transformer-based Pretraining. Transformer-based pretraining has become the dominant approach in natural language processing [11, 20], and is being actively introduced to other research fields such as vision-language [23, 32, 14], images [12, 13, 7, 4], and videos [36, 21]. The H-transformer [8] is the first transformer-based pretraining method for motion sequences. The proposed pretraining strategy predicts the direction of the instantaneous velocity of the joints in each frame. This strategy focuses on learning attention from adjacent frames rather than from distant frames. The model is designed to learn spatial attention between five body part features, where global body movement is not considered. To address these limitations, we propose a GL-Transformer that contains a global and local attention mechanism and a novel pretraining strategy. We aim to train GL-Transformer to generate a representation of input motion sequences suitable for the downstream action recognition task by modeling local and global attention effectively in the pretraining process.

3 Proposed Method

3.1 Overall Scheme

Our goal is to build a transformer architecture suitable for the skeleton motion sequence (Sec. 3.2) and design a novel pretraining strategy (Sec. 3.3) for encoding both the internal dynamics and the global context of the motion sequence. As illustrated in Fig. 1, the proposed framework comprises two stages: unsupervised pretraining and downstream action recognition stages. In the first stage, we pretrain the proposed transformer-based model, GL-Transformer, with unlabeled motion sequences. Next, we verify that GL-Transformer generates the

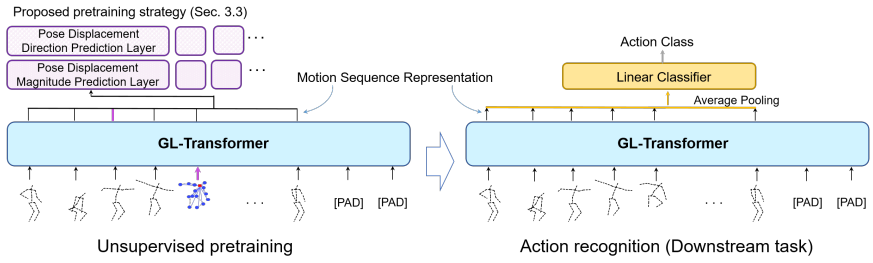

Fig. 1. Overall scheme of the proposed framework. GL-Transformer is pretrained with unlabeled motion sequences, and then evaluated in downstream action recognition task

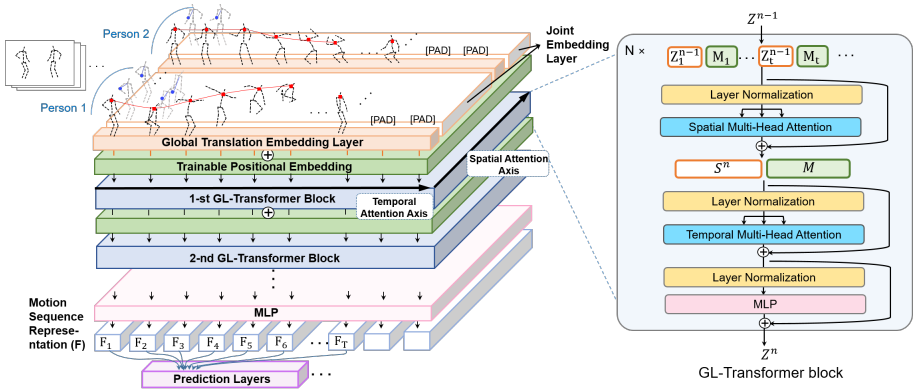

Fig. 2. Model architecture. The input motion sequence is disentangled into global translational motions (red dots) and local motions (blue dots). The proposed model comprises N stacked GL-Transformer blocks. Global and local attention mechanism is implemented in both spatial-MHA and temporal-MHA modules in each block.

appropriate motion representation required for the action recognition. A single linear classifier is attached after GL-Transformer. After average pooling is applied to the motion sequence representation for the temporal axis, it is passed to the classifier.

3.2 Model Architecture

Our model comprises N stacked GL-Transformer blocks, as illustrated in Fig. 2, and each block contains spatial multi-head attention (spatial-MHA) and temporal multi-head attention (temporal-MHA) modules sequentially, as illustrated by the blue boxes on the right side in Fig. 2.

Input Motion Sequences. As illustrated at the top of the left figure in Fig. 2, the input human motion sequence is expressed by two types of information: global translational motion (red dots) of the body and local motions of the body joints (blue dots). The global translational motion represents the trajectory of the center joint of the body, and the local motions represent the relative motions of the body joints from the center joint. The center joint is defined in each dataset, for example, NTU datasets [28, 19] define the spine joint as the center joint. The original 3D skeleton motion sequence is expressed by tensor $X = [X_1, X_2, \cdots, X_T]^T$, where X_t is a matrix representing the skeleton pose at the t-th frame. The pose matrix X_t is defined by $X_t = [q_t^1, q_t^2, \cdots, q_t^K]^T$, where $q_t^k \in \mathbb{R}^3$ indicates the 3-dimensional vector for the k-th joint coordinate. The relative position of the k-th joint is $r_t^k = q_t^k - q_t^c$, where q_t^c denotes the coordinate of the center joint. Using the relative joint positions, the t-th frame of local motion is expressed by a matrix $R_t = [r_t^1, \cdots, r_t^K]^T$, in which we remove $r_t^c = (0, 0, 0)$ and re-index it to $K-1$ dimensional matrix as $R_t = [r_t^1, \cdots, r_t^{K-1}]^T$. The t-th frame

of the global translational motion is calculated using the vector $g_t = q_t^c - q_0^c$. As in Fig. 2, g_t and r_t^k are projected into D dimensional embedding vectors as

$$\bar{g}_t = W_g g_t + b_g, \quad \bar{r}_t^k = W_r r_t^k + b_r, \quad k = 1, \cdots, K-1, \tag{1}$$

where $W_g, W_r \in \mathbb{R}^{(D\times 3)}$ and $b_g, b_r \in \mathbb{R}^{(D\times 1)}$ denote trainable weights and biases of the global translation and joint embedding layers respectively. In the case of an action dataset containing the interaction between two or more persons, vector g_t and matrix R_t are expressed by $g_{t,p}$ and $R_{t,p}$ respectively, where p denotes an index of the person. Similarly, the embedding vectors are expressed as $\bar{g}_{t,p}$, and $\bar{r}_{t,p}^k$. In the following, we describe our method which considers the interaction among multiple persons in the sequence.

Trainable and Tight Positional Embedding. By extending the concept of the positional embedding matrix [33] containing the order information of a sequence, we introduce a trainable spatial-temporal positional embedding tensor $M \in \mathbb{R}^{T\times PK\times D}$ to learn the order information of both the temporal frames and spatial joints from the training data. Note that PK is the dimension for the joint indices of P persons, and D is the dimension of embedding vectors, same as D in $\bar{g}_{t,p}$ and $\bar{r}_{t,p}^k$. Joint order information plays a more important role in skeleton motion sequences than in the case of sentences or images, in that individual joint positions are not meaningful until we know which part of the body the joint belongs to. Furthermore, frame order also plays an important role in detecting the action. To this end, we propose a tight positional embedding method to use order information explicitly in every GL-Transformer block. Previous transformer-based models [11, 12, 39] apply positional embedding once before the first transformer block. In contrast, we apply it to the input tensors of every block, as illustrated in Fig. 2. In each GL-Transformer Block, the positional embedding is explicitly applied in both the spatial-MHA and temporal-MHA modules, as illustrated in the right figure of Fig. 2.

Global and Local Attention (GLA) Mechanism. We aim to construct a global and local attention (GLA) mechanism to extract global semantic information along with capturing the local relationships between the joints within the skeleton motion sequence. GLA is implemented in both the spatial-MHA and temporal-MHA modules. The spatial-MHA module learns spatial dependency within one frame. In the module, global(body)-from/to-local(joint) dependencies are learned by an attention operation between the features corresponding to $g_{t,p}$ and $R_{t,p}$ of each person. Likewise, person-to-person dependencies are learned by the attention among the features of multiple persons: $\{g_{t,p}, R_{t,p} | p = 1, \cdots, P\}$, where P is the number of persons. The temporal-MHA module learns the temporal dependencies across the sequence using pose features aggregated by the spatial-MHA. The temporal-MHA module learns whole-body motion information from distant frames as well as local joint dynamics from the adjacent frames.

Input pose features of the spatial-MHA module at the t-th frame in the n-th block is denoted by $Z_t^n \in \mathbb{R}^{PK\times D}$. For the first block, embeddings of multiple

people are concatenated along the spatial attention axis (see Fig. 2) as

$$Z_t^0 = ||_{p=1}^{P} Z_{t,p}^0, \tag{2}$$

$$Z_{t,p}^0 = [\bar{g}_{t,p}, \bar{r}_{t,p}^1, \cdots \bar{r}_{t,p}^{K-1}]^T, \quad t = 1, \cdots, T, \tag{3}$$

where $||$ indicates the concatenation operation. The spatial-MHA in $n(\geq 2)$-th block receives the output (Z_t^{n-1}) of the previous block. The spatial-MHA module updates the pose features as

$$S_t^n = \text{spatial-MHA}(LN(Z_t^{n-1} + M_t)) + (Z_t^{n-1} + M_t), \tag{4}$$

where $M_t \in \mathbb{R}^{PK \times D}$ is t-th slice of the positional embedding tensor M. $LN(\cdot)$ denotes the layer normalization operator [2]. For the spatial-MHA$(\cdot)$, we borrow the multi-head self-attention (MHA) mechanism from [33], which is described below. For simplicity, we denote $LN(Z_t^{n-1} + M_t)$ as $\hat{Z}_t^{n-1}$. First, $\hat{Z}_t^{n-1}$ is projected to *query* Q, *key* K, *value* V matrices as

$$Q = \hat{Z}_t^{n-1} W^Q, \quad K = \hat{Z}_t^{n-1} W^K, \quad V = \hat{Z}_t^{n-1} W^V, \tag{5}$$

where $W^Q, W^K, W^V \in \mathbb{R}^{D \times d}$ are weight matrices for the projection and d indicates the projection dimension. The attention mechanism is expressed as

$$\text{Attention}(Q, K, V) = softmax(QK^T/\sqrt{d})V. \tag{6}$$

Note that QK^T refers to the dot-product similarity of each projected joint vector in *query* Q to *key* K. High attention weight is given for high similarity. In the MHA, the i-th head performs the attention mechanism in Eq.(6) with different weight matrices W_i^Q, W_i^K, W_i^V from those of other heads as

$$H_i = \text{Attention}(\hat{Z}_t^{n-1} W_i^Q, \hat{Z}_t^{n-1} W_i^K, \hat{Z}_t^{n-1} W_i^V), \quad i = 1, \cdots, h. \tag{7}$$

The concatenation of $\{H_i\}$ is projected to an aggregated pose features as

$$\text{spatial-MHA}(\hat{Z}_t^n) = (||_{i=1}^{h} H_i) W_H, \tag{8}$$

where $W_H \in \mathbb{R}^{dh \times dh}$ is a projection matrix.

To perform temporal-MHA in the n-th block, we vectorize the pose feature of the t-th frame $S_t^n \in R^{PK \times D}$ into $s_t^n \in R^{PK \cdot D}$. Then, the vectorized pose features are stacked to form a pose feature sequence matrix $S^n = [s_1^n, s_2^n, \ldots, s_T^n]^T \in \mathbb{R}^{T \times (PK \cdot D)}$. In the temporal-MHA module, the same MHA mechanism in Eq.(8) is used, but different weight matrices are applied. Then, the output pose sequence feature of the n-th GL-Transformer (Z^n) is obtained through MLP$(\cdot)$, that is,

$$\bar{Z}^n = \text{temporal-MHA}(LN(S^n + \bar{M})) + (S^n + \bar{M}), \tag{9}$$

$$Z^n = MLP(LN(\bar{Z}^n)) + \bar{Z}^n, \tag{10}$$

where $\bar{M} \in \mathbb{R}^{T \times (PK \cdot D)}$ is a matrix in which the dimension of the positional embedding tensor $M \in \mathbb{R}^{T \times PK \times D}$ is changed. In the N-th GL-Transformer block, the final motion sequence representation F for the input motion sequence X is obtained by passing Z^N through a 2-layer MLP as

$$F = \text{GL-Transformer}(X) = MLP(Z^N). \tag{11}$$

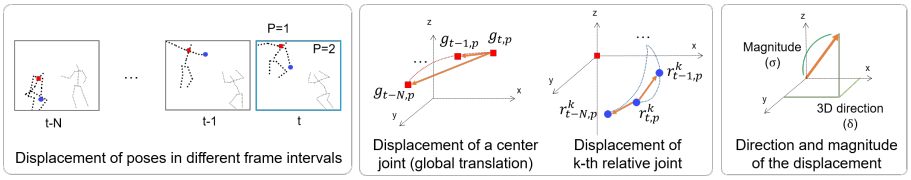

Fig. 3. Description of multi-interval pose displacement

Masked Attention for Natural-Speed Motion Sequence. Most of the existing action recognition methods [8, 30, 17, 38] employ a fixed length of motion sequences, which overlooks the importance of the speed of the motion. To handle natural-speed motion sequences, we utilize an attention mask [33], so that our model can learn the natural joint dynamics across frames and capture speed characteristics from diverse actions. To this end, we define the maximum sequence length as T_{max}. If the length of the original sequence X_{ori} is shorter than T_{max}, the rest of the frames are filled with padding dummy tokens $[PAD] \in \mathbb{R}^{PK\times 3}$, which yields $X = [X_{ori}^T, [PAD], ..., [PAD]]^T \in \mathbb{R}^{T_{max}\times PK\times 3}$. Elements of $[PAD]$ are set to arbitrary numbers because the loss corresponding to the $[PAD]$ token is excluded. To exclude attention from the dummy values, we mask (setting to $-\infty$) columns corresponding to the $[PAD]$ tokens in the QK^T matrix.

3.3 Multi-interval Pose Displacement Prediction (MPDP) Strategy

We design a novel pretraining strategy, multi-interval pose displacement prediction (MPDP), which estimates the whole-body and joint motions at various time intervals at different scales. H-transformer [8] introduces a pretraining strategy that estimates the direction of the instantaneous joint velocity. The instantaneous velocity of the joint in a specific frame can be easily obtained from the adjacent frame so that the model is guided to learn local attention rather than long-range global attention. To overcome this limitation, we propose an MPDP strategy to effectively learn global attention as well as local attention.

As illustrated in Fig. 3 (a), we first select multiple frame intervals $t-N, \cdots, t-n, \cdots, t$. GL-Transformer is trained to predict the magnitude and direction of the pose displacement between the t-th and $(t-n)$-th frame. Local motion (relative joint displacement) is predicted with the help of global motion and vice versa. In addition, the motion of other people is considered when predicting one's motion. The displacements are represented by the orange arrows in Fig. 3 (b) and (c). We design the pose displacement prediction as a classification task using *softmaxed* linear classifiers. The model is trained to predict both the direction and magnitude classes for each interval. The predictions of the t-th frame for the interval n are expressed as

$$\hat{\Delta}_{t,n} = softmax(W_n^{\delta} F_t + b_n^{\delta}), \quad \hat{\Sigma}_{t,n} = softmax(W_n^{\sigma} F_t + b_n^{\sigma}), \tag{12}$$

where F_t denotes t-th slice of the motion sequence representation F, as shown in the left side of Fig. 2. $\hat{\Delta}_{t,n} = ||_{p=1}^{P} \hat{\Delta}_{t,p,n}$ where $\hat{\Delta}_{t,p,n} = [\hat{\delta}_{t,p,n}^{g}, \hat{\delta}_{t,p,n}^{1}, \cdots, \hat{\delta}_{t,p,n}^{K-1}]^T$,

and $\hat{\delta}^g_{t,p,n}$, $\hat{\delta}^k_{t,p,n} \in \mathbb{R}^{C_\delta}$ denotes the predicted direction class vector of global translation and k-th joints, respectively. C_δ is the number of direction classes. $\hat{\Sigma}_{t,n} = ||_{p=1}^{P} \hat{\Sigma}_{t,p,n}$ where $\hat{\Sigma}_{t,p,n} = [\hat{\sigma}^g_{t,p,n}, \hat{\sigma}^1_{t,p,n}, \cdots, \hat{\sigma}^{K-1}_{t,p,n}]^T$, and $\hat{\sigma}^g_{t,p,n}, \hat{\sigma}^k_{t,p,n} \in \mathbb{R}^{C_\sigma}$ denotes the predicted magnitude class vector of the global translation and k-th joints, respectively. C_σ denotes the number of magnitude classes. $W^\delta_n, W^\sigma_n, b^\delta_n$, and b^σ_n are the trainable weights and biases of the linear classifiers for interval n. To train the model parameters, we define the ground truth classes of direction δ and magnitude σ at the t-th frame for the p-th person and interval n as

$$\delta^g_{t,p,n} = class(\angle(g_{t,p} - g_{(t-n),p})), \quad \delta^k_{t,p,n} = class(\angle(r^k_{t,p} - r^k_{(t-n),p})), \tag{13}$$

$$\sigma^g_{t,p,n} = class(\|g_{t,p} - g_{(t-n),p}\|), \quad \sigma^k_{t,p,n} = class(\|r^k_{t,p} - r^k_{(t-n),p}\|), \tag{14}$$

where we set $g_{(t-n),p} = g_{t,p}$ and $r^k_{(t-n),p} = r^k_{t,p}$ at $t \leq n$ because we do not have the information of the $(t-n)$-th frame in this case. $class(\cdot)$ denotes the class label vector of $\cdot$, where the magnitude is quantized into one of the C_σ classes, and the direction is designated as one of the $C_\delta = 27$ classes, in which each of the xyz direction has three classes: $+$,$-$, and no movement. The classification loss is calculated for all intervals and frames except the [PAD] tokens. The total loss is defined as follows:

$$L_{total} = \sum_{t=1}^{T} \sum_{p=1}^{P} \sum_{n}^{N} \Big(\lambda_\delta L_\delta(t,p,n) + \lambda_\sigma L_\sigma(t,p,n) \Big), \tag{15}$$

where direction loss $L_\delta(t,p,n)$ and magnitude loss $L_\sigma(t,p,n)$ are the weighted sum of cross entropy loss to train each component of $\hat{\Delta}_{t,p,n}$ and $\hat{\Sigma}_{t,p,n}$, whereas λ_δ and λ_σ denote the weighting factors of L_δ and L_σ, respectively.

4 Experiments

4.1 Datasets & Evaluation Protocol

NTU-RGB+D. NTU-RGB+D 60 (NTU-60) [28] is a large-scale dataset containing 56,880 3D skeleton motion sequences performed by up to two actors and categorized into 60 action classes. Each person has 25 joints. We follow two standard evaluation criteria: cross-subject (**xsub**) and cross-view (**xview**). In **xsub**, the training and test set are collected by different subjects. **xview** splits the training and testing set according to the camera view. NTU-RGB+D 120 (NTU-120) [19] is an extension of NTU- 60 which contains 113,945 sequences for 120 action classes. The new evaluation criterion cross-setup (**xset**) is added for NTU-120, whose training and testing sets are split by the camera setup IDs.
North-Western UCLA. North-Western UCLA (NW-UCLA) [34] contains 1,494 motion sequences captured by 10 subjects. Each sequence is performed by one actor and each person has 20 joints. The actions are categorized into 10 action classes. Following the standard evaluation protocol, the training set comprises samples from camera views 1 and 2, and the remaining samples from

view 3 are arranged in the testing set.

Evaluation Protocol. We adopt linear evaluation protocol [40, 18, 37, 27, 16, 25, 8] which is the standard for the evaluation of unsupervised learning tasks. Following the protocol, the weight parameters of the pretrained model are fixed, and only the attached single linear classifier is trained with the training data. In addition, we evaluate the proposed model in semi-supervised settings [31, 29, 17, 38]. The pretrained model is fine-tuned with 5% and 10% of the training data, and then the action recognition accuracy is evaluated.

4.2 Implementation Details

We set $T_{max} = 300$ for the NTU dataset and $T_{max} = 50$ for the NW-UCLA dataset. The sequence is augmented by applying a shear [27] and interpolation. For interpolation, the sequence is interpolated into a random length within $\pm$ 10% of the original sequence length. Since the NTU dataset includes two persons, we set it to $P = 2$. Four transformer blocks are utilized, the hidden dimension $D = 6$ for each joint, and eight heads ($h = 8$) are used for self-attention. The H-transformer [8] uses four transformer blocks with $D = 256$ for each of the five body parts. We set $\lambda_\delta, \lambda_\sigma = 1$. In the unsupervised pretraining phase, we utilize the AdamW [22] optimizer with an initial learning rate of $5e^{-4}$ and decay it by multiplying by 0.99 every epoch. The model is trained for 120 epochs for the NTU and 300 epochs for the NW-UCLA with a batch size of 128. In the linear evaluation protocol, we utilize Adam [15] optimizer with a learning rate of $3e^{-3}$. The linear layer is trained for 120 and 300 epochs for NTU and NW-UCLA, respectively, with a batch size of 1024.

4.3 Ablation Study

We conduct ablation studies using the NTU-60 dataset to demonstrate the effectiveness of the main components of our method. The final performance of GL-Transformer substantially exceeds that of the H-transformer [8] in the linear evaluation protocol, exceeding 7.0% for **xsub** and 11.0% for **xview**. The effectiveness of each component is explained as follows:

Effectiveness of GLA and MPDP. In Table 1, Experiment **(1)** exploits the original pose sequence X, Experiment **(2)** utilizes local motion $R_t(t = 1, \cdots, T)$, and Experiment **(3)** utilizes both global translational motion $g_t(t = 1, \cdots, T)$ and local motion $R_t(t = 1, \cdots, T)$. Regarding **(1)**, because global and local motions are mixed in X, it is difficult to model both global and local motions. The result of **(2)** is higher than that of **(1)** when the model learns local dynamics between the joints from local motions. The result of **(3)** is further improved demonstrating that GLA plays an important role in extracting the representation of the entire motion sequence effectively.

In Table 1, Experiment **(3)** does not adopt the displacement magnitude prediction loss, that is $\lambda_\sigma = 0$. For Experiment **(4)**, $\lambda_\sigma = 1$, and predicting both directions and magnitudes exhibits a higher performance. Experiments **(4)** to **(7)** are performed by altering the frame intervals that are utilized in MPDP.

Table 1. Ablation study for verifying the effectiveness of GLA and MPDP in the NTU-60 dataset with the linear evaluation protocol

	Displacement direction	Disentangle global translation	Displacement magnitude	Frame interval	Accuracy(%) xsub	Accuracy(%) xview
H-transformer [8]	✓			{1}	69.3	72.8
Experiment **(1)**	✓			{1}	71.1	73.5
Experiment **(2)**	✓	✓(only local motion)		{1}	74.2	81.9
Experiment **(3)**	✓	✓		{1}	75.4	82.8
Experiment **(4)**	✓	✓	✓	{1}	75.7	82.9
Experiment **(5)**	✓	✓	✓	{1, 5}	75.9	83.3
Experiment **(6)**	✓	✓	✓	{1, 5, 10}	**76.3**	**83.8**
Experiment **(7)**	✓	✓	✓	{1, 5, 10, 15}	75.7	83.4

Table 2. Ablation study for verifying the effectiveness of person-to-person attention in NTU-120 **xsub** (left) and trainable and tight positional embedding (right) in NTU-60 with the linear evaluation protocol

p2p attention	Accuracy(%) one person cat.	two people cat.	total
w/o	63.0	71.6	64.9
w/	**63.7**	**73.5**	**66.0**

Type	Accuracy(%) xsub	xview
Fixed (sinusoidal)	75.5	83.3
Trainable	76.0	83.6
Trainable tight	**76.3**	**83.8**

The performance gradually increases from the interval $n = \{1\}$ to $n = \{1, 5, 10\}$, demonstrating that long-range global attention is effective in aggregating the context of the entire motion sequence. The accuracy corresponding to the interval $n = \{1, 5, 10, 15\}$ is lower than that of $n = \{1, 5, 10\}$. This implies the maximum interval relies on the inter-frame dependency of the given sequence.

Effectiveness of Person-to-person Attention. To verify the effect of person-to-person (p2p) attention, we report the model performance trained with and without p2p attention in Table 2. NTU-120 has 120 action categories, 26 among them are two-person interactions and the rest are one-person actions. The p2p attention improves the performance of both groups, especially the performance increases more in the group of two people categories.

Effectiveness of Trainable Tight Positional Embedding. For positional embedding, the performance increases when a trainable embedding is employed instead of a fixed sinusoidal embedding, as presented in the right table of Table 2. The use of tight embedding further increases the performance. We also verify that frames close to each other are trained to have similar positional embeddings. The corresponding figures are added in the supplementary material. In addition, the experiment demonstrating the effectiveness of natural-speed input is added to the supplementary material.

4.4 Analysis of Learned Attention

We analyze the attention map, $softmax(QK^T/\sqrt{d})$ in Eq.(6), of each pretrained GL-Transformer block. The spatial and temporal attention maps are extracted from the spatial-MHA and temporal-MHA modules, respectively. The attention

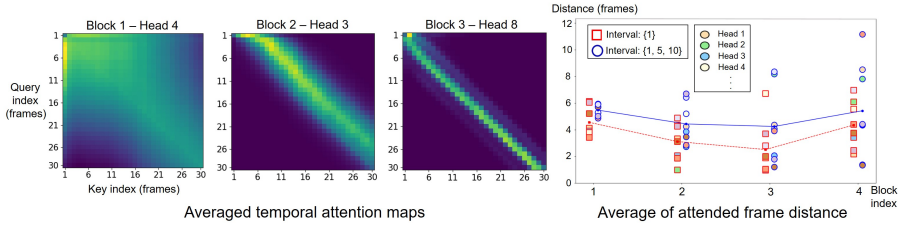

Fig. 4. Examples of learned temporal attention maps averaged over 300 evaluation sequences (left) and average of attended frame distance (right). Yellow color indicates a large value in the left figure. Blue (interval {1,5,10}) and red (interval {1}) lines indicate the average values over heads in each block

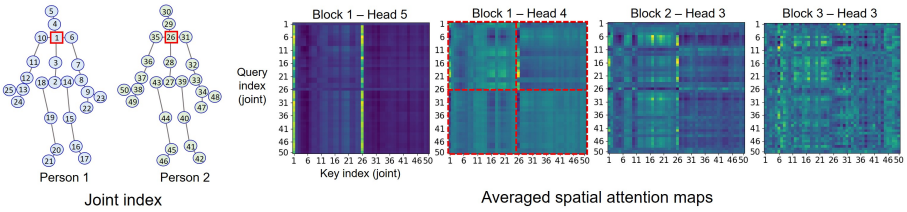

Fig. 5. Examples of learned spatial attention maps averaged over 300 evaluation sequences. Yellow color indicates a large value

maps are averaged over 300 motion sequences from the evaluation data. Each head of each transformer block indicates various types of attention maps, and representative samples are shown in Fig. 4 and Fig. 5. In Fig. 4, we indicate the averaged temporal attention map for the first 30 frames, because the length of test sequences varies from each other. The vertical and horizontal axes represent the *query* and *key* indices, respectively, and the color of each pixel indicates the degree to which the *query* attends to the *key*. Each head attends a different temporal range, for example, approximately neighboring 10 frames and 5 frames are highlighted in the attention maps of Block2-Head3 and Block3-Head8, respectively, whereas a wide range is highlighted in the attention map of Block1-Head4. The figure on the right in Fig. 4 illustrates the average attended frame distances [12] of each head. The average of the attended frame distances [12] is calculated as a weighted sum of the frame distances, where attention is regarded as the weight. Red squares indicate each head when using frame interval $\{1\}$, and blue circles indicate each head when using intervals $\{1, 5, 10\}$. In each block, more heads attend to distant frames when the model is pretrained with intervals $\{1, 5, 10\}$ as compared to when the model is pretrained with interval $\{1\}$.

An example of the spatial attention map is illustrated in Fig. 5. The 1-st and 26-th indices are utilized for the global translations corresponding to $g_{t,1}$ and $g_{t,2}$, respectively, which are represented as red squares in the left figure. In some heads, the 1-st and 26-th indices appear to be attended differently from other joints. For example, in Block1-Head5, *queries* of all joints pay attention to the 1-st and 26-th *keys* of more than *keys* of other joints. In Block1-Head4,

Table 3. Action recognition results with linear evaluation protocol in NTU-60 dataset

Method	Network	Accuracy(%) xsub	xview
LongT GAN (2018) [40]	GRU (encoder-decoder)	39.1	48.1
P&C (2020) [30]	GRU (encoder-decoder)	50.7	76.3
MS^{2}L (2020) [18]	GRU (encoder-decoder)	52.6	-
PCRP (2021) [37]	GRU (encoder-decoder)	54.9	63.4
AS-CAL (2021) [27]	LSTM (contrastive learning)	58.5	64.6
CRRL (2021) [35]	LSTM (contrastive learning)	67.6	73.8
EnGAN-PoseRNN (2019) [16]	RNN (encoder-decoder)	68.6	77.8
SeBiReNet (2020) [25]	GRU (encoder-decoder)	-	79.7
'TS' Colorization (2021) [38]	GCN (encoder-decoder)	71.6	79.9
CrosSCLR-joint (2021) [17]	GCN (contrastive learning)	72.9	79.9
CrosSCLR-bone (2021) [17]	GCN (contrastive learning)	75.2	78.8
H-transformer (2021) [8]	Transformer	69.3	72.8
GL-Transformer	Transformer	**76.3**	**83.8**

correlations between the joints corresponding to each person are observed as 4 divisions in the attention map, as indicated in red dotted lines. Overall, the proposed model learns the global relationships at shallow blocks (i.e. Block1) and learns fine-grained relationships at deeper blocks (i.e. Block2 and Block3).

4.5 Comparison with State-of-the-art Methods

We compared our approach with the state-of-the-art (SOTA) methods for unsupervised action recognition: methods using RNN-based encoder-decoder models [40, 30, 18, 37, 16, 25], a method using GRU-based encoder-decoder model [38], methods using contrastive learning scheme [27, 17], and a transformer-based method [8]. We use a linear evaluation protocol to measure the action recognition accuracy. The performance of our method substantially exceeds that of the H-transformer [8] which focuses only on the local relationship between body parts and between frames. As presented in Table 3, the performance of GL-Transformer exceeds that of the H-transformer by 7.0% in **xsub** and 11.0% in **xview** in the NTU-60 dataset. Furthermore, our method outperforms all previous methods by a notable margin.

On the NTU-120 dataset, GL-Transformer outperforms the SOTA methods with a significant margin, as presented in the left table of Table 4. It is verified that the proposed method operates robustly on datasets that include more detailed actions. On the NW-UCLA dataset, GL-Transformer achieved the highest performance among the previous methods, demonstrating that the proposed model is effective even with a small amount of training data, as presented in the right table of Table 4. In addition, we compare the results from the semi-supervised setting on the NTU-60 and NW-UCLA datasets in Table 5. The results of the SOTA semi-supervised action recognition methods [29, 31] are also compared in conjunction with the unsupervised methods aforementioned. GL-Transformer exceeds SOTA performance in both evaluations using 5% and 10% of the training data.

Table 4. Action recognition results with linear evaluation protocol in the NTU-120 dataset (left) and NW-UCLA dataset (right)

Method	Accuracy(%)	
	xsub	xset
P&C (2020) [30]	41.7	42.7
PCRP (2021) [37]	43.0	44.6
AS-CAL (2021) [27]	48.6	49.2
CrosSCLR-bone (2021) [17]	53.3	50.6
CRRL (2021) [35]	56.2	57.0
CrosSCLR-joint (2021) [17]	58.8	53.3
GL-Transformer	**66.0**	**68.7**

Method	Accuracy(%)
LongT GAN (2018) [40]	74.3
MS^{2}L (2020) [18]	76.8
SeBiReNet (2020) [25]	80.3
CRRL (2021) [35]	83.8
P&C (2020) [30]	84.9
PCRP (2021) [37]	86.1
'TS' Colorization (2021) [38]	90.1
H-transformer (2021) [8]	83.9
GL-Transformer	**90.4**

Table 5. Results with semi-supervised setting in the NTU-60 and NW-UCLA datasets

Methods	NTU-60 (xsub)		NTU-60 (xview)		NW-UCLA	
	5%	10%	5%	10%	5%	10%
MCC-ST-GCN (2021) [31]	42.4	55.6	44.7	59.9	-	-
MCC-2s-AGCN (2021) [31]	47.4	60.8	53.3	65.8	-	-
MCC-AS-GCN (2021) [31]	45.5	59.2	49.2	63.1	-	-
LongT GAN (2018) [40]	-	62.0	-	-	-	59.9
ASSL (2020) [29]	57.3	64.3	63.6	69.8	52.6	-
MS^{2}L (2020) [18]	-	65.2	-	-	-	60.5
CrosSCLR-bone (2021) [17]	59.4	67.7	57.0	67.3	-	-
'TS' Colorization (2021) [38]	60.1	66.1	63.9	73.3	55.9	71.3
CrosSCLR-joint (2021) [17]	61.3	67.6	64.4	73.5	-	-
GL-Transformer	**64.5**	**68.6**	**68.5**	**74.9**	**58.5**	**74.3**

5 Conclusions

We introduce a novel transformer architecture and pretraining strategy suitable for motion sequences. The proposed GL-Transformer successfully learns global and local attention, so that the model effectively captures the global context and local dynamics of the sequence. The performance of our model substantially exceeds those of SOTA methods in the downstream action recognition task in both unsupervised and self-supervised manners. In future studies, our model can be extended to a model for learning various skeleton features together, such as the position and bone, to encode richer representations. The memory usage and computation of the model are expected to be reduced by using the concept of sparse attention [9, 5], which sparsely pays attention to each other among tokens. Furthermore, our model can be extended to a large-parameter model and pretrained with a large number of skeleton sequences extracted from unspecified web videos to be more generalized, and can be applied to various downstream tasks dealing with human actions.

Acknowledgement. This work was supported by IITP/MSIT [B0101-15-0266, Development of high performance visual bigdata discovery platform for large-scale real-time data analysis, 1/4; 2021-0-01343, AI graduate school program (SNU), 1/4; 2021-0-00537, Visual common sense through self-supervised learning for restoration of invisible parts in images, 1/4; 1711159681, Development of high-quality AI-AR interactive media service through deep learning-based human model generation technology, 1/4]

References

1. Aksan, E., Kaufmann, M., Cao, P., Hilliges, O.: A spatio-temporal transformer for 3d human motion prediction. In: 2021 International Conference on 3D Vision (3DV). pp. 565–574. IEEE (2021)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Bai, R., Li, M., Meng, B., Li, F., Ren, J., Jiang, M., Sun, D.: Gcst: Graph convolutional skeleton transformer for action recognition. arXiv preprint arXiv:2109.02860 (2021)
4. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
5. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020)
6. Cai, Y., Huang, L., Wang, Y., Cham, T.J., Cai, J., Yuan, J., Liu, J., Yang, X., Zhu, Y., Shen, X., et al.: Learning progressive joint propagation for human motion prediction. In: European Conference on Computer Vision. pp. 226–242. Springer (2020)
7. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12299–12310 (2021)
8. Cheng, Y.B., Chen, X., Chen, J., Wei, P., Zhang, D., Lin, L.: Hierarchical transformer: Unsupervised representation learning for skeleton-based human action recognition. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)
9. Child, R., Gray, S., Radford, A., Sutskever, I.: Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509 (2019)
10. Cho, S., Maqbool, M., Liu, F., Foroosh, H.: Self-attention network for skeleton-based human action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 635–644 (2020)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021)
14. Huang, Z., Zeng, Z., Liu, B., Fu, D., Fu, J.: Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849 (2020)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Kundu, J.N., Gor, M., Uppala, P.K., Radhakrishnan, V.B.: Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In: 2019 IEEE winter conference on applications of computer vision (WACV). pp. 1459–1467. IEEE (2019)
17. Li, L., Wang, M., Ni, B., Wang, H., Yang, J., Zhang, W.: 3d human action representation learning via cross-view consistency pursuit. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4741–4750 (2021)

18. Lin, L., Song, S., Yang, W., Liu, J.: Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2490–2498 (2020)
19. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: 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)
20. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
21. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. arXiv preprint arXiv:2106.13230 (2021)
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
23. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)
24. Mazzia, V., Angarano, S., Salvetti, F., Angelini, F., Chiaberge, M.: Action transformer: A self-attention model for short-time pose-based human action recognition. Pattern Recognition **124**, 108487 (2022)
25. Nie, Q., Liu, Z., Liu, Y.: Unsupervised 3d human pose representation with viewpoint and pose disentanglement. In: European Conference on Computer Vision. pp. 102–118. Springer (2020)
26. Plizzari, C., Cannici, M., Matteucci, M.: Spatial temporal transformer network for skeleton-based action recognition. In: International Conference on Pattern Recognition. pp. 694–701. Springer (2021)
27. Rao, H., Xu, S., Hu, X., Cheng, J., Hu, B.: Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. Information Sciences **569**, 90–109 (2021)
28. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1010–1019 (2016)
29. Si, C., Nie, X., Wang, W., Wang, L., Tan, T., Feng, J.: Adversarial self-supervised learning for semi-supervised 3d action recognition. In: European Conference on Computer Vision. pp. 35–51. Springer (2020)
30. Su, K., Liu, X., Shlizerman, E.: Predict & cluster: Unsupervised skeleton based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9631–9640 (2020)
31. Su, Y., Lin, G., Wu, Q.: Self-supervised 3d skeleton action representation learning with motion consistency and continuity. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13328–13338 (2021)
32. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7464–7473 (2019)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
34. Wang, J., Nie, X., Xia, Y., Wu, Y., Zhu, S.C.: Cross-view action modeling, learning and recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2649–2656 (2014)

35. Wang, P., Wen, J., Si, C., Qian, Y., Wang, L.: Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition. arXiv preprint arXiv:2111.11051 (2021)
36. Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Jiang, Y.G., Zhou, L., Yuan, L.: Bevt: Bert pretraining of video transformers. arXiv preprint arXiv:2112.01529 (2021)
37. Xu, S., Rao, H., Hu, X., Cheng, J., Hu, B.: Prototypical contrast and reverse prediction: Unsupervised skeleton based action recognition. IEEE Transactions on Multimedia (2021)
38. Yang, S., Liu, J., Lu, S., Er, M.H., Kot, A.C.: Skeleton cloud colorization for unsupervised 3d action representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13423–13433 (2021)
39. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11656–11665 (2021)
40. Zheng, N., Wen, J., Liu, R., Long, L., Dai, J., Gong, Z.: Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)