Masked Motion Predictors are Strong 3D Action Representation Learners

Yunyao Mao
Institution1
Institution1 address
firstauthor@i1.org

Jiajun Deng
Institution2
First line of institution2 address
secondauthor@i2.org

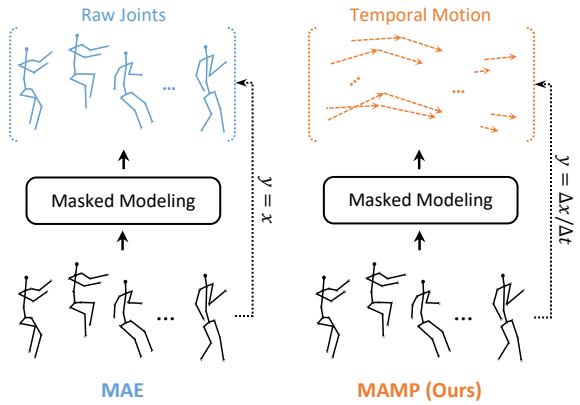
Abstract

In 3D human action recognition, limited supervised data makes it challenging to fully tap into the modeling potential of powerful networks such as transformers. As a result, researchers have been actively investigating effective self-supervised pre-training strategies. In this work, we show that instead of following the prevalent pretext task to perform masked self-component reconstruction in human joints, explicit contextual motion modeling is key to the success of learning effective feature representation for 3D action recognition. Formally, we propose the Masked Motion Prediction (MAMP) framework. To be specific, the proposed MAMP takes as input the masked spatio-temporal skeleton sequence and predicts the corresponding temporal motion of the masked human joints. Considering the high temporal redundancy of the skeleton sequence, in our MAMP, the motion information also acts as an empirical semantic richness prior that guide the masking process, promoting better attention to semantically rich temporal regions. Extensive experiments on NTU-60, NTU-120, and PKU-MMD datasets show that the proposed MAMP pre-training substantially improves the performance of the adopted vanilla transformer, achieving state-of-the-art results without bells and whistles. The source code of our MAMP will be made available for public research.

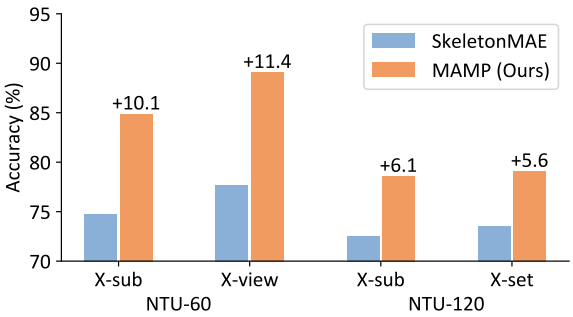
1. Introduction

How to accurately recognize human actions has been a long-standing challenge in computer vision. Recently, with the advances in techniques of depth sensing and pose estimation [3, 14, 52], skeleton-based 3D human action recognition has become an emerging problem to the community, which is of great significance in a series of applications such as human-computer interaction, video surveillance, virtual reality, *etc*. Despite the computation efficiency and background robustness of skeletons, existing supervised 3D action recognition methods [5, 7, 13, 17, 22, 23, 30, 37, 40, 41, 58] heavily rely on well-annotated training sequences, which are labor-intensive and time-consuming to acquire. Furthermore, limited supervision also leads to the overfitting issue in general models, especially for transformers that are with weak inductive bias and high model capacity. These facts motivate the exploration of self-supervised 3D action representation learning.

In the literature, the prevalent pretext tasks originally developed for images have been adapted for 3D action repre-



(a) Comparison of pre-training objectives.



(b) Comparison of linear probing accuracy.

Figure 1. Illustration of (a) pre-training objective comparison between masked auto encoders (MAE) and our masked motion predictors (MAMP) and (b) performance comparison between the typical MAE method, *i.e.*, SkeletonMAE [51], and our MAMP under the linear evaluation protocol.

sentation learning, such as colorization [56], reconstruction [60, 44, 25], contrastive learning [21, 46, 47], *etc*. Among them, contrastive learning once dominated 3D action representation learning with its concise framework and promising performance. Nevertheless, as a global representation learner, it still suffers from certain limitations, such as the lack of explicit constraints for temporal context modeling and the over-reliance on heuristic action data augmentations [29], impeding its further exploration of 3D actions.

Recently, as transformers flourish in computer vision, masked autoencoder (MAE) [15] has attracted a surge of research interest for its exceptional performance. Given that a 3D skeleton serves as an abstract representation of human behaviors, there has been growing interest in applying the MAE concept to 3D action representation learning, to capture the underlying spatio-temporal dynamics of skeleton sequences. Early attempts generally followed the practice of images, employing masked self-reconstruction of human joints as the pre-training pretext. Despite considerable effort, we argue that the network is not effectively directed to prioritize contextual motion modeling in such a self-reconstruction objective, which is, however, crucial for comprehending 3D actions as the appearance information is greatly erased in human skeletons. How to better explore the contextual motion clue in self-supervised 3D action representation learning is a valid problem.

By consolidating this idea, we introduce Masked Motion Prediction (MAMP), a simple yet effective framework to address the problem of self-supervised 3D action representation learning. Specifically, the proposed MAMP takes as input the masked spatio-temporal skeleton sequence and turns to predict the corresponding temporal motion of the masked human joints. In this way, the network is directly encouraged for contextual motion modeling. Moreover, given the observation that moments with significant motion are often critical for human action understanding, in our MAMP, the temporal motion is used not only as the pre-training objective but also as an empirical semantic richness prior that effectively guiding the skeleton masking process. Compared to the random version, the proposed motion-aware masking strategy takes additional temporal motion intensity as input. It first converts the input intensity into a probability distribution and then utilizes the re-parameterization technique for efficient probability-guided masked token sampling. As a result, joints with significant motion are masked with a higher probability, facilitating better attention to semantically rich temporal regions.

As illustrated in Figure 1, compared to masked self-reconstruction of human joints, masked motion prediction acts as a more effective pretext task for 3D action representation learning. It substantially alleviates the problem that the transformers cannot fully unleash their modeling potential for human actions due to the scarcity of annotated 3D skeletons. The adopted vanilla transformer sets a series of state-of-the-art records in 3D action recognition after MAMP pre-training, without the need for bells and whistles such as multi-stream ensembling. Specifically, compared to training from scratch, our MAMP demonstrates significant absolute performance improvements of 10.0% and 13.2% on the challenging cross-subject protocol of NTU RGB+D 60 [36] and NTU RGB+D 120 [26] datasets, resulting in top-1 accuracy of 93.1% and 90.0%, respectively. We hope this simple yet effective framework will serve as a strong baseline that facilitates future research on 3D action pre-training and beyond.

Overall, we make the following three-fold contributions:

- We present masked motion prediction to learn 3D action representation, which substantially alleviates the insufficient contextual motion modeling issue in the conventional masked self-reconstruction paradigm.
- We devise the motion-aware masking strategy, which incorporates motion intensity as an empirical semantic richness prior for adaptive joint masking.
- We conduct extensive experiments on three prevalent benchmarks to verify the effectiveness of our method. Remarkably, with our proposed MAMP, the vanilla transformer, for the first time, achieves the top-performing record for 3D action recognition.

2. Related Work

2.1. Supervised 3D Action Recognition

How to better model the dynamic skeletons for supervised action recognition is an extensively studied problem. In many early works, RNNs are favored for their excellent sequential modeling capability, such as the hierarchical RNN model proposed in [13] and the 2D Spatio-Temporal LSTM in [28, 27]. In view of the great success of CNNs [19, 16] in image understanding, some methods also try to apply it to 3D action recognition. To cater for the input format, [12] and [20] treat the skeleton sequence as a three-channel (x, y, and z coordinates) pseudo-image, with the number of frames and joints as height and width, respectively. Considering the natural connections between joints, ST-GCN [53] introduces the Graph Neural Networks (GCNs) for skeleton modeling, where the convolution kernels are elaborately designed according to the skeleton topology. The astonishing performance of ST-GCN has led the trend of GCN-based 3D action recognition, with numerous subsequent improvements emerging in input streams [49, 24, 38], kernel design [5, 59, 37, 30], *etc*.

Recent approaches [34, 33, 39] try to introduce the popular vision transformer into 3D action recognition. However, under limited training data, vanilla transformers with weak inductive bias cannot be fully trained. Therefore, many customized designs are required in existing supervised at-

tempts, such as temporal convolution [34], graph convolution [33, 34], space-time separation [39], *etc*. In our approach, we demonstrate that pre-training with masked motion prediction is key to the success of transformers in 3D action recognition. The proposed MAMP framework endows the vanilla transformer with unrivaled performance.

2.2. Self-supervised 3D Action Recognition

Self-supervised representation learning aims to capture the domain priors from unlabeled data so as to facilitate the application of the model in downstream tasks. In 3D human action recognition, many pretext tasks have been utilized to explore the action context that resides in the skeleton sequence. Among them, LongT GAN [60] and P&C [44] try to learn 3D action representation by autoencoder-based sequence reconstruction, where the decoder in P&C is further weakened to promote the learning of the feature encoder. In Colorization [56], the skeleton sequences are treated as point clouds and action representation is learned by colorizing each joint based on its spatial and temporal orders.

Recently, many contrastive learning-based approaches [25, 46, 21, 47, 57, 32] have emerged, showing superior performance compared to earlier works. To learn better 3D action representation, they either try to dig helpful supervision across different skeleton modalities [21, 32], or explore better action data augmentation [47] and positive sample mining strategies [57]. Nevertheless, as a global feature learner originally designed for images, contrastive learning lacks explicit constraints on the exploration of temporal motion context, limiting its further development for 3D actions.

SkeletonMAE [51] first introduces the idea of MAE[15] into transformer-based 3D action representation learning, where the original joint coordinates of masked regions are predicted. In our approach, we demonstrate that such a self-reconstruction objective is sub-optimal for learning 3D action representation. Therefore, we introduce the Masked Motion Prediction (MAMP) framework for explicit contextual motion modeling, resulting in significantly better performance compared to raw skeleton reconstruction.

2.3. Masked Visual Prediction

With the development of vision transformers [4, 11, 48], the masked prediction derived from the autoencoder [1] has revived again. Similar to the BERT [10] pre-training in NLP, the input tokens are randomly masked and corresponding objectives are predicted, which can be the raw pixels [15], HOG features [50], or token ids from offline learned dVAEs [2]. Recently, there have also been attempts [43, 55] to use optical flow or temporal difference of images as the auxiliary reconstruction objectives, but inferior performance is observed when they are applied alone. This is largely attributed to the high redundancy of the raw images, where the key foreground motion is difficult to be pre-extracted accurately. In our approach, we employ the idea of masked visual prediction for 3D action representation learning, with the temporal skeleton motion adopted as the only reconstruction target. Different from images, the explicit temporal correspondence of joints in the human skeleton sequence enables the ready extraction of their accurate motion context. Furthermore, we also incorporate motion intensity as the semantic richness prior to guide the masking process.

3. Our Method

3.1. Overview

Figure 2 illustrates the overall pipeline of our proposed Masked Motion Prediction (MAMP) framework. It takes a skeleton sequence $S \in \mathbb{R}^{T_s \times V \times C_s}$ as input, which is randomly cropped from the original data and is resized to a fixed temporal length T_s. V and C_s are the number of joints and coordinate channels, respectively. The motion sequence $M \in \mathbb{R}^{T_s \times V \times C_s}$ of the input is also extracted, which is defined as the differential on temporal dimension (manually padded for the first frame).

As in most vision transformers, the input joints are linearly mapped into joint embedding $E \in \mathbb{R}^{T_e \times V \times C_e}$. After that, the motion-aware masking strategy is applied to mask most of the embedding features under the guidance of temporal motion intensity. The remaining features are processed by the encoder-decoder architecture, where the transformer encoder learns representation from unmasked joint embedding and the transformer decoder performs contextual modeling based on the learnable mask tokens and latent representation from the transformer encoder. Different from MAE [15] that reconstructs the original signal for representation learning, in MAMP, a motion prediction head is adopted, which takes decoded features as input and predicts the temporal motion of the input skeleton sequence.

After the aforementioned pre-training, only the joint embedding layer and the transformer encoder are reserved for downstream applications.

3.2. Joint Embedding

In most transformer-based attempts [33, 39, 34], each spatio-temporal skeleton joint is embedded separately, resulting in a large number of input tokens. Considering the temporal redundancy, in our approach, the input skeleton sequence $S \in \mathbb{R}^{T_s \times V \times C_s}$ is divided into temporally non-overlapping segments $S' \in \mathbb{R}^{T_e \times V \times l \times C_s}$, where l is the length of each segment and $T_e = T_s/l$. In each segment, joints with the same spatial position are embedded together:

$$E = \texttt{JointEmbed}(S') \in \mathbb{R}^{T_e \times V \times C_e}, \tag{1}$$

where C_e is the dimension of the embedding features. Compared to the original skeleton sequence, the temporal resolu-

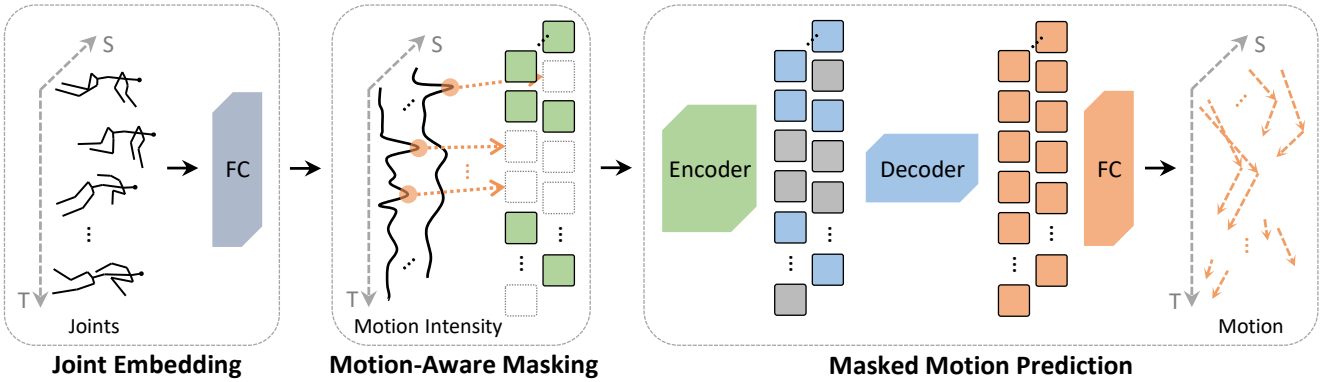


Figure 2. The overall pipeline of the proposed MAMP framework. Different from the self-reconstruction scheme adopted in previous works [60, 51], our MAMP learns 3D action representation by predicting the corresponding motion sequence for masked joint input. Moreover, motion information also acts as an empirical semantic richness prior that effectively guides the masking process, enabling more attention to be applied to regions with significant temporal motion intensity.

tion of the embedding E is reduced by a factor of l, resulting in higher computational efficiency.

3.3. Motion Extraction

Different from RGB frames with heavy spatial redundancy, the human skeleton sequence is highly semantic, with explicit correspondence between neighboring frames. Therefore, we can easily obtain its motion $M \in \mathbb{R}^{T_s \times V \times C_s}$ by applying temporal difference on joint coordinates:

$$M_{i,:,:} = S_{i,:,:} - S_{i-m,:,:}, \ i \in m, m+1, \ldots, T_s - 1, \quad (2)$$

where stride m controls the step size of the motion. For convenience, the motion sequence is padded to be consistent with the length of the original input:

$$M_{0:m-1,:,:} = \begin{cases} 0, & constant \\ M_{m:2m-1,:,:}, & replicate \end{cases}, \quad (3)$$

where $constant$ and $replicate$ denote constant padding (with zeros) and replicate padding for the first m frames of the motion sequence, respectively.

3.4. Motion-Aware Masking

In the proposed approach, the motion information is used not only as the reconstruction target during pre-training but also as the empirical semantic richness prior that guiding the masking of embedding features. Considering that the skeleton sequence is segment-wise embedded with length l, we extract the motion sequence $M^{\text{mask}} \in \mathbb{R}^{T_i \times V \times C_i}$ with stride $m = l$ and replicate padding according to Eq. (2) and Eq. (3), which is further reshaped into $M' \in \mathbb{R}^{T_e \times V \times l \times C_i}$ as is done for S'. Then, the motion intensity I, which indicates the motion significance of each spatio-temporal segment, is computed as follows:

$$I = \sum_{i=0}^{l} \sum_{j=0}^{C_i} |M'_{:,:,i,j}| \in \mathbb{R}^{T_e \times V}. \quad (4)$$

Since human actions are composed of a series of temporal movements, we argue that the intensity of motion largely reflects the semantic richness. Therefore in MAMP, the motion intensity is further converted into probability distribution with a temperature hyper-parameter τ:

$$\pi = \texttt{Softmax}(I/\tau), \quad (5)$$

which indicates the probability that each embedding feature is masked. In MAMP, the idea of gumble max is adapted for efficient probability-guided mask index sampling:

$$\begin{aligned} g &= -\log(-\log \epsilon), \ \epsilon \in U[0,1]^{T_e \times V}, \\ idx^{\text{mask}} &= \texttt{Index-of-Top-K}(\log \pi + g), \end{aligned} \quad (6)$$

where $U[0, 1]$ denotes uniform distribution between 0 and 1. The obtained idx^{mask} indicates which joints are masked and is used for unmasked token selection in Section 3.5. Based on the above operations, the network is encouraged to focus more on semantically rich regions, so as to learn more discriminative 3D action representation.

3.5. Masked Motion Prediction

We follow the encoder-decoder design in MAE [15], where the transformer encoder focuses on representation learning, while the decoder is responsible for the implementation of the pre-training pretext.

Encoder: In the encoder, separate spatio-temporal positional embedding $P_e^s \in \mathbb{R}^{1 \times V \times C_e}$ and $P_e^t \in \mathbb{R}^{T_e \times 1 \times C_e}$ are first element-wise added (with broadcasting) to the input joint embedding E:

$$E_p = E + P_e^s + P_e^t. \quad (7)$$

Then, the unmasked tokens in E_p are selected according to the idx^{mask} extracted in Eq. (6) and are flattened to $E_p^u \in \mathbb{R}^{N_u \times C_e}$, where $N_u = T_e \times V \times (1 - \text{mask ratio})$ is

the number of unmasked tokens. After that, the latent representation is extracted by L_e vanilla transformer blocks:

$$\begin{aligned} H_0 &= E_p^u, \\ H_l' &= \texttt{MSA}(\texttt{LN}(H_{l-1})) + H_{l-1}, \quad l \in 1, \cdots, L_e \\ H_l &= \texttt{MLP}(\texttt{LN}(H_l')) + H_l', \quad l \in 1, \cdots, L_e \\ H_e^u &= \texttt{LN}(H_{L_e}), \end{aligned} \tag{8}$$

where MSA, MLP, and LN denote multi-head self-attention, multilayer perceptron, and layer norm, respectively.

Decoder: In the decoder, the learnable mask tokens are inserted into H_e^u according to the mask indices idx^{mask}. The result is reshaped back to $H_e \in \mathbb{R}^{T_e \times V \times C_e}$, which is processed by L_d decoder layers for masked modeling:

$$\begin{aligned} Z_0 &= H_e + P_d^s + P_d^t, \\ Z_l' &= \texttt{MSA}(\texttt{LN}(Z_{l-1})) + Z_{l-1}, \quad l \in 1, \cdots, L_d \\ Z_l &= \texttt{MLP}(\texttt{LN}(Z_l')) + Z_l', \quad l \in 1, \cdots, L_d \\ Z_d &= \texttt{LN}(Z_{L_d}), \end{aligned} \tag{9}$$

where P_d^s and P_d^t are the spatial and temporal positional embedding of the transformer decoder, respectively.

Motion Prediction: In our method, the reconstruction target is not the original skeletons, but the motion sequence M^{target} pre-extracted according to Eq. (2) and Eq. (3), which is normalized by its segment-wise mean and standard deviation as in [15]. Therefore, given the decoded feature $Z_d \in \mathbb{R}^{T_e \times V \times C_d}$, we additionally adopt a prediction head to predict the temporal motion of masked human joints:

$$M^{\text{pred}} = \texttt{MotionPredHead}(Z_d), \tag{10}$$

where we empirically find that a simple fully connected layer just works well. For masked human joints, we compute the mean squared error (MSE) between the predicted result M^{pred} and the reconstruction target M^{target}:

$$\mathcal{L} = \frac{1}{|idx^{\text{mask}}|} \sum_{(i,j) \in idx^{\text{mask}}} \|(M_{i,j,:}^{\text{pred}} - M_{i,j,:}^{\text{target}})\|_2^2. \tag{11}$$

4. Experiments

4.1. Datasets and Evaluation Protocols

NTU-RGB+D 60 [36]: NTU-RGB+D 60 (NTU-60) is a large-scale dataset for human action recognition. It contains 60 action categories performed by 40 different subjects, with a total of 56,880 3D skeleton sequences. In this paper, we adopt the evaluation protocols recommended by the authors, namely cross-subject (X-sub) and cross-view (X-view). The former, X-sub uses the action sequences performed by half of the 40 subjects as training samples and the rest as test samples. For X-view, the training samples are action sequences captured by cameras 2 and 3, and the test samples are those captured by camera 1.

NTU-RGB+D 120 [26]: NTU-RGB+D 120 (NTU120) is an extended version of NTU-60, in which the number of action categories is increased from 60 to 120, the number of total skeleton sequences and subjects are also increased to 114,480 and 106, respectively. Furthermore, the authors also introduce a more challenging evaluation protocol named cross-setup (X-set) to substitute for the original X-view in NTU-60. Specifically, X-set divides sequences into 32 different setups based on the camera distance and background. Samples from half of these setups are used as the training set and the remainder constitute the test set.

PKU-MMD [8]: Following [46], to perform 3D action classification on PKU-MMD, we crop out action instances based on temporal annotations and divide them into training and test sets according to the cross-subject protocol. PKU-MMD contains two phases: PKU-MMD I (PKU-I) and PKU-MMD II (PKU-II). In PKU-I, the number of samples in training and test sets are 18,841 and 2,704, respectively. Due to the more noise introduced by the larger view variation, PKU-II is more challenging, with 5,332 samples for training and 1,613 for testing.

4.2. Experimental Setup

Network Architecture: In our MAMP framework, we adopt a vanilla vision transformer [11] as the backbone network, which consists of $L_e = 8$ identical building blocks. In each block, the embedding dimension is 256, the head number of the multi-head self-attention module is 8, and the hidden dimension of the feed-forward network is 1024. We employ learnable spatio-temporal separated positional embedding to the embedded inputs. The settings of the transformer decoder used during pre-training are consistent with those of the backbone encoder except that the number of layers L_d is reduced to 5.

Pre-training Details: During pre-training, the skeleton sequences are resized to a temporal length of 120 frames. The masking ratio of the input token is 90%. The target motion sequence M^{target} has stride $m = 1$ and is padded with zeros. We adopt the AdamW [31] optimizer with a weight decay of 0.05 and betas of (0.9, 0.95). We pre-train the network for 400 epochs with a batch size of 128. The learning rate is linearly increased to 1e-3 from 0 in the first 20 warm-up epochs and then decreased to 5e-4 by the cosine decay schedule. The experiments are conducted using the PyTorch framework on four NVIDIA RTX 3090 GPUs.

4.3. Comparison with State-of-the-art Methods

Linear Evaluation Results: In linear evaluation protocol, the pre-trained backbone is fixed and a post-attached linear classifier is trained with supervision for 100 epochs with a batch size of 256 and a learning rate of 0.1. The learning rate is decreased to 0 by the cosine decay schedule. As shown in Table 1, the performance on three datasets are re-

Method	Input stream	NTU-60		NTU-120		PKU-MMD	
		X-sub	X-view	X-sub	X-set	Phase I	Phase II
3s-SkeletonCLR [21]	Joint+Motion+Bone	75.0	79.8	60.7	62.6	85.3	-
3s-CrosSCLR [21]	Joint+Motion+Bone	77.8	83.4	67.9	66.7	84.9	21.2
3s-AimCLR [47]	Joint+Motion+Bone	78.9	83.8	68.2	68.8	87.4	39.5
LongT GAN [60]	Joint only	39.1	48.1	-	-	67.7	26.0
P&C [44]	Joint only	50.7	76.3	42.7	41.7	59.9	25.5
MS^{2}L [25]	Joint only	52.6	-	-	-	64.9	27.6
AS-CAL [35]	Joint only	58.5	64.8	48.6	49.2	-	-
ISC [46]	Joint only	76.3	85.2	67.1	67.9	80.9	36.0
GL-Transformer [18]	Joint only	76.3	83.8	66.0	68.7	-	-
CPM [57]	Joint only	78.7	84.9	68.7	69.6	88.8	48.3
CMD [32]	Joint only	79.4	86.9	70.3	71.5	-	43.0
SkeletonMAE* [51]	Joint only	74.8	77.7	72.5	73.5	82.8	36.1
MAMP (Ours)	Joint only	**84.9**	**89.1**	**78.6**	**79.1**	**92.2**	**53.8**

Table 1. Performance comparison on the NTU-60, NTU-120, and PKU-MMD datasets under the linear evaluation protocol. * indicates the re-implemented version under our framework, where improved performance is achieved.

Method	Backbone	NTU-60	
		X-sub	X-view
CPM [57]	ST-GCN	84.8	91.1
CrosSCLR [21]	3s-ST-GCN	86.2	92.5
AimCLR [47]	3s-ST-GCN	86.9	92.8
CrosSCLR [21]	STTFormer	84.6	90.5
AimCLR [47]	STTFormer	83.9	90.4
SkeletonMAE [51]	STTFormer	86.6	92.9
Colorization [56]	3s-DGCNN	88.0	94.9
MCC [45]	2s-AGCN	89.7	96.3
ViA [54]	2s-UINK	89.6	96.4
Hi-TRS [6]	3s-Transformer	90.0	95.7
W/o pre-training	Transformer	83.1	92.6
SkeletonMAE* [51]	Transformer	88.5	94.7
MAMP (Ours)	Transformer	**93.1**	**97.5**

Table 2. Performance comparison on the NTU-60 dataset under the fine-tuned evaluation protocol.

Method	Backbone	NTU-120	
		X-sub	X-set
CPM [57]	ST-GCN	78.4	78.9
CrosSCLR [21]	3s-ST-GCN	80.5	80.4
AimCLR [47]	3s-ST-GCN	80.1	80.9
CrosSCLR [21]	STTFormer	75.0	77.9
AimCLR [47]	STTFormer	74.6	77.2
SkeletonMAE [51]	STTFormer	76.8	79.1
MCC [45]	2s-AGCN	81.3	83.3
ViA [54]	2s-UINK	85.0	86.5
Hi-TRS [6]	3s-Transformer	85.3	87.4
W/o pre-training	Transformer	76.8	79.7
SkeletonMAE* [51]	Transformer	87.0	88.9
MAMP (Ours)	Transformer	**90.0**	**91.3**

Table 3. Performance comparison on the NTU-120 dataset under the fine-tuned evaluation protocol.

ported, they are NTU-60, NTU-120, and PKU-MMD, respectively. We include latest high-performance approaches for comparison, *e.g.*, GL-Transformer [18], CPM [57], CMD [32], 3s-CrosSLR [21], and 3s-AimCLR [47]. As we can see, with the joint stream as the only input, our proposed MAMP outperforms these methods on all the datasets. Specifically, MAMP outperforms previous state-of-the-art method CMD by 5.5% and 8.3% on the challenging NTU-60 x-sub and NTU-120 x-sub, respectively. For a fair comparison, we also re-implement the SkeletonMAE [51] under the same settings as our approach (denote as SkeletonMAE*), where improved performance is achieved. We can find that our MAMP significantly outperforms SkeletonMAE* on all six subsets of the three datasets, demonstrating the superiority of masked motion prediction compared to the self-reconstruction of joints.

Fine-tuned Evaluation Results: In fine-tuned evaluation protocol, an MLP head is attached to the pre-trained backbone and the whole network is fully fine-tuned for 100 epochs with a batch size of 48. The learning rate is linearly increased to 3e-4 from 0 in the first 5 warm-up epochs and then decreased to 1e-5 by the cosine decay schedule. We also adopt layer-wise lr decay [9] following [2]. As shown in Table 2 and Table 3, we evaluate the fine-tuned performance on NTU-60 and NTU-120, respectively. The vanilla transformer does not show satisfactory performance when trained directly from scratch, which is under expectation as weakly biased transformers require a large amount of training data to effectively prevent overfitting. After being pre-trained with the proposed MAMP framework, the network exhibits significant performance improvements ranging from 5% to 13% on the four subsets of the NTU-60 and NTU-120 datasets. The final results exceed all previ-

Method	NTU-60			
	X-sub		X-view	
	(1%)	(10%)	(1%)	(10%)
LongT GAN [60]	35.2	62.0	-	-
MS^{2}L [25]	33.1	65.1	-	-
ASSL [42]	-	64.3	-	69.8
ISC [46]	35.7	65.9	38.1	72.5
3s-CrosSCLR [21]	51.1	74.4	50.0	77.8
3s-Colorization [56]	48.3	71.7	52.5	78.9
CMD [32]	50.6	75.4	53.0	80.2
3s-Hi-TRS [6]	49.3	77.7	51.5	81.1
3s-AimCLR [47]	54.8	78.2	54.3	81.6
3s-CMD [32]	55.6	79.0	55.5	82.4
CPM [57]	56.7	73.0	57.5	77.1
W/o pre-training	38.8	70.8	40.4	76.0
SkeletonMAE* [51]	54.4	80.6	54.6	83.5
MAMP (Ours)	**66.0**	**88.0**	**68.7**	**91.5**

Table 4. Performance comparison on the NTU-60 dataset under the semi-supervised evaluation protocol. We report the average of five runs as the final performance.

Method	To PKU-II		
	NTU-60	NTU-120	PKU-I
LongT GAN [60]	44.8	-	43.6
MS^{2}L [25]	45.8	-	44.1
ISC [46]	51.1	52.3	45.1
CMD [32]	56.0	57.0	-
SkeletonMAE* [51]	58.4	61.0	62.5
MAMP (Ours)	**70.6**	**73.2**	**70.1**

Table 5. Performance comparison on the PKU-II dataset under the transfer learning evaluation protocol. The source datasets are the NTU-60, NTU-120, and PKU-I datasets.

ous methods, even those with multi-stream ensembling such as Colorization [56], MCC [45], and ViA [54]. Moreover, our MAMP also outperforms the re-implemented SkeletonMAE* by a considerable margin.

Semi-supervised Evaluation Results: Following previous works [21, 32, 46], in semi-supervised evaluation protocol, the post-attached classification layer and the pre-trained encoder are fine-tuned together with only a small fraction of the training set. Apart from that, we keep other training settings consistent with the fine-tuned evaluation protocol. As in [21, 47, 57], we report the performance on the NTU-60 dataset when using 1% and 10% of the training set. Note that considering the randomness during training data selection, we report the average of five runs as the final results. As shown in Table 4, our proposed MAMP significantly outperforms previous works like 3s-AimCLR [47], CPM [57], CMD [32], and SkeletonMAE* [51]. When using only 1% of the training data, MAMP outperforms SkeletonMAE* by 11.6% and 14.0% in X-sub and X-view respectively. Compared to training from scratch, MAMP pre-training brings performance improvements of more than 15.5% on all subsets of NTU-60.

Input	Target	NTU-60	NTU-120
Joint	Joint	74.8	72.5
	Motion	**84.9**	**78.6**
Motion	Joint	76.5	71.0
	Motion	75.9	70.5

Table 6. Ablation study on the superiority of masked motion prediction. The performance is evaluated on the NTU-60 X-sub and NTU-120 X-sub datasets under the linear evaluation protocol.

Strategy	NTU-60	NTU-120
Random masking	83.7	77.3
Motion-aware masking	**84.9**	**78.6**

Table 7. Ablation study on the mask sampling strategy. The performance is evaluated on the NTU-60 X-sub and NTU-120 X-sub datasets under the linear evaluation protocol.

Transfer Learning Evaluation Results: In transfer learning evaluation protocol, the network is pre-trained on a source dataset and then finetuned on a different target dataset. In this way, the generalizability of the learned representation is verified. In this paper, the target dataset is PKU-MMD II and the source datasets are NTU-60, NTU-120, and PKU-MMD I, respectively. Results in Table 5 show that, compared to previous methods, the representation learned by our proposed MAMP framework exhibits the best transferability, outperforming the reproduced SkeletonMAE* [51] by 12.2%, 12.2%, and 7.6% on the three source datasets, respectively.

4.4. Ablation Study

Superiority of Masked Motion Prediction: As shown in Table 6, to verify the superiority of masked motion prediction, we designed four different ablative experiments. Given the joint and motion streams of the original data, all possible choices of the model input and reconstruction target are traversed. We can find that our joint-to-motion prediction significantly outperforms other strategies. Under the linear evaluation protocol, our MAMP exceeds joint-to-joint prediction (adopted by [51]) by 10.1% and 6.1% on NTU-60 and NTU-120, respectively. This suggests that predicting dynamic motion from static skeletons during pre-training facilitates better contextual modeling of 3D human actions.

We also visualize the pre-training loss in Figure 3. Unlike the fast convergence of SkeletonMAE, masked motion prediction serves as a much harder pre-training objective.

Mask Sampling Strategy: In our approach, we employ the vanilla transformer as the backbone network, where embedding features at any spatio-temporal location can be freely masked as in MAE [15]. To verify the effectiveness of the

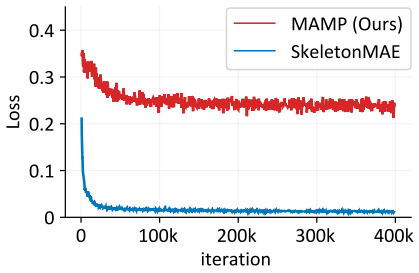


Figure 3. Pre-training loss plot. Compared to masked self-reconstruction of joints in SkeletonMAE, masked motion prediction acts as a harder pre-training objective.

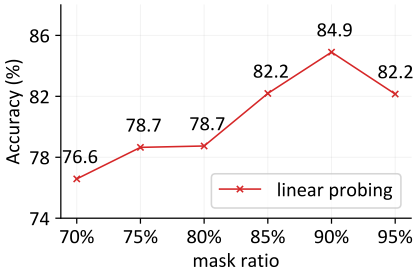


Figure 4. Ablation study on the masking ratio. The performance is evaluated on the NTU-60 X-sub dataset under the linear evaluation protocol.

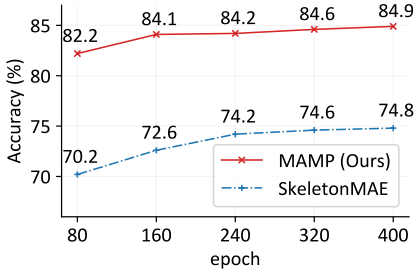


Figure 5. Pre-training schedule of SkeletonMAE and our MAMP. The performance is evaluated on the NTU-60 X-sub dataset under the linear evaluation protocol.

$T_e = T_s/l$	T_s	l	NTU-60	NTU-120
30	60	2	84.6	77.8
	120	4	**84.9**	**78.6**
	180	6	83.3	78.6
	240	8	84.5	77.9

Table 8. Ablation study on the segment length l used in the joint embedding process. For a fair comparison, the input length T_s is adjusted to ensure that the embedded features have a fixed length T_e. The performance is evaluated on the NTU-60 X-sub and NTU-120 X-sub datasets under the linear evaluation protocol.

L_d	NTU-60	NTU-120
2	83.3	77.2
3	**84.9**	77.6
4	84.6	77.5
5	**84.9**	**78.6**

(a) Decoder depth.

C_d	NTU-60	NTU-120
64	82.3	74.2
128	83.5	77.3
256	**84.9**	**78.6**
512	84.5	77.1

(b) Decoder width.

Table 9. Ablation study on the decoder design. The performance is evaluated on the NTU-60 X-sub and NTU-120 X-sub datasets under the linear evaluation protocol.

proposed motion-aware masking strategy, we compare its performance with that of random masking in Table 7. As we can see, our motion-aware masking strategy brings absolute performance improvements of 1.2% and 1.3% on NTU-60 and NTU-120, respectively. This suggests that the motion information, as an empirical semantic richness prior, can effectively guide the skeleton masking process.

Segment Length: We evaluated the performance of the learned representation under different segment lengths l used in the joint embedding process. For a fair comparison, we resize the original input sequence to ensure that the embedded features have a fixed length $T_e = T_s/l = 30$. As shown in Table 8, a segment length of 4 brings the best performance on both NTU-60 and NTU-120 datasets.

Decoder Design: We experimented with different numbers of layers and widths (feature dimensions) for the transformer decoder. As shown in Table 9 (a), our MAMP exhibits the best performance on the NTU-60 and NTU-120 datasets when the number of decoder layers is 3 and 5, respectively. The experimental results of the decoder width are in Table 9 (b), where a width of 256 brings the best performance. Overall, a decoder with 5 layers and a width of 256 is adopted in our MAMP framework by default.

Masking Ratio: As shown in Figure 4, we experimented with different masking ratios. Results on NTU-60 X-sub show that either too large or too small masking ratios have a negative impact on performance. We empirically found that a masking ratio of 90% exhibits the best results.

Pre-training Schedule: We studied the influence of the pre-training schedule length. As shown in Figure 5, both SkeletonMAE and our MAMP exhibit higher performance with longer pre-training schedules. It is worth mentioning that our MAMP significantly outperforms SkeletonMAE for all pre-training schedules with lengths ranging from 80 to 400 epochs, demonstrating the stability and superiority of the proposed masked motion prediction strategy.

5. Conclusion

In this work, we present MAMP, a simple yet effective framework for 3D action representation learning. We show that compared to conventional masked self-reconstruction of human joints, masked joint-to-motion prediction is demonstrated to be more effective for contextual motion modeling of 3D human actions. Given the high temporal redundancy of the skeleton sequence, we further devise the motion-aware masking strategy, which incorporates motion intensity as the empirical semantic richness prior for adaptative skeleton masking, facilitating better attention to semantically rich temporal regions. We conduct extensive experiments on three prevalent benchmarks under four evaluation protocols. Results show that the proposed MAMP brings remarkable performance improvements and sets a series of new state-of-the-art records, unleashing the tremendous potential of vanilla transformers for 3D action modeling.

References

[1] Dana H Ballard. Modular learning in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 647, pages 279–284, 1987. 3

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 3, 6

[3] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(01):172–186, 2021. 1

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 213–229, 2020. 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 (ICCV)*, pages 13359–13368, 2021. 1, 2

[6] Yuxiao Chen, Long Zhao, Jianbo Yuan, Yu Tian, Zhaoyang Xia, Shijie Geng, Ligong Han, and Dimitris N Metaxas. Hierarchically self-supervised transformer for human skeleton representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 185–202, 2022. 6, 7

[7] 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 (CVPR)*, pages 183–192, 2020. 1

[8] Liu Chunhui, Hu Yueyu, Li Yanghao, Song Sijie, and Liu Jiaying. PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017. 5

[9] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020. 6

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 3, 5

[12] Yong Du, Yun Fu, and Liang Wang. Skeleton based action recognition with convolutional neural network. In *Proceedings of the Asian Conference on Pattern Recognition (ACPR)*, pages 579–583, 2015. 2

[13] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015. 1, 2

[14] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2334–2343, 2017. 1

[15] 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 (CVPR)*, pages 16000–16009, 2022. 2, 3, 4, 5, 7

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016. 2

[17] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3D action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3288–3297, 2017. 1

[18] Boeun Kim, Hyung Jin Chang, Jungho Kim, and Jin Young Choi. Global-local motion transformer for unsupervised skeleton-based action learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 209–225, 2022. 6

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2012. 2

[20] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Skeleton-based action recognition with convolutional neural networks. In *Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 597–600. IEEE, 2017. 2

[21] 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 (CVPR)*, pages 4741–4750, 2021. 2, 3, 6, 7

[22] 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 (CVPR)*, pages 3595–3603, 2019. 1

[23] Tianjiao Li, Qiuhong Ke, Hossein Rahmani, Rui En Ho, Henghui Ding, and Jun Liu. Else-Net: Elastic semantic network for continual action recognition from skeleton data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13434–13443, 2021. 1

[24] Duohan Liang, Guoliang Fan, Guangfeng Lin, Wanjun Chen, Xiaorong Pan, and Hong Zhu. Three-stream convolutional neural network with multi-task and ensemble learning for 3D action recognition. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 934–940, 2019. 2

[25] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. MS2L: Multi-task self-supervised learning for skeleton based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, pages 2490–2498, 2020. 2, 3, 6, 7

[26] 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 (TPAMI)*, 42(10):2684–2701, 2020. 2, 5

[27] Jun Liu, Amir Shahroudy, Dong Xu, Alex C Kot, and Gang Wang. Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 40(12):3007–3021, 2017. 2

[28] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal LSTM with trust gates for 3D human action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 816–833. Springer, 2016. 2

[29] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 35(1):857–876, 2023. 2

[30] 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 (CVPR)*, pages 143–152, 2020. 1, 2

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 5

[32] Yunyao Mao, Wengang Zhou, Zhenbo Lu, Jiajun Deng, and Houqiang Li. CMD: Self-supervised 3D action representation learning with cross-modal mutual distillation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–752, 2022. 3, 6, 7

[33] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208:103219, 2021. 2, 3

[34] Helei Qiu, Biao Hou, Bo Ren, and Xiaohua Zhang. Spatio-temporal tuples transformer for skeleton-based action recognition. *arXiv preprint arXiv:2201.02849*, 2022. 2, 3

[35] 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. 6

[36] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016. 2, 5

[37] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7912–7921, 2019. 1, 2

[38] 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 (CVPR)*, pages 12026–12035, 2019. 2

[39] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 2, 3

[40] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. AdaSGN: Adapting joint number and model size for efficient skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13413–13422, 2021. 1

[41] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1227–1236, 2019. 1

[42] Chenyang Si, Xuecheng Nie, Wei Wang, Liang Wang, Tieniu Tan, and Jiashi Feng. Adversarial self-supervised learning for semi-supervised 3D action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2020. 7

[43] Yuxin Song, Min Yang, Wenhao Wu, Dongliang He, Fu Li, and Jingdong Wang. It takes two: Masked appearance-motion modeling for self-supervised video transformer pre-training. *arXiv preprint arXiv:2210.05234*, 2022. 3

[44] 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 (CVPR)*, pages 9631–9640, 2020. 2, 3, 6

[45] Yukun Su, Guosheng Lin, and Qingyao Wu. Self-supervised 3D skeleton action representation learning with motion consistency and continuity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13328–13338, October 2021. 6, 7

[46] Fida Mohammad Thoker, Hazel Doughty, and Cees GM Snoek. Skeleton-contrastive 3D action representation learning. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 1655–1663, 2021. 2, 3, 5, 6, 7

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

[48] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, pages 10347–10357, July 2021. 3

[49] Hongsong Wang and Liang Wang. Modeling temporal dynamics and spatial configurations of actions using

two-stream recurrent neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 499–508, 2017. 2

[50] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14668–14678, 2022. 3

[51] Wenhan Wu, Yilei Hua, Ce zheng, Shiqian Wu, Chen Chen, and Aidong Lu. SkeletonMAE: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition. *arXiv preprint arXiv:2209.02399*, 2022. 1, 3, 4, 6, 7

[52] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3D human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 899–908, 2020. 1

[53] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 7444–7452, 2018. 2

[54] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. ViA: View-invariant skeleton action representation learning via motion retargeting. *arXiv preprint arXiv:2209.00065*, 2022. 6, 7

[55] Haosen Yang, Deng Huang, Bin Wen, Jiannan Wu, Hongxun Yao, Yi Jiang, Xiatian Zhu, and Zehuan Yuan. Self-supervised video representation learning with motion-aware masked autoencoders. *arXiv preprint arXiv:2210.04154*, 2022. 3

[56] 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 (ICCV)*, pages 13423–13433, 2021. 2, 3, 6, 7

[57] Haoyuan Zhang, Yonghong Hou, Wenjing Zhang, and Wanqing Li. Contrastive positive mining for unsupervised 3D action representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 36–51, 2022. 3, 6, 7

[58] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1112–1121, 2020. 1

[59] Xikun Zhang, Chang Xu, and Dacheng Tao. Context aware graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14333–14342, 2020. 2

[60] 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 (AAAI)*, pages 2644–2651, 2018. 2, 3, 4, 6, 7