Rethinking Masked Data Reconstruction Pretraining for Strong 3D Action Representation Learning

Tao Gong[1, 2, 3], **Qi Chu**[1, 2, 3*], **Bin Liu**[1, 2, 3], **Nenghai Yu**[1, 2, 3]

[1]School of Cyber Science and Technology, University of Science and Technology of China
[2]Anhui Province Key Laboratory of Digital Security
[3]the CCCD Key Lab of Ministry of Culture and Tourism
{tgong, qchu, flowice, ynh}@ustc.edu.cn

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. For example, MAMP shows that instead of following the prevalent masked joint reconstruction, explicit masked motion reconstruction is key to the success of learning effective feature representation for 3D action recognition. However, we find that if we make a simple and effective change to the reconstructed target of masked joint reconstruction, masked joint reconstruction can achieve the same results as masked motion reconstruction. The devil is in the special characteristic of 3D skeleton data and the normalization process of training targets. We need to dig for all effective information of targets during normalization. Besides, considering that mask data reconstruction focuses more on learning local relations in input data for fulfilling the reconstruction task, instead of modeling the relation among samples, we further employ contrastive learning to learn more discriminative 3D action representations. We show that contrastive learning can consistently boost the performance of model pre-trained by masked joint prediction under various settings, especially in the semi-supervised setting that has a very limited number of labeled samples. Extensive experiments on NTU-60, NTU-120, and PKU-MMD datasets show that the proposed pre-training strategy achieves state-of-the-art results without bells and whistles.

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 (Cao et al. 2021; Fang et al. 2017; Xu et al. 2020), 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 (Chen et al. 2021; Cheng et al. 2020; Li et al. 2021b; Liu et al. 2020b; Shi et al. 2021; Zhang et al. 2020) 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. 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 representation learning, such as colorization (Yang et al. 2021), contrastive learning (Li et al. 2021a; Thoker, Doughty, and Snoek 2021; Tianyu et al. 2022), 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 over-reliance on heuristic action data augmentations (Liu et al. 2023), impeding its further exploration of 3D actions.

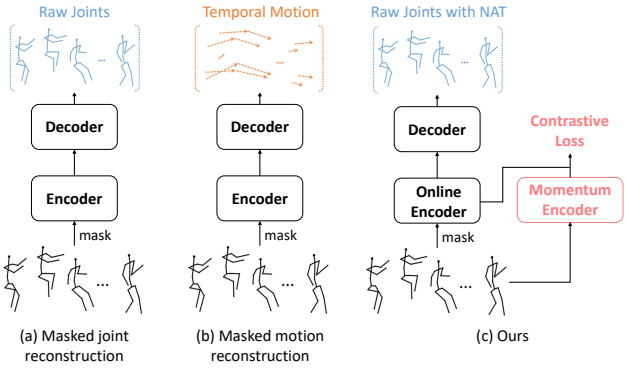

Figure 1: Illustration of pre-training objective comparison among prevalent masked joint reconstruction, masked motion reconstruction, and ours.

Recently, as transformers flourished in computer vision, masked autoencoder (MAE) (He et al. 2022) 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, such as SkeletonMAE (Wu et al. 2022), generally followed the practice

*Corresponding Author

of images, employing masked self-reconstruction of human joints (i.e. masked joint reconstruction) as the pre-training pretext. Subsequently, MAMP (Mao et al. 2023) shows that masked motion reconstruction results in significantly better performance compared to masked joint reconstruction.

However, in this paper, we find that masked joint reconstruction can achieve the same results as masked motion reconstruction if we apply a simple and effective change to the reconstructed target of masked joint reconstruction. To be specific, the devil is in the normalization of the reconstructed target. In the prevalent masked joint reconstruction and masked motion reconstruction, normalization by mean and standard deviation is usually applied among each temporal segment which contains multiple 3D coordinates (x, y, and z coordinates) in a small temporal window. For the reconstructed target in masked motion reconstruction, each value in the temporal segment is different. However, for the reconstructed target in masked joint reconstruction, each value among different timestamps is similar, while each value among different coordinate axes is different. Therefore, the reconstructed target in masked joint reconstruction shares similar values among different timestamps after normalization, which is the key to result in poor performance. We name this phenomenon in the mask joint reconstruction of 3D action representation learning as **"The Loss of Effective Reconstructed Target"**. To deal with the problem, We make a simple and effective change to the reconstructed target in masked joint reconstruction, i.e. **N**ormalization **A**mong **T**imestamps (**NAT**) for each coordinate axis. By doing so, the normalized reconstructed target in masked joint reconstruction is different in the temporal segment, just like masked motion reconstruction, and the performance of masked joint reconstruction with NAT is the same as masked motion reconstruction. In conclusion, the rationale is that we need to dig for all effective information of targets during normalization. We hope this finding could correct the previous not quite accurate conclusion in the 3D action representation learning community.

Masked data reconstruction focuses on learning local relations in the input sample for fulfilling the reconstruction task, instead of modeling the relation among different samples (Li et al. 2023). It is suspected that masked data reconstruction is less efficient in learning discriminative representations. This issue has been manifested by experimental results in (He et al. 2022; Xie et al. 2022). Therefore, can we leverage contrastive learning (Bachman, Hjelm, and Buchwalter 2019) to further strengthen the 3D action representation learned by masked data reconstruction methods? To answer this question, we aim to explore a possible way to boost masked data reconstruction with contrastive learning in a unified framework. An overview of the proposed method is shown in Figure 1. Specifically, our method adopts a siamese architecture (Bromley et al. 1993). One branch is an online updated asymmetric encoder-decoder that learns latent representations to reconstruct masked joints from a few visible segments. The other branch is a momentum encoder that provides contrastive learning supervision. Different from online encoder whose inputs only contain the visible segments, the momentum encoder is fed with the full set of 3D skeleton data. This design ensures the semantic integrity of its output features to guide the online encoder. With the above novel designs, the online encoder of our method can learn more discriminative features of holistic information and achieve state-of-the-art performance in various settings, especially in the semi-supervised setting that has a very limited number of labeled samples.

Our contributions are summarized as follows:

1. We find that the key to increase the performance of masked joint reconstruction lies in the normalization process of the reconstruction target. By applying a simple but effective change to the normalization process, we show that masked joint reconstruction can achieve the same results as masked motion reconstruction.
2. We improve the representation of masked joint reconstruction by using contrastive learning. Its learned representations not only preserve the local context-sensitive features but also model the instance discriminativeness among different 3D skeleton data.
3. We conduct extensive experiments on three prevalent benchmarks to verify the effectiveness of our method, and show that our method achieves state-of-the-art results without bells and whistles for 3D action recognition.

Related Work

Supervised 3D Action Recognition

In recent years, deep learning has been widely used in skeleton action recognition fields due to its powerful ability of feature extraction and representation learning. RNN-based methods (e.g., LSTMs) (Du, Wang, and Wang 2015; Liu et al. 2017) were widely utilized to process skeleton data. Nevertheless, the data representations extracted by RNNs were too simple to present the comprehensive spatial-temporal features of skeleton data. Thus, GCN-based methods (Zhang et al. 2020; Cheng et al. 2020) were naturally introduced to model the topological graph features from the skeleton data. Recently, with the success of the vision transformer (ViT) (Dosovitskiy et al. 2020), the transformer-based model has become a powerful architecture for skeleton data analysis (Qiu et al. 2022; Chen et al. 2022) due to the ability to learn global representations.

Self-supervised 3D Action Recognition

Many previous works have been proposed to perform self-supervised 3D action representation learning. In LongT GAN (Zheng et al. 2018), an autoencoder-based model along with an additional adversarial training strategy are proposed. Similarly, P&C (Su, Liu, and Shlizerman 2020) trains an encoder-decoder network to both predict and cluster skeleton sequences. MS2L (Lin et al. 2020) integrates multiple pretext tasks to learn better representation. Recently, many contrastive learning-based approaches (Lin et al. 2020; Thoker, Doughty, and Snoek 2021; Li et al. 2021a; Tianyu et al. 2022; Zhang et al. 2022; Mao et al. 2022) have emerged, showing superior performance compared to earlier works. To learn better 3D action representation, they either try to dig for helpful supervision across

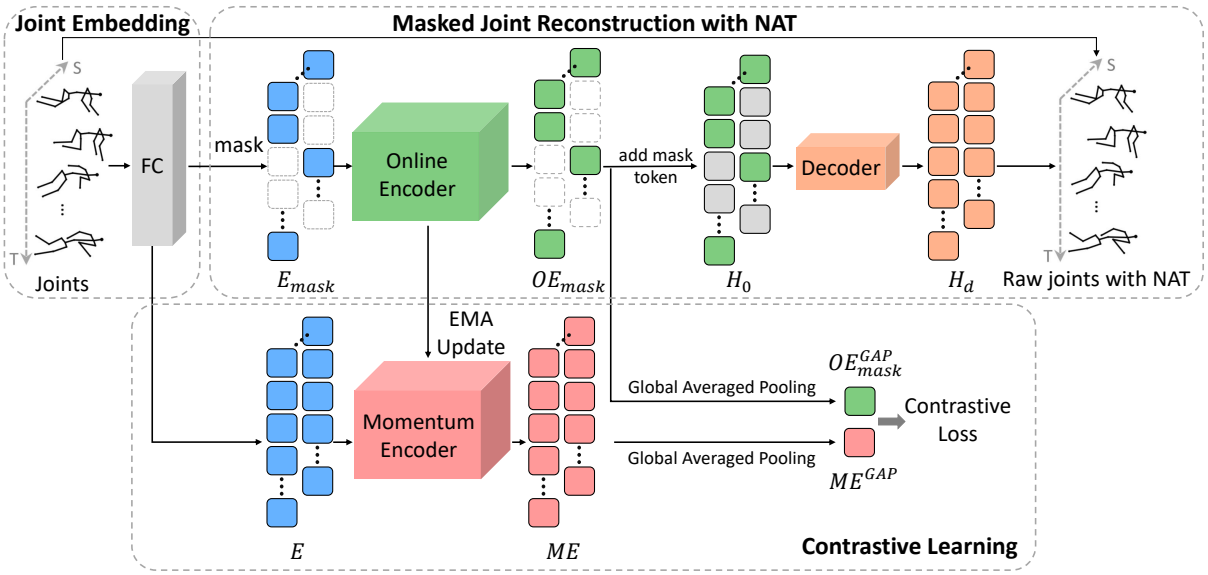

Figure 2: The overall pipeline of our method.

different skeleton modalities (Li et al. 2021a; Mao et al. 2022), or explore better action data augmentation (Tianyu et al. 2022) and positive sample mining strategies (Zhang et al. 2022). Nevertheless, as a global feature learner originally designed for images, the over-reliance on heuristic action data augmentations in contrastive learning limits its further development for 3D actions. SkeletonMAE (Wu et al. 2022) first introduces the idea of MAE (He et al. 2022) into transformer-based 3D action representation learning, where the original joint coordinates of masked regions are predicted. MAMP (Mao et al. 2023) shows that masked motion reconstruction results in significantly better performance compared to raw joint reconstruction. In this paper, we show that masked joint reconstruction can achieve the same results as masked motion reconstruction.

Constrastive Learning

Contrastive learning (Chen et al. 2020a; He et al. 2020; Huang et al. 2023) aims to learn feature representation via instance discrimination. It pulls positive pairs closer and pushes negative pairs away. Since no labels are available during self-supervised contrastive learning, two different augmented versions of the same sample are treated as a positive pair, and samples from different instances are considered to be negative. In MoCo (He et al. 2020) and MoCo v2 (Chen et al. 2020b), the negative samples are taken from previous batches and stored in a queue-based memory bank. In contrast, SimCLR (Chen et al. 2020a) and MoCo v3 (Chen, Xie, and He 2021) rely on a larger batch to provide enough negative samples. This paper explores boosting masked data prediction with contrastive learning in a unified framework to learn more discriminative 3D action representation.

Method

Framework Overview

Figure 2 illustrates the overall pipeline of our framework. It takes a skeleton sequence $S \in R^{T_s \times V \times C_s}$ as input, which is randomly cropped from the original data and is resized to a fixed temporal length T_s. V denotes the number of joints. $C_s = 3$ denotes the x, y, z coordinate, respectively.

As in most vision transformers, the input joints are linearly mapped into joint embedding E. After that, we use two branches where the online branch is an asymmetric encoder-decoder and the momentum branch is a momentum-updated encoder. For the online branch, we mask most of the embedding features. The remaining features E_{mask} are processed by the encoder-decoder architecture, where the online transformer encoder learns representation OE_{mask} from unmasked joint embedding E_{mask} and the transformer decoder performs masked joint prediction based on the features H_0 which consist of learnable mask tokens and latent representation OE_{mask} from the transformer online encoder. For the momentum branch, the whole joint embedding E is taken as input to reserve the semantic integrity and the discriminativeness of the learned representations. The momentum encoder outputs the whole latent joint representation ME. Different from the online encoder, we update the parameters of the momentum encoder by exponential moving average (EMA). We perform global averaged pooling (GAP) to the unmasked latent features OE_{mask} from the online encoder and the latent features ME from the momentum encoder to get features OE^{GAP}_{mask} and ME^{GAP}, respectively, and then utilize contrastive learning to model the instance discrimination among different 3D skeleton data.

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

Joint Embedding

Following MAMP (Mao et al. 2023), the input skeleton sequence $S \in R^{T_s \times V \times C_s}$ is divided into temporally non-overlapping segments $S' \in R^{T_e \times V \times (l \times C_s)}$, where l is the temporal length of each segment and $T_e = T_s/l$. Joints of the same segment are embedded together:

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

where C_e is the dimension of the embedding features and FC denotes a linear layer.

Masked Joint Reconstruction with NAT

We follow the encoder-decoder design in MAE (He et al. 2022) and MAMP (Mao et al. 2023), where the online transformer encoder focuses on representation learning, while the decoder is responsible for the implementation of the pre-training pretext.

Online Encoder: We mask most of the embedding features to get the unmasked joint embedding E_{mask}, and flat it to $E^u_{mask} \in R^{N_u \times C_e}$, where $N_u = T_e \times V \times$(1 - mask ratio) is the number of unmasked tokens. After that, the latent representation OE_{mask} is extracted by N_e vanilla transformer blocks:

$$\begin{aligned} G_0 =& E^u_{mask}, \\ G'_n =& MSA(LN(G_{n-1})) + G_{n-1}, \quad n \in 1, ..., N_e \\ G_n =& MLP(LN(G'_n)), \quad n \in 1, ..., N_e \\ OE_{mask} =& LN(G_{N_e}), \end{aligned} \tag{2}$$

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

Decoder: In the decoder, the learnable mask tokens are inserted into OE_{mask}. The result is reshaped back to $H_0 \in R^{T_e \times V \times C_e}$, which is processed by N_d decoder layers for masked modeling:

$$\begin{aligned} H'_n =& MSA(LN(H_{n-1})) + H_{n-1}, \quad n \in 1, ..., N_d \\ H_n =& MLP(LN(H'_n)), \quad n \in 1, ..., N_d \\ H_d =& LN(H_{N_d}), \end{aligned} \tag{3}$$

where $H_d \in R^{T_e \times V \times C_d}$ denotes the decoded feature which is used to reconstruct masked joints, and C_d is the dimension of features H_d.

Joints Prediction with NAT: Previous work (Mao et al. 2023) shows that compared with masked joint reconstruction, masked motion reconstruction is key to the success of learning effective feature representation for 3D action recognition. However, we show that if we make a simple and effective change to the reconstructed target of masked joint reconstruction, masked joint reconstruction can achieve the same results as masked motion reconstruction.

Given the decoded feature H_d, we additionally adopt a prediction head (a simple linear layer) to predict the masked human joints $O^{pred} \in R^{T_e \times V \times (l \times C_s)}$. The target of O^{pred} is the the input skeleton sequence $S' \in R^{T_e \times V \times (l \times C_s)}$. S' is usually normalized by its segment-wise mean and standard deviation as in (He et al. 2022):

$$\begin{aligned} \hat{S}'_{t,v} =& \frac{S'_{t,v} - mean(S'_{t,v})}{\sqrt{var(S'_{t,v})}}, \quad \hat{S}'_{t,v} \in R^{l \times C_s} \\ & t = 1, ..., T_e, \;\; v = 1, ..., V \end{aligned} \tag{4}$$

where $mean(S'_{t,v})$ denotes the mean of $S'_{t,v}$ and $var(S'_{t,v})$ denotes the variance of $S'_{t,v}$. We now analyze the rational of the poor performance of using $\hat{S}'_{t,v}$ as target.

$S'_{t,v}$ contains multiple 3D coordinates (x, y, z coordinates) in a continuous and short temporal window, Therefore, we rewrite $S'_{t,v}$ to $S'_{t,v} = \{(x_i, y_i, z_i) | i = 1, ..., l\}$ and $\hat{S}'_{t,v}$ to $\hat{S}'_{t,v} = \{(\hat{x}_i, \hat{y}_i, \hat{z}_i) | i = 1, ..., l\}$ for convenience. The values of x_i, y_i, and z_i for a given i are usually quite different, since a human joint can locate at any position in the 3D space. However, the values among x_i where $i = 1, .., l$ are quite similar, since a human joint can only move a small distance in a continuous and short temporal window. After normalizing by equation 4, the values of $\hat{x}_i$, $\hat{y}_i$, and $\hat{z}_i$ are quite different, while the values among $\hat{x}_i$ where $i = 1, .., l$ are still quite similar. Therefore, the reconstructed normalization target in masked joint reconstruction shares similar values among different timestamps, which is the key to result in poor performance. We name this phenomenon in the mask joint reconstruction of 3D action representation learning as **"The Loss of Effective Reconstructed Target"**.

To deal with the problem, considering the characteristic of 3D skeleton data, we make a simple and effective change to the reconstructed target in masked joint reconstruction, i.e. **N**ormalization **A**mong **T**imestamps (**NAT**) for each coordinate axis to get the normalized target $\bar{S}'$:

$$\begin{aligned} \bar{S}'_{t,v,k} =& \frac{S'_{t,v,k} - mean(S'_{t,v,k})}{\sqrt{var(S'_{t,v,k})}}, \quad \bar{S}'_{t,v,k} \in R^{l} \\ & t = 1, ..., T_e, \;\; v = 1, ..., V, \;\; k = x, y, z \end{aligned} \tag{5}$$

After normalizing by equation 5, the values of the reconstructed normalization target in masked joint reconstruction are quite different in both different timestamps and different coordinate axes. In the experiments section, we show that masked joint reconstruction with NAT can achieve the same results as masked motion reconstruction.

This finding also holds when using masked motion reconstruction as the target (Mao et al. 2023). The normalization process of masked motion reconstruction is as follows:

$$\begin{aligned} M'_{t,v,i} =& S'_{t,v,i+1} - S'_{t,v,i}, \quad M'_{t,v,i} \in R^{C_s}, \\ \hat{M}'_{t,v} =& \frac{M'_{t,v} - mean(M'_{t,v})}{\sqrt{var(M'_{t,v})}}, \quad M'_{t,v} \in R^{l \times C_s} \\ & t = 1, ..., T_e, \;\; v = 1, ..., V, \;\; i = 1, ..., l \end{aligned} \tag{6}$$

By calculating the first derivative of S' along the temporal dimension (i.e., the motion information), the values of normalized motion $\hat{M}'_{t,v}$ are also quite different in both different timestamps and different coordinate axes.

In conclusion, the rationale is not to hide the information contained in the reconstructed target. We hope this finding could correct the previous not quite accurate conclusion in the 3D action representation learning community and draw the attention of researchers in this field to the special characteristics of 3D skeleton data.

Finally, we compute the mean squared error (MSE) between the predicted result O^{pred} and the reconstruction target $\bar{S}'$ from equation 5:

$$L_r = \frac{1}{N_{mask}} \sum_{t,v} (O_{t,v}^{pred} - \bar{S}'_{t,v})^2 \tag{7}$$

where the loss is only calculated in the masked tokens and N_{mask} denotes the number of masked tokens.

Contrastive Learning

Masked data reconstruction focuses on learning local relations in the input sample for fulfilling the reconstruction task, instead of modeling the relation among different samples (Li et al. 2023). To learn more about discriminative 3D action representation, we explore a possible way to boost masked data reconstruction with contrastive learning in a unified framework.

Momentum Encoder: The momentum encoder is introduced to provide contrastive supervision for the online encoder to learn discriminative representations. It shares the same architecture as the online encoder, but takes the whole 3D skeleton data as input, in order to reserve the semantic integrity and the discriminativeness of the learned representations. We perform GAP to the features ME from the momentum encoder to get the global representations ME^{GAP} for contrastive learning. Similarly, we get the global features OE_{mask}^{GAP} from the unmasked latent features OE_{mask} from the online encoder. Different from the online encoder, we update the parameters of the momentum encoder by exponential moving average (EMA). The momentum hyper-parameter for updating the momentum encoder is denoted as μ. Momentum update is used since it stabilizes the training by fostering smooth feature changes, as found in MoCo (He et al. 2020).

Constrastive Loss: We use InfoNCE (He et al. 2020; Chen et al. 2020a) loss as contrastive loss. InfoNCE loss seeks to simultaneously pull close positive views from the same sample and push away negative samples. We then compute the cosine similarity ρ between OE_{mask}^{GAP} and ME^{GAP}:

$$\rho = \frac{OE_{mask}^{GAP} \cdot ME^{GAP}}{||OE_{mask}^{GAP}||_2 ||ME^{GAP}||_2} \tag{8}$$

We denote ρ^+ as the positive pairs cosine similarity, which is constructed by OE_{mask}^{GAP} and ME^{GAP} from the same 3D skeleton sample. ρ_j^- indicates the cosine similarity for the j-th negative pair. Due to the limitation of GPU memory, instead of using other 3D skeleton features ME^{GAP} in the same large batch to construct negative pairs, we use 3D skeleton features ME^{GAP} stored in queue Q to construct negative pairs. The queue Q is defined on the fly by a set of features ME^{GAP} from 3D skeleton samples, with the current batch enqueued and the oldest batch dequeued, during training. The loss function of InfoNCE loss is:

$$L_c = -log \frac{\exp(\rho^+/\tau)}{\exp(\rho^+/\tau) + \sum_{j=1}^{N_Q} \exp(\rho_j^-/\tau)} \tag{9}$$

where τ is the temperature constant, which is set to 0.07. N_Q is the number of features stored in queue Q.

In the experiments, we show that contrastive learning can consistently boost the performance of the model in various settings, especially in the semi-supervised setting that has a very limited number of labeled samples.

Training Loss

The overall training loss is a weighted combination of reconstruction loss L_r and contrastive loss L_c defined as:

$$L = L_r + \lambda L_c \tag{10}$$

where λ is used to balance the loss between L_r and L_c.

Experiments

Datasets and Implementation Details

Following previous works (Mao et al. 2022, 2023; Shah et al. 2023), in this paper, we adopt the evaluation protocols cross-subject (X-sub) and cross-view (X-view) for NTU-RGB+D 60 (Shahroudy et al. 2016) dataset, cross-subject (X-sub) and cross-setup (X-set) for NTU-RGB+D 120 (Liu et al. 2020a) dataset. We also evaluate our method on the PKU-MMD II (Chunhui et al. 2017) (PKU-II) phase. In all experiments, we report the top-1 accuracy following previous works (Mao et al. 2023; Lin, Zhang, and Liu 2023).

The momentum hyper-parameter μ, the number of features N_Q, and the loss weight λ are set to 0.999, 65536, and 0.0001, respectively. Other hyper-parameters follow MAMP (Mao et al. 2023)

Main Results

Linear Evaluation Results. Following MAMP (Mao et al. 2023), 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 NTU-60, NTU-120, and PKU-MMD datasets are reported. We include the latest high-performance approaches for comparison, *e.g.*, 3s-ActCLR (Lin, Zhang, and Liu 2023), HaLP (Shah et al. 2023), and MAMP (Mao et al. 2023). As we can see, with the joint stream as the only input, our method outperforms these methods on all datasets. Specifically, our method outperforms previous state-of-the-art method MAMP by 2.0% and 2.4% on the challenging NTU-60 x-sub and NTU-120 x-set, respectively.

Fine-tuned Evaluation Results. Following MAMP (Mao et al. 2023), in finetuned 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

Method	Input stream	NTU-60		NTU-120		PKU-MMD
		X-sub	X-view	X-sub	X-set	Phase II
3s-SkeletonCLR (Li et al. 2021a)	Joint+Motion+Bone	75.0	79.8	60.7	62.6	-
3s-CrosSCLR (Li et al. 2021a)	Joint+Motion+Bone	77.8	83.4	67.9	66.7	21.2
3s-AimCLR (Tianyu et al. 2022)	Joint+Motion+Bone	78.9	83.8	68.2	68.8	39.5
3s-RVTCLR+ (Zhu et al. 2023)	Joint+Motion+Bone	79.7	84.6	68.0	68.9	-
3s-ActCLR (Lin, Zhang, and Liu 2023)	Joint+Motion+Bone	84.3	88.8	74.3	75.7	-
AS-CAL (Rao et al. 2021)	Joint only	58.5	64.8	48.6	49.2	-
ISC (Thoker, Doughty, and Snoek 2021)	Joint only	76.3	85.2	67.1	67.9	36.0
GL-Transformer (Kim et al. 2022)	Joint only	76.3	83.8	66.0	68.7	-
CPM (Zhang et al. 2022)	Joint only	78.7	84.9	68.7	69.6	48.3
CMD (Mao et al. 2022)	Joint only	79.4	86.9	70.3	71.5	43.0
HaLP (Shah et al. 2023)	Joint only	79.7	86.8	71.1	72.2	43.5
ActCLR (Lin, Zhang, and Liu 2023)	Joint only	80.9	86.7	69.0	70.5	-
SkeletonMAE (Wu et al. 2022)	Joint only	74.8	77.7	72.5	73.5	36.1
MAMP (Mao et al. 2023)	Joint only	84.9	89.1	78.6	79.1	53.8
Ours	Joint only	**86.9**	**91.0**	**80.0**	**81.5**	**55.3**

Table 1: Performance comparison on the NTU-60, NTU-120, and PKU-MMD datasets under linear evaluation protocol.

Method	NTU-60		NTU-120	
	X-sub	X-view	X-sub	X-set
CTR-CGN (Chen et al. 2021)	92.4	96.8	88.9	90.6
PoseC3D (Duan et al. 2022)	93.7	96.6	86.0	89.6
InfoGCN (Chi et al. 2022)	93.0	97.1	89.8	91.2
STC-Net (Lee et al. 2023a)	93.3	97.3	90.2	91.7
HD-GCN (Lee et al. 2023b)	93.4	97.2	90.1	91.6
GAP (Xiang et al. 2023)	92.9	97.0	89.9	91.1
SkeleTR (Duan et al. 2023)	94.8	97.7	87.8	88.3
BlockGCN (Zhou et al. 2024)	90.9	95.4	86.9	88.2
LLM-AR (Qu, Cai, and Liu 2024)	**95.0**	98.4	88.7	91.5
SkateFormer (Do and Kim 2024)	93.5	97.8	89.8	91.4
Ours	93.9	**98.5**	**90.4**	**92.6**

Table 2: Performance comparison with fully-supervised methods on the NTU-60 and NTU-120 datasets. Note that our method does not perform multi-stream ensembling during evaluation.

Method	NTU-60			
	X-sub		X-view	
	(1%)	(10%)	(1%)	(10%)
3s-CrosSCLR (Li et al. 2021a)	51.1	74.4	50.0	77.8
3s-Colorization (Yang et al. 2021)	48.3	71.7	52.5	78.9
3s-Hi-TRS (Chen et al. 2022)	49.3	77.7	51.5	81.1
3s-AimCLR (Tianyu et al. 2022)	54.8	78.2	54.3	81.6
3s-CMD (Mao et al. 2022)	55.6	79.0	55.5	82.4
CPM (Zhang et al. 2022)	56.7	73.0	57.5	77.1
3s-RVTCLR+ (Zhu et al. 2023)	54.9	79.5	53.6	83.7
SkeletonMAE (Wu et al. 2022)	54.4	80.6	54.6	83.5
MAMP (Mao et al. 2023)	66.0	88.0	68.7	91.5
Ours w/o Con	66.2	88.3	68.8	91.8
Ours	**78.2**	**90.4**	**81.3**	**94.5**

Table 3: Performance comparison on the NTU-60 dataset under semi-supervised evaluation protocol. "Ours w/o Con" denotes not using contrastive learning to learn discriminative representation.

the cosine decay schedule. We also adopt layer-wise lr decay (Clark et al. 2020) following (Bao et al. 2022). We compare our method with the top-performing supervised methods like SkeleTR (Duan et al. 2023), SkateFormer (Do and Kim 2024), and LLM-AR (Qu, Cai, and Liu 2024) in Table 2. Results show that without ensembling, our method outperforms most top-performing methods, especially on larger NTU-120 dataset.

Semi-supervised Evaluation Results. Following previous works (Li et al. 2021a; Thoker, Doughty, and Snoek 2021), in the 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. We also keep other training settings consistent with the fine-tuned evaluation protocol. As in (Tianyu et al. 2022; Mao et al. 2023), we report the performance on the NTU-60 dataset when using 1% and 10% of the training set. Considering the randomness during training data selection, we report the average of five runs as the final results. As shown in Table 3, Our method significantly outperforms previous works like 3s-CMD (Mao et al. 2022), 3s-RVTCLR+ (Zhu et al. 2023) and MAMP (Mao et al. 2023). When using only 1% of the training data, our method significantly outperforms previous state-of-the-art method MAMP by 12.2% and 12.6% in X-sub and X-view, respectively.

Transfer Learning Evaluation Results. In the transfer learning evaluation protocol, the network is pre-trained on

Method	To PKU-II	
	NTU-60	NTU-120
ISC (Thoker, Doughty, and Snoek 2021)	51.1	52.3
CMD (Mao et al. 2022)	56.0	57.0
3s-ActCLR (Lin, Zhang, and Liu 2023)	55.9	-
HaLP (Shah et al. 2023)	54.8	55.4
SkeletonMAE (Wu et al. 2022)	58.4	61.0
MAMP (Mao et al. 2023)	70.6	73.2
Ours w/o Con	70.7	73.2
Ours	**72.6**	**73.7**

Table 4: Performance comparison on the PKU-II dataset under transfer learning evaluation protocol. The source datasets are the NTU-60 and NTU-120 datasets. "Ours w/o" denotes not using contrastive learning.

Input	Target	NTU-60	NTU-120
Joint	Joint	74.8	72.5
	Motion	84.9	78.6
	Joint w NAT	**85.1**	**78.8**

Table 5: The effectiveness of NAT in masked joint reconstruction. Contrastive learning is not used for all experiments. The results are evaluated on NTU-60 X-sub and NTU-120 X-sub datasets under linear evaluation protocol.

method	NTU-60	NTU-120
Ours w/o Con	85.1	78.8
Ours	**86.9**	**80.0**

Table 6: The effectiveness of contrastive learning. "Ours w/o Con" denotes not using contrastive learning. The performance is evaluated on the NTU-60 X-sub and NTU-120 X-sub datasets under linear evaluation protocol.

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 and NTU-120, respectively. Results in Table 4 show that, compared to previous methods, the representation learned by our method exhibits the best transferability, outperforming the previous state-of-the-art method MAMP (Mao et al. 2023) by 2.0% and 0.5% on two source datasets, respectively.

Ablation Study

NAT in Masked Joints Reconstruction. As shown in Table 5, when using NAT for the joint target, the performance significantly improves by 10.3 % and 6.3 % on NTU-60 and NTU-120 datasets. This shows the effectiveness of the proposed NAT. Table 5 also shows that masked joint reconstruction with NAT achieves the same results as masked motion reconstruction on NTU-60 and NTU-120 datasets.

Contrastive learning. The semi-supervised setting with only 1% labeled samples for training needs that the pre-trained network can extract powerful and discriminative representation. As shown in Table 3, compared with "Ours w/o Con" and "Ours", the network with contrastive learning exhibits significant performance improvements (around 12 points) under the very limited (1%) amount of supervised training dataset. This shows that contrastive learning can improve feature discriminability effectively. With only a few labeled samples, the network pre-trained with contrastive learning can achieve good results in the downstream classification tasks.

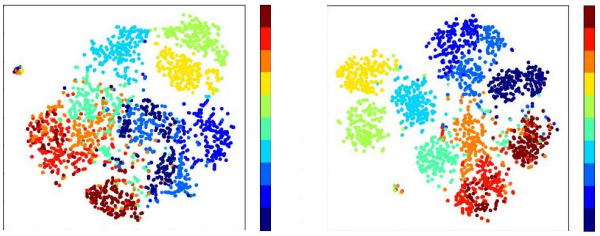

Figure 3: The t-SNE visualization of embeddings. "Ours w/o Con" denotes not using contrastive learning.

Table 6 also shows that using contrastive learning can improve the performance by 1.8 % and 1.2 % on NTU-60 and NTU-120 datasets under the linear evaluation protocol. This further demonstrates that even though there are a lot of training samples, using contrastive learning can consistently boost the performance of the network.

We also apply t-SNE (Van der Maaten and Hinton 2008) to show the embedding distribution of "Ours w/o Con" and "Ours" on the NTU-60 X-sub benchmark. For a fair comparison, we select the same 10 classes for visualization, where a dot of the same color represents the same class. From Figure 3, we can see that the embeddings extracted from our method have better inter-class separability and intra-class compactness, further indicating that our contrastive learning can learn more discriminative features.

Conclusion

In this paper, we rethink the masked data reconstruction pretraining for 3D action representation learning. We conclude with two points that could benefit the community of this research field. (1) Masked joint reconstruction with the proposed Normalization Among Timestamps (NAT) can achieve the same results as masked motion reconstruction. The devil is in the special characteristic of 3D skeleton data and the normalization process of training targets. We need to dig for all effective information of targets during normalization. (2) Contrastive learning can help the model pre-trained by masked data prediction learn more discriminative 3D action representations, which can benefit the performance of the model, especially on the downstream classification tasks with a very limited number of labeled samples. We hope our findings will benefit the researchers of the 3D action representation learning community.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62121002, No. U20B2047) and Anhui Provincial Science and Technology Major Project (No. 2023z020006).

References

Bachman, P.; Hjelm, R. D.; and Buchwalter, W. 2019. Learning representations by maximizing mutual information across views. *Advances in Neural Information Processing Systems*, 32.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2022. BEiT: BERT Pre-Training of Image Transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; and Shah, R. 1993. Signature verification using a" siamese" time delay neural network. *Advances in Neural Information Processing Systems*, 6.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.; and Sheikh, Y. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(01): 172–186.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 1597–1607.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. In *ICCV*, 9640–9649.

Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; and Hu, W. 2021. Channel-Wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. In *ICCV*, 13359–13368.

Chen, Y.; Zhao, L.; Yuan, J.; Tian, Y.; Xia, Z.; Geng, S.; Han, L.; and Metaxas, D. N. 2022. Hierarchically Self-supervised Transformer for Human Skeleton Representation Learning. In *ECCV*, 185–202.

Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; and Lu, H. 2020. Skeleton-based action recognition with shift graph convolutional network. In *CVPR*, 183–192.

Chi, H.-g.; et al. 2022. InfoGCN: Representation learning for human skeleton-based action recognition. In *CVPR*.

Chunhui, L.; Yueyu, H.; Yanghao, L.; Sijie, S.; and Jiaying, L. 2017. PKU-MMD: A Large Scale Benchmark for Continuous Multi-Modal Human Action Understanding. *arXiv preprint arXiv:1703.07475*.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR*.

Do, J.; and Kim, M. 2024. SkateFormer: Skeletal-Temporal Transformer for Human Action Recognition. In *ECCV*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Du, Y.; Wang, W.; and Wang, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 1110–1118.

Duan, H.; Xu, M.; Shuai, B.; Modolo, D.; Tu, Z.; Tighe, J.; and Bergamo, A. 2023. Skeletr: Towards skeleton-based action recognition in the wild. In *ICCV*, 13634–13644.

Duan, H.; et al. 2022. Revisiting skeleton-based action recognition. In *CVPR*.

Fang, H.-S.; Xie, S.; Tai, Y.-W.; and Lu, C. 2017. RMPE: Regional multi-person pose estimation. In *ICCV*, 2334–2343.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *CVPR*, 16000–16009.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 9729–9738.

Huang, Z.; Jin, X.; Lu, C.; Hou, Q.; Cheng, M.-M.; Fu, D.; Shen, X.; and Feng, J. 2023. Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kim, B.; Chang, H. J.; Kim, J.; and Choi, J. Y. 2022. Global-Local Motion Transformer for Unsupervised Skeleton-Based Action Learning. In *ECCV*, 209–225.

Lee, J.; Lee, M.; Cho, S.; Woo, S.; Jang, S.; and Lee, S. 2023a. Leveraging spatio-temporal dependency for skeleton-based action recognition. In *ICCV*, 10255–10264.

Lee, J.; Lee, M.; Lee, D.; and Lee, S. 2023b. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *ICCV*, 10444–10453.

Li, L.; Wang, M.; Ni, B.; Wang, H.; Yang, J.; and Zhang, W. 2021a. 3D human action representation learning via cross-view consistency pursuit. In *CVPR*, 4741–4750.

Li, S.; Wu, D.; Wu, F.; Zang, Z.; and Li, S. Z. 2023. Architecture-agnostic masked image modeling—from ViT back to CNN. In *Proceedings of the 40th International Conference on Machine Learning*, 20149–20167.

Li, T.; Ke, Q.; Rahmani, H.; Ho, R. E.; Ding, H.; and Liu, J. 2021b. Else-Net: Elastic Semantic Network for Continual Action Recognition From Skeleton Data. In *ICCV*, 13434–13443.

Lin, L.; Song, S.; Yang, W.; and Liu, J. 2020. MS2L: Multi-Task Self-Supervised Learning for Skeleton Based Action Recognition. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, 2490–2498.

Lin, L.; Zhang, J.; and Liu, J. 2023. Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In *CVPR*, 2363–2372.

Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.-Y.; and Kot, A. C. 2020a. 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.

Liu, J.; Shahroudy, A.; Xu, D.; Kot, A. C.; and Wang, G. 2017. 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.

Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; and Tang, J. 2023. Self-Supervised Learning: Generative or Contrastive. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 35(1): 857–876.

Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; and Ouyang, W. 2020b. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, 143–152.

Mao, Y.; Deng, J.; Zhou, W.; Fang, Y.; Ouyang, W.; and Li, H. 2023. Masked motion predictors are strong 3d action representation learners. In *ICCV*, 10181–10191.

Mao, Y.; Zhou, W.; Lu, Z.; Deng, J.; and Li, H. 2022. CMD: Self-Supervised 3D Action Representation Learning with Cross-Modal Mutual Distillation. In *ECCV*, 734–752.

Qiu, H.; Hou, B.; Ren, B.; and Zhang, X. 2022. Spatio-Temporal Tuples Transformer for Skeleton-Based Action Recognition. *arXiv preprint arXiv:2201.02849*.

Qu, H.; Cai, Y.; and Liu, J. 2024. Llms are good action recognizers. In *CVPR*, 18395–18406.

Rao, H.; Xu, S.; Hu, X.; Cheng, J.; and Hu, B. 2021. Augmented skeleton based contrastive action learning with momentum LSTM for unsupervised action recognition. *Information Sciences*, 569: 90–109.

Shah, A.; Roy, A.; Shah, K.; Mishra, S.; Jacobs, D.; Cherian, A.; and Chellappa, R. 2023. Halp: Hallucinating latent positives for skeleton-based self-supervised learning of actions. In *CVPR*, 18846–18856.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *CVPR*, 1010–1019.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2021. AdaSGN: Adapting Joint Number and Model Size for Efficient Skeleton-Based Action Recognition. In *ICCV*, 13413–13422.

Su, K.; Liu, X.; and Shlizerman, E. 2020. PREDICT & CLUSTER: Unsupervised skeleton based action recognition. In *CVPR*, 9631–9640.

Thoker, F. M.; Doughty, H.; and Snoek, C. G. 2021. Skeleton-Contrastive 3D Action Representation Learning. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 1655–1663.

Tianyu, G.; Hong, L.; Zhan, C.; Mengyuan, L.; Tao, W.; and Runwei, D. 2022. Contrastive Learning from Extremely Augmented Skeleton Sequences for Self-supervised Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wu, W.; Hua, Y.; zheng, C.; Wu, S.; Chen, C.; and Lu, A. 2022. SkeletonMAE: Spatial-Temporal Masked Autoencoders for Self-supervised Skeleton Action Recognition. *arXiv preprint arXiv:2209.02399*.

Xiang, W.; Li, C.; Zhou, Y.; Wang, B.; and Zhang, L. 2023. Generative action description prompts for skeleton-based action recognition. In *ICCV*, 10276–10285.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. Simmim: A simple framework for masked image modeling. In *CVPR*, 9653–9663. IEEE.

Xu, J.; Yu, Z.; Ni, B.; Yang, J.; Yang, X.; and Zhang, W. 2020. Deep kinematics analysis for monocular 3D human pose estimation. In *CVPR*, 899–908.

Yang, S.; Liu, J.; Lu, S.; Er, M. H.; and Kot, A. C. 2021. Skeleton cloud colorization for unsupervised 3D action representation learning. In *ICCV*, 13423–13433.

Zhang, H.; Hou, Y.; Zhang, W.; and Li, W. 2022. Contrastive Positive Mining for Unsupervised 3D Action Representation Learning. In *ECCV*, 36–51.

Zhang, P.; Lan, C.; Zeng, W.; Xing, J.; Xue, J.; and Zheng, N. 2020. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *CVPR*, 1112–1121.

Zheng, N.; Wen, J.; Liu, R.; Long, L.; Dai, J.; and Gong, Z. 2018. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2644–2651.

Zhou, Y.; Yan, X.; Cheng, Z.-Q.; Yan, Y.; Dai, Q.; and Hua, X.-S. 2024. BlockGCN: Redefine Topology Awareness for Skeleton-Based Action Recognition. In *CVPR*, 2049–2058.

Zhu, Y.; Han, H.; Yu, Z.; and Liu, G. 2023. Modeling the relative visual tempo for self-supervised skeleton-based action recognition. In *ICCV*, 13913–13922.