Frequency Decoupled Masked Auto-encoder for Self-supervised Skeleton-based Action Recognition

Ye Liu, *Member, IEEE*, Tianhao Shi, Mingliang Zhai and Jun Liu, *Senior Member, IEEE*

***Abstract*—In 3D skeleton-based action recognition, the limited availability of supervised data has driven interest in self-supervised learning methods. The reconstruction paradigm using masked auto-encoder (MAE) is an effective and mainstream self-supervised learning approach. However, recent studies indicate that MAE models tend to focus on features within a certain frequency range, which may result in the loss of important information. To address this issue, we propose a frequency decoupled MAE. Specifically, by incorporating a scale-specific frequency feature reconstruction module, we delve into leveraging frequency information as a direct and explicit target for reconstruction, which augments the MAE's capability to discern and accurately reproduce diverse frequency attributes within the data. Moreover, in order to address the issue of unstable gradient updates caused by more complex optimization objectives with frequency reconstruction, we introduce a dual-path network combined with an exponential moving average (EMA) parameter updating strategy to guide the model in stabilizing the training process. We have conducted extensive experiments which have demonstrated the effectiveness of the proposed method.**

***Index Terms*—Skeleton-based action recognition, Masked auto-encoder, Self-supervised learning, Frequency domain.**

I. Introduction

HUMAN action recognition is vital for applications like surveillance, human-robot interaction, and virtual reality. The development of depth sensors and pose estimation algorithms has made skeleton data easily accessible. Skeleton data exhibits computational efficiency and a robust spatio-temporal correlation, thereby emerging as an indispensable data source for action recognition [1], [2].

In the early stages of research, foundational models tailored for capturing spatio-temporal features and trained using supervised learning were predominantly employed for action recognition tasks, such as H-RNN [3], ST-LSTM [4] and some CNN-based methods [5], [6]. In 2018, Yan et al. [7] proposed ST-GCN which introduced graph convolution network (GCN) for dynamic skeleton modeling. Since then, GCNs have been extensively applied and remain the mainstream approach to this day [8]–[12]. Recently, several studies [13]–[15] have leveraged the strong contextual modeling capabilities of Transformers. In addition to these advances, hybrid models [16], [17] have also shown promising results.

Ye Liu, Tianhao Shi and Mingliang Zhai are with College Of Automation & Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210023, China, (e-mails: yeliu@njupt.edu.cn, 1223055915@njupt.edu.cn, zhaimingliang@njupt.edu.cn).

Jun Liu is with School of Computing and Communications, Lancaster University, UK. (Email: j.liu81@lancaster.ac.uk)

Corresponding author: Ye Liu.

Due to the high cost and effort required to obtain labeled data for supervised learning, researchers have also been exploring self-supervised learning methods for skeleton-based action recognition, which mainly involves two phases: pre-training and fine-tuning. In the unsupervised pre-training phase, the model learns features from skeleton sequences in pretext tasks. In the fine-tuning phase, the pre-trained model with learned representations is used for feature extraction, and a classification head is trained with labeled data. Pretext tasks are mainly categorized into two types: contrastive learning based and reconstruction based. Contrastive learning harnesses the power of learning discriminative features by comparing different augmentations of the same data against augmentations from different data. Previous studies, such as HiCLR [18], HiCo [19], and ActCLR [20], have demonstrated the effectiveness of contrastive learning as a pretext task for self-supervised skeleton-based action recognition. Unlike contrastive learning, the reconstruction based methods learn to restore the original signal from altered versions, such as those that have been masked or noised. Among them, masked auto-encoder (MAE) [21] learns visual representations by randomly masking a large portion of input images and training an encoder-decoder architecture to reconstruct the original images from the remaining visible parts. MAE has also been introduced in skeleton-based action recognition, for example, SkeletonMAE [22], [23], MAMP [24], which achieved impressive performance.

However, recent studies [25], [26] suggest that reconstruction based methods tend to focus on and utilize the high-frequency information in the input signal. While high-frequency components excel at capturing detailed motion, low-frequency components contribute to identifying broader motion trends and directions, which is also crucial for action recognition.

To overcome this limitation, we introduce an innovative frequency decoupled masked auto-encoder (MAE) architecture, the key of which is a pivotal scale-specific frequency reconstruction module subsequent to the decoder stage. This module endeavors to reconstruct distinct frequency characteristics with tailored temporal resolutions: enhancing high-frequency features at a heightened temporal precision while preserving low-frequency details with a more subdued temporal granularity. By explicitly disentangling and targeting these two in frequency domains within the reconstruction objectives, the module fosters a dual-pronged approach within MAE: it not only prompts the model to intensify its focus on low-frequency information but also harnesses its innate sensitivity to high-frequency nuances, ultimately yielding a more holistic and comprehensive feature representation. Since a more com-

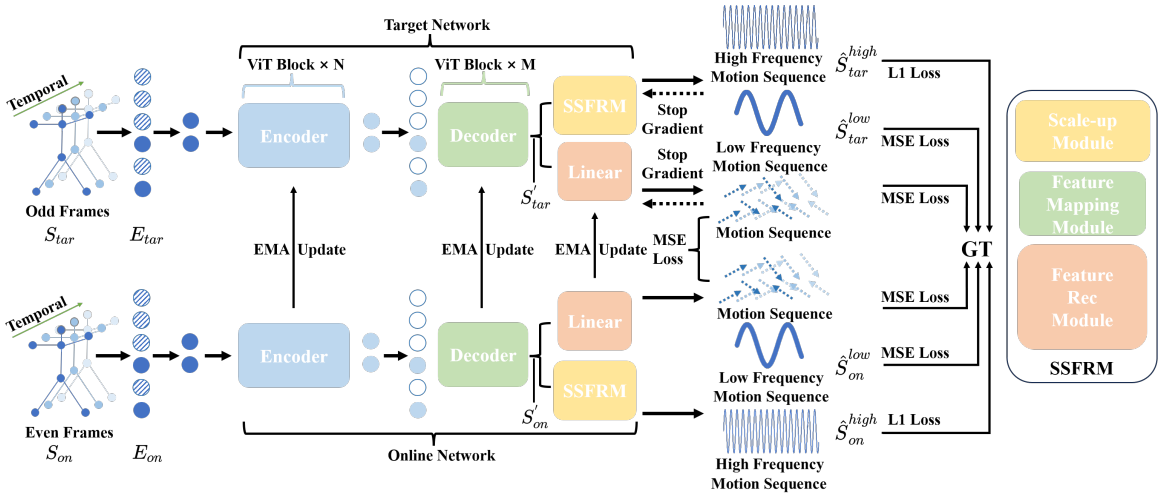

Fig. 1. Pipeline of our method (pre-training phase): Frequency Decoupled MAE. The lower part of the diagram represents the online network, which is updated using gradients. The upper part of the diagram represents the target network, which is updated through guidance from the online network and the exponential moving average (EMA) strategy. Decoupled frequency reconstruction is performed in both the online network and the target network.

plex reconstruction objective is introduced, instability may be induced in the training process, we thus employ a dual-path structure strategy to assist the model in stable and efficient training. In summary, the main contributions of this letter are:

- To focus on diverse frequency attributes in skeleton data, we propose a scale-specific frequency reconstruction module, enabling the MAE model to comprehensively learn the frequency representation of skeleton sequence.
- We introduce a dual-path MAE network structure and set distinct input for two sub-networks, which improves the model's training stability while training with more complex loss with frequency reconstruction.
- We conducted extensive experiments to validate the effectiveness of the proposed method.

II. Method

A. Overview

The proposed method's overall pipeline is illustrated in Fig. 1. We first preprocess the original 3D skeleton sequence dividing it into two subsequences with equal length which serve as inputs to the dual-path network: S_{tar} of odd frames to the target network and S_{on} of even frames to the online network. Subsequently, S_{tar} and S_{on} are masked and embedded, yielding masked token sequences E_{tar}, E_{on}, which remove semantically rich parts to force the model to learn high-level semantic features. Then these masked tokens are encoded and decoded to reconstruct the original inputs at two temporal scales. In this process, a novel scale-specific frequency reconstruction module is incorporated. Afterward, a tailored frequency reconstruction loss is incorporated into the loss function to supervise and enhance the reconstruction process.

B. Multi-tube Masking and Tokenization

Tube Masking, originally proposed for video MAE [27], masks parts of frame pixels of the same locations across the temporal dimension to avoid information leakage. For skeleton sequence data, a similar masking strategy is employed. Given a skeleton sequence of size $T \times V \times C$, we firstly temporally divide it into N equal-length tubes of size $\frac{T}{N} \times V \times C$. For each tube, K joints out of V are randomly picked for masking, which means the indices of masked joints are shared across the $\frac{T}{N}$ frames in each tube.

Following previous work [24], the masked motion sequences are then pachified as masked token sequences E_{tar} and E_{on}, which are fed to an encoder-decoder stage that consists of a series of Transformer blocks. And the decoded data are unpatchified as S'_{tar} and S'_{on} to recover spatial and temporal structure.

C. Scale-specific Frequency Reconstruction

As a further decoding stage, we introduce a scale-specific frequency reconstruction module (SSFRM), which takes S'_{tar} and S'_{on} as inputs and functions to reconstruct high-frequency data at a larger time scale while simultaneously reconstructing low-frequency data at a smaller time scale. SSFRM consists of three sub-modules: scale-up module, feature mapping module, and feature reconstruction module, the details of these sub-modules are shown in Fig. 2.

Scale-up Module: This module is responsible for increasing the temporal resolution of the decoded embeddings. It consists of two parts: the first includes two transposed convolution and $GELU$ layers that upscale the decoded embeddings in the temporal dimension by $2\times$, resulting in a low-temporal resolution output. The second comprises a transposed convolution and a $GELU$ layer that upscale the embeddings by $4\times$, producing a high-temporal resolution output.

Feature Mapping Module: The feature mapping module includes a grouped convolution layer and a point-wise convolution layer, designed to further map the decoded embeddings into representations that are more closely aligned with frequency characteristics.

Feature Reconstruction Module: The feature reconstruction module also utilizes grouped and point-wise convolution layers, along with transposed convolution layers, aiming to restore the frequency feature sequence as the final output.

The final outputs of the SSFRM are $\hat{S}_{tar}^{high}$, $\hat{S}_{tar}^{low}$ for target branch and $\hat{S}_{on}^{high}$, $\hat{S}_{on}^{low}$ for online branch.

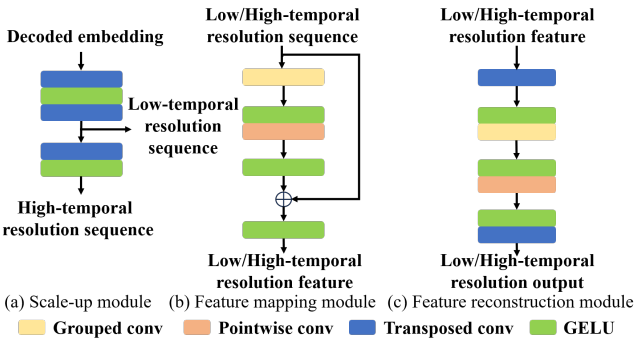

Fig. 2. The three sub-modules of the scale-specific frequency reconstruction module.

D. Dual-path Network Architecture

Since a more complex reconstruction target is introduced, the training process may become unstable. To address this problem, inspired by BYOL [28], we design a two-branch architecture: an online branch with parameters updating with gradients, and a target branch that updates its parameters using the EMA strategy. We aim to use the online network to guide the target network while exploiting the characteristics of EMA to alleviate the adverse effects of unstable gradients. The EMA update strategy can be written as follows:

$$\theta_{tar}^{'} = \gamma\theta_{tar} + (1-\gamma)\theta_{on}^{'}, \tag{1}$$

where $\theta_{tar}^{'}$ is the updated target network parameters and γ is a hyperparameter controlling the influence of the online network's parameters on the target network.

E. Loss Function

The loss consists of three parts: motion reconstruction loss L_m, motion consistency loss L_{mc}, and motion frequency feature reconstruction loss L_{mf}. Instead of reconstructing the original skeleton sequence, we set one of the reconstruction targets as the motion sequence, which is a more challenging task and compels the model to learn the dynamic properties of skeleton data. The motion reconstruction loss is written as follows:

$$L_m = MSE(sg(\varphi(S_{tar}^{\prime})), d(S_{tar})) + MSE(\varphi(S_{on}^{\prime}), d(S_{on})), \tag{2}$$

where $\varphi(.)$ is a linear layer that adjusts the number of feature channels. $d(.)$ is the differential operator that computes the motion of skeletons, and $sg(x)$ refers to the operation of stopping the gradient update of x.

L_{mc} is added to enforce the consistency between the two branches:

$$L_{mc} = MSE(sg(\varphi(S_{tar}^{\prime})), \varphi(S_{on}^{\prime})), \tag{3}$$

L_{mf} is added to supervise the frequency reconstruction process. We apply discrete cosine transform (DCT) and truncation operation to the motion sequence to obtain high/low motion frequency coefficients, then apply inverse DCT to the high/low coefficients to obtain ground truth:

$$F_{tar}^{high} = iDCT(\tau^{\uparrow}(DCT(d(S_{tar})))), \tag{4}$$

$$F_{tar}^{low} = iDCT(\tau^{\downarrow}(DCT(d(S_{tar})))), \tag{5}$$

where $\tau^{\uparrow}$ and $\tau^{\downarrow}$ are a pair of truncation operations, which set the elements in a tensor to zeros if they are less/greater than a certain threshold value. F_{on}^{high}, F_{on}^{low} can be computed similarly. The frequency feature reconstruction loss can be written as:

$$\begin{aligned} L_{mf} =& |sg(\hat{S}_{tar}^{high}) - F_{tar}^{high}| + |\hat{S}_{on}^{high} - F_{on}^{high}| + \\ & MSE(sg(\hat{S}_{tar}^{low}), F_{tar}^{low}) + MSE(\hat{S}_{on}^{low}, F_{on}^{low}), \end{aligned} \tag{6}$$

The final loss function is:

$$L = L_m + \eta_1 * L_{mc} + \eta_2 * L_{mf}, \tag{7}$$

where η_1 and η_2 are hyperparameters used to tune the intensity of the motion consistency loss and the motion frequency feature reconstruction loss.

III. Experiment

A. Dataset

NTU-RGB+D 60: The NTU-60 dataset [29] developed by NTU ROSE Lab includes 60 human action categories with 56,880 samples. It has two evaluation protocols: cross-subject (X-sub) and cross-view (X-view). X-sub divides 40 subjects into training and testing sets (20 each). X-view uses samples from cameras 2 and 3 for training and camera 1 for testing.

NTU-RGB+D 120: The NTU-120 dataset [30] extends NTU-60 by adding 57,600 samples for a total of 114,480 samples. It covers 120 action categories and includes X-sub and cross-setup (X-set) evaluations. X-set uses samples with even setup IDs for training and odd setup IDs for testing.

TABLE I
PERFORMANCE COMPARISON ON THE NTU-60 DATASET UNDER THE LINEAR EVALUATION. UNDERLINED DATA INDICATES THE SECOND-BEST RESULT.

Method	NTU-60	
	X-sub	X-view
SkeletonMAE [22]	74.8	77.7
HiCLR-GCN [18]	80.4	85.5
HiCo-LSTM [19]	81.4	88.8
ActCLR [20]	84.3	88.8
MAMP [24]	<u>84.9</u>	<u>89.1</u>
FD-MAE (Ours)	**86.4**	**90.4**

B. Comparison With Previous Methods

Linear Evaluation: Linear evaluation primarily involves freezing the model parameters obtained during the pre-training phase and adding a linear classifier head to the output of the pre-trained model's encoder. Only this linear head undergoes

TABLE II
PERFORMANCE COMPARISON ON THE NTU-120 DATASET UNDER THE LINEAR EVALUATION.

Method	NTU-120	
	X-sub	X-set
SkeletonMAE [22]	72.5	73.5
HiCLR-GCN [18]	70.0	70.4
HiCo-LSTM [19]	73.7	74.5
ActCLR [20]	74.3	75.7
MAMP [24]	78.6	79.1
FD-MAE (Ours)	**78.9**	**79.9**

supervised training. Tables I and Table II compare the performance of our method with previous self-supervised methods, our method shows significant improvements over the previous works.

Semi-supervised Evaluation: Semi-supervised evaluation refers to fine-tuning all parameters of the pre-trained model's encoder and the newly added linear head, using only a portion of the training set for supervised training while employing the full test set for evaluation. The experimental results are presented in Table III.

TABLE III
PERFORMANCE COMPARISON ON THE NTU-60 DATASET UNDER THE SEMI-SUPERVISED EVALUATION.

Method	NTU-60			
	X-sub		X-view	
	(1%)	(10%)	(1%)	(10%)
SkeletonMAE [22]	54.4	80.6	54.6	83.5
HiCo-Transformer [19]	54.4	73.0	54.8	78.3
HiCLR-Transformer [18]	54.7	82.1	53.7	84.8
HiCLR-GCN [18]	58.5	79.6	58.3	84.0
MAMP [24]	66.0	88.0	68.7	91.5
FD-MAE (Ours)	**70.5**	**88.4**	**74.4**	**92.9**

TABLE IV
ABLATION ON PROPOSED METHOD.

DP	FR	MTM	NTU-60
-	-	-	84.9
✓	-	-	85.4
✓	✓	-	85.8
✓	✓	✓	**86.4**

TABLE V
ABLATION ON MASK RATIO.

Mask Ratio	NTU-60
80%	85.1
88%	**86.4**
90%	86.0
95%	82.9

TABLE VI
ABLATION ON TUBE LENGTH.

Tube Length	NTU-60
2	85.9
3	**86.4**
5	86.2

TABLE VII
ABLATION ON DECODER DEPTH.

Decoder Depth	NTU-60
1	84.7
2	85.4
3	**86.4**

C. Ablation Study

To assess the impact of each module, we conducted ablation studies in which we incrementally added three key components: the dual-path network, the frequency reconstruction module, and the multi-tube masking. The results are presented in Table IV. All ablation experiments were conducted under the NTU-60 X-sub linear evaluation. We also compared the pre-training stability of dual-path MAE and classic MAE. Benefiting from the EMA strategy, the pre-training motion reconstruction loss of the dual-path MAE is relatively smoother, as shown in Fig. 3. Additionally, we examined the frequency characteristic of the improved model. Results in Fig. 4 show that the model with the frequency reconstruction module better utilizes low-frequency signal. We also conducted ablation experiments on some important hyperparameters, selecting the optimal hyperparameters by considering both computational efficiency and performance. The results are presented in Tables V, VI, and VII.

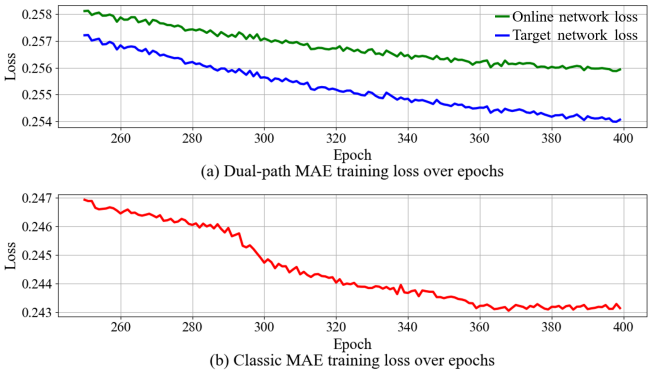

Fig. 3. Training stability comparison between our dual-path MAE and the classic MAE.

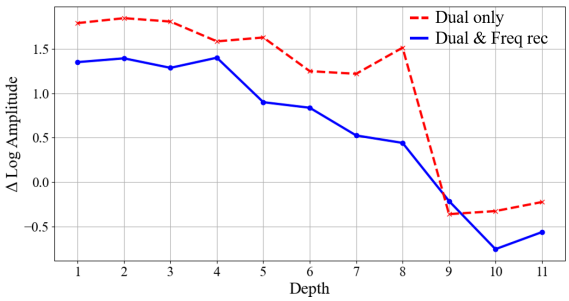

Fig. 4. Comparison of relative log amplitude (the amplitude difference between the highest and lowest frequencies of embedding [25], lower value means smaller variation in frequency) across network layers (depth) with and without the frequency reconstruction module.

IV. CONCLUSION

We propose an enhanced MAE model for self-supervised 3D skeleton-based action recognition. The introduced frequency reconstruction module is designed to address the imbalance in attention between different frequency components of input signal caused by MAE's characteristics, while a dual-path structure is applied to enhance training stabilization. To prevent information leakage, we use multi-tube masking. Experimental results indicate that the proposed method achieves performance improvements over previous methods.

REFERENCES

[1] R. Yue, Z. Tian, and S. Du, "Action recognition based on rgb and skeleton data sets: A survey," *Neurocomputing*, vol. 512, pp. 287–306, Nov. 2022.

[2] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3200–3225, Mar. 2022.

[3] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1110–1118.

[4] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4263–4270.

[5] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3288–3297.

[6] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, and M. He, "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2017, pp. 601–604.

[7] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7444–7452.

[8] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13 359–13 368.

[9] H.-g. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani, "Infogcn: Representation learning for human skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20 186–20 196.

[10] W. Peng, J. Shi, and G. Zhao, "Spatial temporal graph deconvolutional network for skeleton-based human action recognition," *IEEE Signal Process. Lett.*, vol. 28, pp. 244–248, 2021.

[11] J. Lee, M. Lee, D. Lee, and S. Lee, "Hierarchically decomposed graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 10 444–10 453.

[12] Y. Zhang, Z. Sun, M. Dai, J. Feng, and K. Jia, "Cross-scale spatiotemporal refinement learning for skeleton-based action recognition," *IEEE Signal Process. Lett.*, vol. 31, pp. 441–445, 2024.

[13] L. Wang and P. Koniusz, "3mformer: Multi-order multi-mode transformer for skeletal action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5620–5631.

[14] J. Do and M. Kim, "Skateformer: Skeletal-temporal transformer for human action recognition," 2024, *arXiv:2403.09508*.

[15] D. Ahn, S. Kim, H. Hong, and B. C. Ko, "Star-transformer: a spatio-temporal cross attention transformer for human action recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 3330–3339.

[16] J. Kong, Y. Bian, and M. Jiang, "Mtt: Multi-scale temporal transformer for skeleton-based action recognition," *IEEE Signal Process. Lett.*, vol. 29, pp. 528–532, 2022.

[17] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Comput. Vis. Image Underst.*, vol. 208-209, Jul. 2021, Art. no. 103219.

[18] J. Zhang, L. Lin, and J. Liu, "Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 3427–3435.

[19] J. Dong, S. Sun, Z. Liu, S. Chen, B. Liu, and X. Wang, "Hierarchical contrast for unsupervised skeleton-based action representation learning," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 525–533.

[20] L. Lin, J. Zhang, and J. Liu, "Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2363–2372.

[21] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16 000–16 009.

[22] W. Wu, Y. Hua, C. Zheng, S. Wu, C. Chen, and A. Lu, "Skeletonmae: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2023, pp. 224–229.

[23] H. Yan, Y. Liu, Y. Wei, Z. Li, G. Li, and L. Lin, "Skeletonmae: graph-based masked autoencoder for skeleton sequence pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 5606–5618.

[24] Y. Mao, J. Deng, W. Zhou, Y. Fang, W. Ouyang, and H. Li, "Masked motion predictors are strong 3d action representation learners," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 10 181–10 191.

[25] N. Park, W. Kim, B. Heo, T. Kim, and S. Yun, "What do self-supervised vision transformers learn?" in *Proc. 11th Int. Conf. Learn. Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=azCKuYyS74

[26] A. Vanyan, A. Barseghyan, H. Tamazyan, V. Huroyan, H. Khachatrian, and M. Danelljan, "Analyzing local representations of self-supervised vision transformers," 2023, *arXiv:2401.00463*.

[27] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 10 078–10 093.

[28] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21 271–21 284.

[29] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1010–1019.

[30] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2019.