Supplementary for "Self-supervised 3D Skeleton Action Representation Learning with Motion Consistency and Continuity"

Yukun Su[1,2], Guosheng Lin[3†], and Qingyao Wu[1,2†]

[1]School of Software and Engineering, South China University of Technology
[2]Key Laboratory of Big Data and Intelligent Robot, Ministry of Education
[3]School of Computer Science and Engineering, Nanyang Technological University
suyukun666@gmail.com, gslin@ntu.edu.sg, qyw@scut.edu.cn

A. Appendix

This appendix provides more experiment results and visualizations. Sec. A.1 presents the details of the decoder of the motion continuity modeling module, Sec. A.2 shows more additional experiments on self-supervised learning, Sec. A.3 gives the training curves in more detail and Sec. A.4 shows more qualitative results of the interpolation.

A.1. Interpolation Decoder Architecture

The architecture of the skeleton interpolation decoder is presented in Table 1. Note that we employ different backbone models including ST-GCN [7], 2S-AGCN [5] and AS-GCN [3] as our network architectures. Each network will differ in the details of the convolution operation, however, they all share the main operation termed "spatial-temporal convolution" that is proposed in [7]. Uniformly, the first four convolutional blocks reduce the frame number to aggregate higher-level action features. For the last layer, we adopt a simply modified spatial-temporal deconvolution operation. Finally, the tensor with shape $[25, 3, 64]$ can be obtained from a fully connected layer, which contains the joint position of the interpolated 64 frames.

A.2. Comparison with other methods on PKUMMD

As shown in Table 2, we compare our method with the state-of-the-art self-supervised learning methods. All the networks are self-pretrained on NTU dataset and then initialized the weights on PKUMMD dataset. As we can see, our MCC method achieve the best performance and outperform other existing methods by a large margin, which demonstrate the effectiveness of the proposed method.

†Corresponding authors.

Input-Shape	Operation	Output-Shape
$[25, 256, 8]$	S-GCN T-GCN, stride=1	$[25, 128, 8]$
$[25, 128, 8]$	S-GCN T-GCN, stride=2	$[25, 128, 4]$
$[25, 128, 4]$	S-GCN T-GCN, stride=2	$[25, 128, 2]$
$[25, 128, 2]$	S-GCN T-GCN, stride=2	$[25, 96, 1]$
$[25, 96, 1]$	Deconv S-GCN Deconv T-GCN, stride=1	$[25, 192, 1]$
$[25, 192, 1]$	FC layer	$[25, 3, 64]$

Table 1. The architecture of the skeleton interpolation decoder. S-GCN indicates the spatial-convolution, and T-GCN indicates the temporal-convolution that are both proposed in [7]. The input tensor with shape $[25, 256, 8]$ is obtained from the encoder, where 25 represents the joint number, 256 is the feature channel, and 8 means the frame number.

A.3. Training Process

To further demonstrate the process of training network from scratch and our self-supervised learning for pre-training, Figure 1 shows the accuracy and loss curves. It is noticeable that when employing the self-supervised pre-trained weights, the network can achieve higher accuracy in different datasets with lower loss, which shows the effectiveness of our self-supervised learning method.

A.4. More Visualizations

More skeleton interpolation results of different actions are illustrated in Figure 2, which contains the action of "brush teeth", "stand up" and "kicking something". As

Method	Architecture	PKUMMD (Acc.)
LongT GAN [8]	unidirectional GRUs	44.8
MS^{2}L [4]	BiGRU	45.8
Clip Order prediction [6]$_{\text{CVPR'2019}}$	ST-GCN	51.2
	2S-AGCN	53.8
	AS-GCN	55.7
Jigsaw puzzle recognition [2]$_{\text{AAAI'2019}}$	ST-GCN	50.4
	2S-AGCN	56.6
	AS-GCN	55.4
pace prediction [1]$_{\text{CVPR'2020}}$	ST-GCN	49.7
	2S-AGCN	54.9
	AS-GCN	55.8
MCC (ours)	ST-GCN	**54.5**
	2S-AGCN	**60.8**
	AS-GCN	**58.4**

Table 2. Comparison with other state-of-the-art self-supervised methods on PKUMMD Part-II subset.

we can see, the actions are interpolated with very low error compared with the target ground-truth frames. Note that our self-supervised learning method is not specifically designed for interpolating the human skeleton.

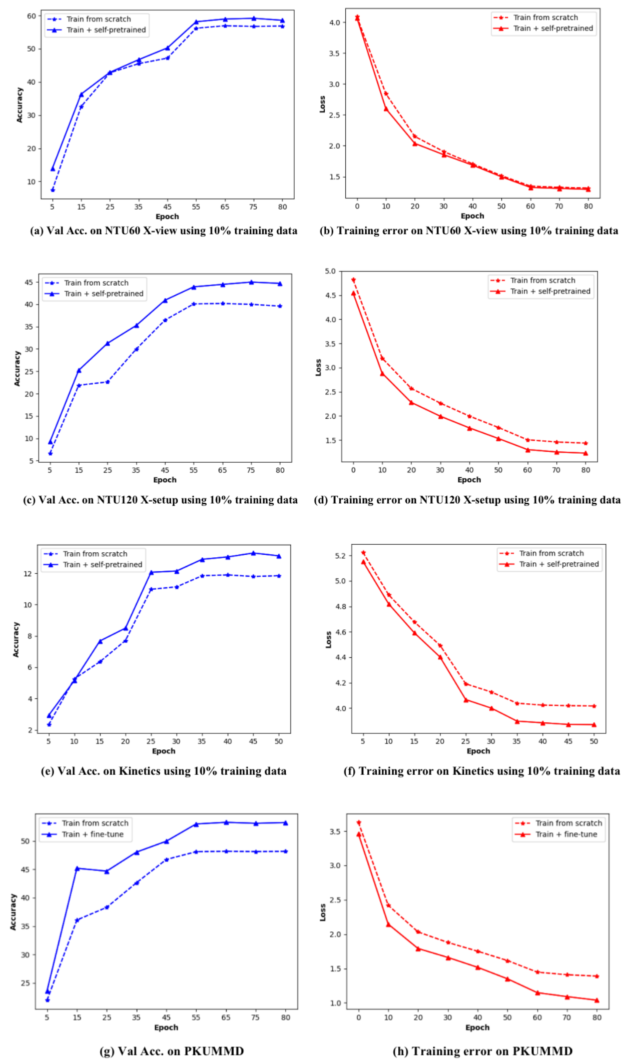

Figure 1. Accuracy (blue) and Loss (red) comparison (backbone: ST-GCN [7]) in different datasets. $1^{st}\sim3^{rd}$ row: the accuracy and loss curves between the network training from scratch and initializing the pre-trained weights by self-supervised learning when using only 10% of labeled training data on NTU60 X-view subset, NTU120 X-setup subset, and Kinetics dataset, respectively. 4^{th} row: the accuracy and loss curves between the network training from scratch and initializing the weights learned on larger datasets through self-supervised pre-training.

References

[1] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9922–9931, 2020. 2

[2] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8545–8552, 2019. 2

[3] 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 Conference on Computer Vision and Pattern Recognition*, pages 3595–3603, 2019. 1

[4] 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*, pages 2490–2498, 2020. 2

[5] 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 Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2019. 1

[6] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019. 2

[7] 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*, volume 32, 2018. 1, 2, 3

[8] 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 *Thirty-Second AAAI conference on artificial intelligence*, 2018. 2

Figure 2. More visualizations of the skeleton action samples from the interpolation module (backbone: ST-GCN [7]) on NTU60-RGB+D dataset. (a) Action of "brush teeth". (b) Action of "stand up". (c) Action of "kicking something".