SkeletonMAE: Graph-based Masked Autoencoder for Skeleton Sequence Pre-training

Supplementary Material

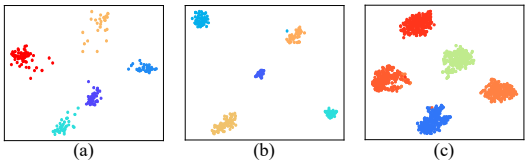
This supplementary material will further detail the following aspects of the submitted manuscript: A. Experimental Implementation Details, B. More Verification, C. More Visualization Results, D. Spatial Modeling Method, E. Supplementary Experiments.

A. Experimental Implementation Details

In this section, we will introduce the details of the experimental application. Firstly, our SkeletonMAE is pre-trained by an Adaptive Moment Estimation (Adam) [4] optimizer with the initial learning rate as 1.5×10^{-4} and the PReLU [3] as an activation function. The batch size is 1024 and the training epoch is 50. And we will set the mask body part for pre-training. Furthermore, the input feature dropout is 0.2 and the attention dropout is 0.1. At the fine-tuning stage, we use the Stochastic Gradient Descent (SGD) [1] with momentum (0.9) and adopt the warm-up strategy for the first 5 epochs. The weight decay for the optimizer is 0.005. The total fine-tuning epoch is 110. The learning rate is initialized to 0.1 and is divided by 10 at the 90 epoch and the 100 epoch. We employ 0.1 for label smoothing and the number of workers for data loader is 4. We use a large batch size of 128 to facilitate training our attention mechanism and enhancing the model's perception for all human action categories. Both our pre-training and fine-tuning models are implemented in PyTorch 1.9.1 [6], and we train our models on a single NVIDIA GeForce RTX 2080Ti 11GB GPU with CUDA 11.1. The number of parameters of our SkeletonMAE model is about 1.5M. Compared to previous self-supervised pre-training models [8, 7, 2], our model is more light-weighted.

B. More Verification

As shown in Fig. 1, both PCA and t-SNE are promising dimensionality reduction methods commonly used for feature visualization. In Fig. 2, we show the distribution of the difference between the average accuracy of all categories and the accuracy of each category. We try to briefly illustrate the effectiveness of the pre-trained SkeletonMAE encoder proposed in this paper through accuracy



(a) (b) (c)

Figure 1. (a) and (b) are visualisations by 2d-PCA and t-SNE on FineGYM, respectively. (c) is visualisations t-SNE on NTU 60.

difference. Firstly, the SkeletonMAE encoder enhances the diversity of the receptive field of the internal expression of the network by providing graph embedding. Besides, we thought that the SkeletonMAE encoder has a regularization effect to a certain extent. What's more, we can find that the model SkeletonMAE with the SkeletonMAE encoder greatly reduces the accuracy difference between classes without adding regularization methods, which means that its diverse receptive field expression is of great significance.

C. More Visualization Results

Long-tailed Distribution Fig. 3 shows the long-tail distribution properties of the FineGym dataset. To verify the effectiveness of our SSL model for classes with relatively small instances, in the case of the pre-training body masked part of 3. We randomly selected five fine-grained labels in the last half of the long-tail distribution for visualization in Fig. 4 (a2)-(a3), where the final embedded features before the classifier are visualized by the 2d-PCA and 3d-PCA. In particular, the id, instances, event and label of these randomly selected five fine-grained labels as Tab. 1.

Fig. 4 shows the confusion matrices (Left), 2d-PCA (Middle) and 3d-PCA (Right) of SSL on FineGym, Diving48, NTU 60 X-Sub and NTU 120 X-Sub datasets. For the FineGym (99 actions) dataset, We randomly selected five fine-grained labels in the last half of the long-tail distribution for 2d-PCA and 3d-PCA visualization. And for Diving48 (48 actions), NTU 60 X-Sub (60 actions) and NTU 120 X-Sub (120 actions), five classes are randomly selected for 2d-PCA and 3d-PCA visualization. For all datasets, the final embedded features before the classifier are visualized

id	# ins	event	label
72	60	balance beam	salto backward stretched with **2 twist**
73	67	balance beam	salto backward stretched with **2.5 twist**
80	76	uneven bars	giant circle forward with **1 turn on one arm before handstand phase**
81	83	uneven bars	giant circle forward with **0.5 turn to handstand**
82	72	uneven bars	giant circle forward

Table 1. The five fine-grained labels in the FineGym dataset, which we selected to visualize by the 2d-PCA and 3d-PCA.

by the 2d-PCA and 3d-PCA. Fig. 4 (a1, b1, c1, d1) shows that our SSL works well for fine-grained action recognition tasks on the three datasets. Fig. 4 (a2, a3, b2, b3, c2, c3, d2, d3) shows that our SSL representation is low intra-class variation and high inter-class variation, which further validates that our SkeletonMAE can learn discriminative fine-grained skeleton representation by reconstructing human body structure.

D. Spatial-Temporal Representation Learning Method

According to the vanilla GCN [5], the update rules for a multi-layer GCN that propagates rules layer-wise are as follows:

$$\mathbf{H}_t^{(l+1)} = \sigma\left(\widetilde{\mathbf{D}}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}_t^{(l)}\mathbf{W}^{(l)}\right), \tag{1}$$

where adjacency matrix $\widetilde{\mathbf{A}} = \mathbf{A}+\mathbf{I}_N$, $\mathbf{I}_N$ is the identity matrix, $\widetilde{\mathbf{D}}$ is the diagonal degree matrix of $\widetilde{\mathbf{A}}$,$\mathbf{H}_t^{(l)} \in \mathbb{R}^{N\times D^{(l)}}$ and $\mathbf{W}^{(l)}$ are the matrix of joints representation and trainable weight matrix in the l^{th} layer, respectively. $\mathbf{t}$ is the time index, $\sigma(\cdot)$ denotes an activation function like the ReLU.

To model multiple human skeleton interactions, for $\mathrm{SM}(\mathbf{H}_t^{(l)})$ it can be formalized as:

$$\begin{aligned}\mathrm{SM}(\mathbf{H}_t^{(l)}) = \mathrm{Concat}(&\mathrm{SM}(\mathbf{H}_{t,0}^{(l)}) \oplus \mathrm{SM}(\mathbf{H}_{t,1}^{(l-1)});\\ &\mathrm{SM}(\mathbf{H}_{t,1}^{(l)}) \oplus \mathrm{SM}(\mathbf{H}_{t,0}^{(l-1)})),\end{aligned} \tag{2}$$

where $\mathrm{SM}(\mathbf{H}_{t,0}^{(l)})$ means the 0-th person skeleton sequence feature in l-layer. $\mathrm{Concat}(\cdot;\cdot)$ means to concatenate two features.

E. Supplementary Experiments

From Tab. 2, in the non-long-tailed dataset NTU 60, our body part mask strategy improves significantly compared to the random mask strategy in all cases except for torso and head reconstruction, since the movements of individual limbs are more important for distinguishing an action (the same in FineGYM dataset).

# Masked Joints Number		5	9	12	15
Ratio of Mask Joints		30%	50%	70%	90%
Accuracy of SSL		90.5	91.9	91.8	91.6
Masked Body Part	**High**	92.8 ($\mathcal{V}_3$,$\mathcal{V}_5$)	92.3 ($\mathcal{V}_0$,$\mathcal{V}_3$,$\mathcal{V}_5$)	92.0 ($\mathcal{V}_1$,$\mathcal{V}_2$,$\mathcal{V}_3$, $\mathcal{V}_4$,$\mathcal{V}_5$)	91.9 ($\mathcal{V}_0$,$\mathcal{V}_1$,$\mathcal{V}_3$, $\mathcal{V}_4$,$\mathcal{V}_5$)
	Low	92.2 ($\mathcal{V}_2$,$\mathcal{V}_3$)	91.3 ($\mathcal{V}_0$,$\mathcal{V}_1$)	92.0 ($\mathcal{V}_1$,$\mathcal{V}_2$,$\mathcal{V}_3$, $\mathcal{V}_4$,$\mathcal{V}_5$)	91.7 ($\mathcal{V}_0$,$\mathcal{V}_1$,$\mathcal{V}_2$, $\mathcal{V}_3$,$\mathcal{V}_5$)

Table 2. The comparison of our body part based masked and the random masked strategies in the NTU 60 X-sub dataset.

References

[1] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012. 1

[2] 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*, pages 16000–16009, 2022. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 1

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[5] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2

[6] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 1

[7] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 1

[8] Wenhan Wu, Yilei Hua, Shiqian Wu, Chen Chen, Aidong Lu, et al. Skeletonmae: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition. *arXiv preprint arXiv:2209.02399*, 2022. 1

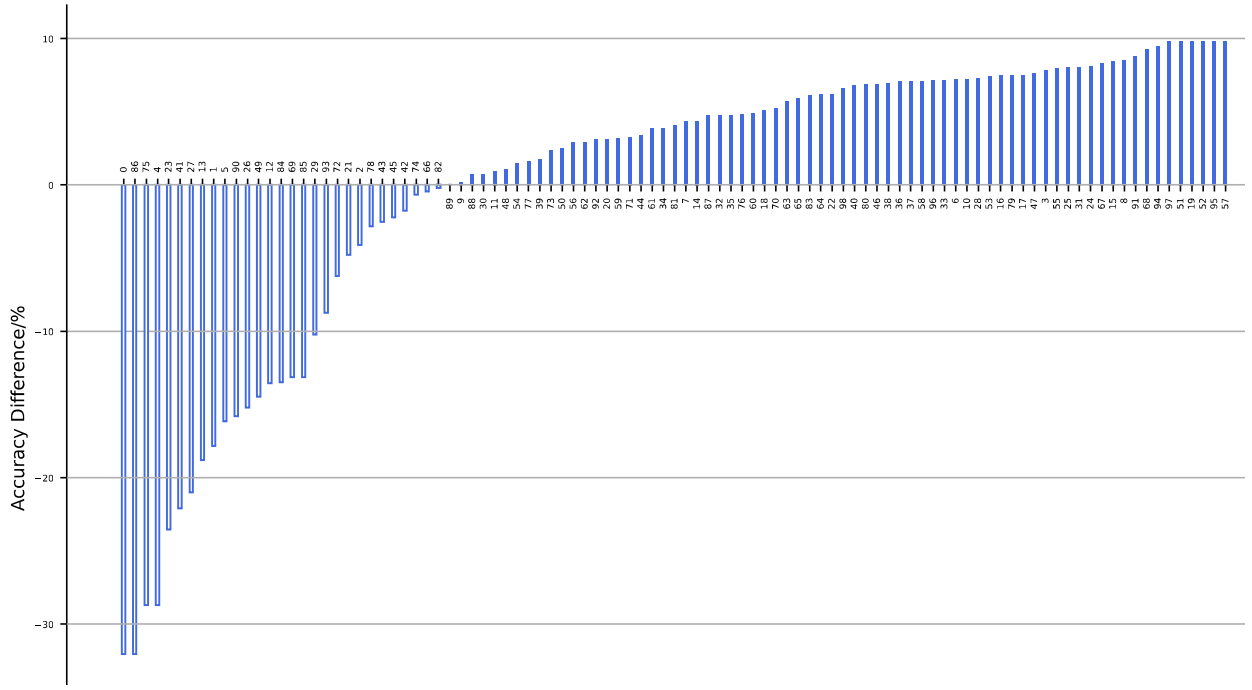


Figure 2. The accuracy difference for our SSL model on the FineGym dataset of 99 fine-grained action labels, accuracy difference for each class obtained by subtracting the average accuracy of all classes for each class.

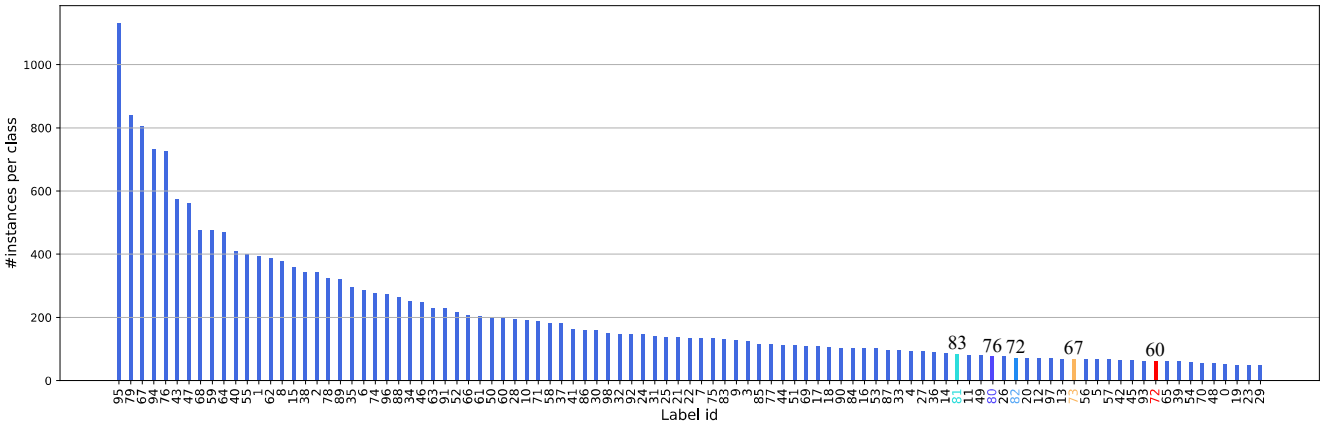


Figure 3. For the long-tailed distribution of the FineGym dataset, we randomly select five fine-grained labels (id: 72, 73, 80, 81 and 82) in the last half of the long-tail distribution for 2d-PCA and 3d-PCA visualization.

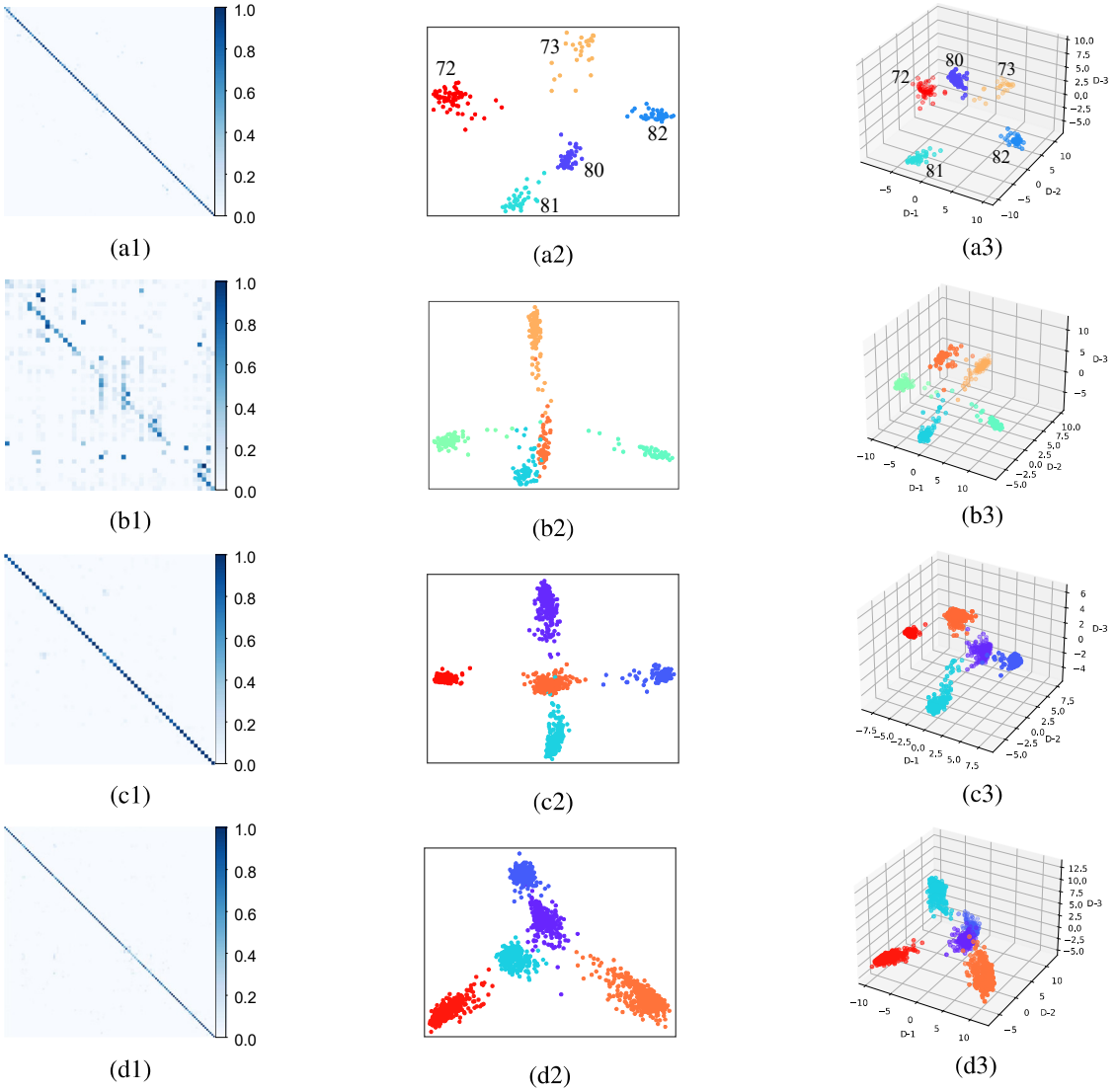


Figure 4. Confusion matrices (Left a1, b1, c1, d1), 2d-PCA (Middle a2, b2, c2, d2) and 3d-PCA (Right a3, b3, c3, d3) of SSL on three datasets. Five classes are randomly selected for 2d-PCA and 3d-PCA visualization. (a1), (a2) and (a3) are for FineGym (99 actions); (b1), (b2) and (b3) are for Diving48 (48 actions); (c1), (c2) and (c3) are for NTU 60 X-Sub (60 actions); (d1), (d2) and (d3) are for NTU-120 X-Sub (120 actions).