Supplementary Material
CVPR 2020 submission ID 9386

1 Selection of the encoder through hyper-parameter search

We estimate the 'training capacity' of P&C network according to recognition accuracy of untrained networks (P&C Rand) and select the best performing model as a baseline for training. As we describe in the main manuscript, due to the clustering properties of random encoder-decoder networks, there could be variations of P&C Rand achieving reasonable baseline performance. In examination of the variants we define the following hyper-parameters: types of recurrent cells (uni-, bi-, GRU, LSTM), number of neurons, number of layers, weights initialization and maximum length of input sequence. We show the evaluation of the hyper-parameter search for NW-UCLA dataset[4] in Fig. 1.

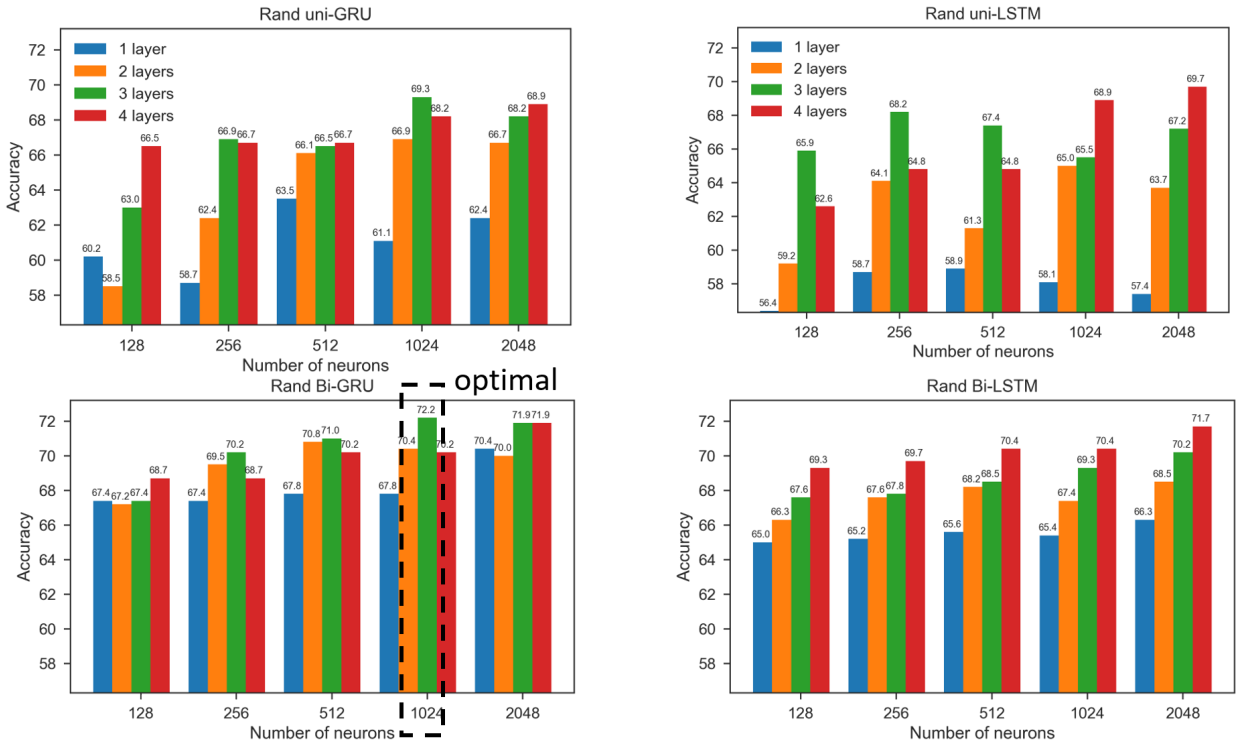

Figure 1: Accuracy of various random network architectures for UCLA dataset.

Specifically, in the hyper-parameter search we consider 4 encoder architectures: Bi-directional GRU, Bi-directional LSTM, Uni-directional GRU and Uni-directional LSTM and select the appropriate number of neurons and layers by evaluating the recognition performance of the randomized encoder. Bi-directional architectures turn out to achieve better performance than Uni-directional ones. We choose the optimal configuration, highlighted in Fig. 1: three layers Bi-GRU with 1024 neurons. Notably, since there is no training involved in the search (forward propagation through the encoder only), the hyper-parameter search is extremely fast (0.7 sec for one batch with size= 64, i.e., $\approx$ 2 sec for the whole NW-UCLA dataset).

We also examine variants of weight initialization. In Fig. 1 we used random uniform initialization $\in [-0.05, 0.05]$) and we compared it with additional variants such as Orthogonal, Random Normal and GloRot Uniform initializations. We find that there is no significant variation in the initialization type (as we show in table 1). We choose the best performing initialization (random uniform) for our final P&C Rand model, however we note that there is no significant difference between the initialization types. Furthermore, we use the same P&C Rand baseline for all three datasets that we evaluate our P&C system on since this optimal architecture appears to perform well as a baseline for various datasets. Since the lengths of the

Initialization	**Random Uniform**	Orthogonal	Random Normal	Glorot Uniform
Rand. Accuracy	**72.2**	71.1	70.2	71.0

Table 1: Comparison of P&C Rand accuracy for various weight initialization for the NW-UCLA dataset.

different action sequences are different we down-sample the sequences to a fixed maximum-length. Therefore, additional hyper-parameter that we optimize for is the maximum length of the input sequence. We have estimated the performance of P&C Rand for various maximum sequence lengths for NW-UCLA dataset. The comparison is shown in Table 2. We obtain that the maximum-length of 50 is the optimal choice for this parameter.

Maximum Length	10	25	35	**50**	75	100
Rand. Accuracy	65.4	66.7	69.5	**72.2**	71.0	70.4

Table 2: Comparison of P&C Rand accuracy for various maximum-length values for the NW-UCLA dataset.

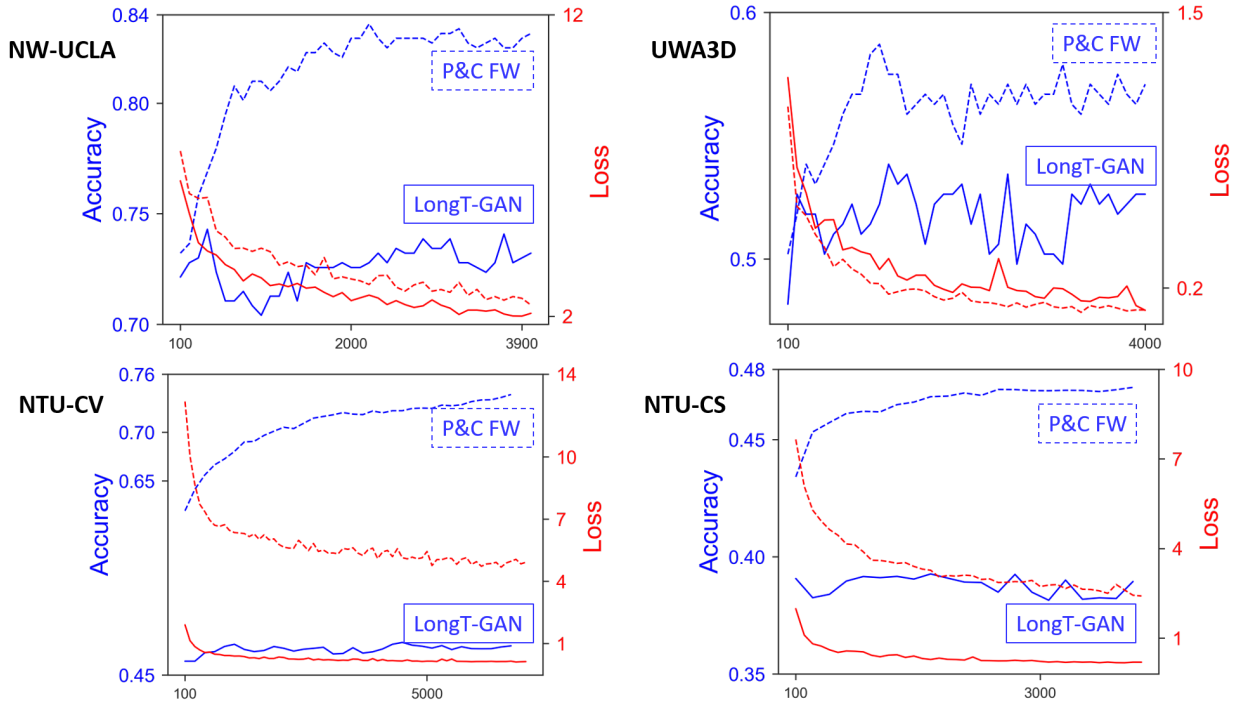

Figure 2: Accuracy(blue) and Loss(red) comparison in all datasets: LongT-GAN(Solid),P&C FW(Dashed)

2 Comparison with LongT-GAN (Unsupervised Skeleton-based method)

We implement the LongT-GAN network by following the description in [5] and then compare our P&C approach trained on all three datasets, see Fig. 2. The comparison shows that LongT-GAN performance is similar to the performance of optimal P&C Rand on these challenging datasets. Notably, the performance of LongT GAN on motion capture datasets (CMU MoCap[1], HDM05 [1] and Berkely MHAD[2]) is much higher than the performance on the three datasets that we consider. Indeed the latter are markerless (noisy), include more classes and captured from different views. We believe the reason for the drop in performance of LongT GAN is due to training the decoder with ground truth mask inputs. This allows the decoder to perform the prediction well (the main purpose of LngT-GAN system), however, the encoder does not learn the features needed for separating different types of actions.

This is the reason that we incorporate training strategies based on weakening the decoder, i.e. FW and FS. Such strategies applied to our P&C network enhance the features learned by the encoder and achieve higher recognition performance in clustering and recognition.

3 Detailed tables for P&C system performance and limitations

In addition, we include all the different variations of P&C performance on the three datasets in Table 3. In the main manuscript we have included a subset of these due to space limits (P&C Rand, P&C FS-AEC, P&C FW-AEC). It can be observed that our training strategies succeed to improve performance from the baseline for up to 20%. In terms of limitations, it can be seen that lowest improvement in performance is on the largest dataset NTU RGB-D (60 Classes) cross subject test. This is due to variations of the movement from subject to subject that the network is unable to fully capture simply from keypoints sequences alone. Incorporating constraints that dictate particular arrangement of keypoints (e.g. skeleton graph [3]) could be a future enhancement of the system for cross subject action recognition.

Method	**NW-UCLA** (%)
Unsupervised Skeleton	
P&C Rand	72.0
P&C no FS	81.8
P&C no FW	82.9
P&C FS	82.3
P&C FW	83.6
P&C FS-AEC	**83.8**
P&C FW-AEC	**84.9**

Method	**UWA3D**	
	V3 (%)	V4 (%)
Unsupervised Skeleton		
P&C Rand	48.5	51.5
P&C no FS	54.6	60.3
P&C no FW	57.1	60.3
P&C FW	58.7	62.3
P&C FS	58.7	63.0
P&C FS-AEC	**59.5**	**63.1**
P&C FW-AEC	**59.9**	**63.1**

Method	**NTU RGB-D 60**	
	CV(%)	CS(%)
Unsupervised Skeleton		
P&C Rand	56.4	39.6
P&C no FS	74.0	47.3
P&C no FW	74.2	50.4
P&C FW	75.2	50.3
P&C FS	75.3	49.2
P&C FW-AEC	**76.1**	**50.7**
P&C FS-AEC	**76.3**	**50.6**

Table 3: Comparison of action recognition performance of our P&C system.

References

[1] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn, June 2007.

[2] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 53–60. IEEE, 2013.

[3] 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 Conference on Computer Vision and Pattern Recognition*, pages 1227–1236, 2019.

[1] http://mocap.cs.cmu.edu

[4] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2649–2656, 2014.

[5] 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.