Supplementary Material– PECoP: Parameter Efficient Continual Pretraining for Action Quality Assessment

Amirhossein Dadashzadeh[1], Shuchao Duan[1], Alan Whone[2], Majid Mirmehdi[1]
[1]School of Computer Science [2]Translational Health Sciences
University of Bristol, UK
{a.dadashzadeh, shuchao.duan, alan.whone, m.mirmehdi}@bristol.ac.uk

In this supplementary document, we provide additional details about the PD4T dataset and present further experiments and evaluations using another AQA baseline, TPT [1], on the PD4T dataset.

1. More details on PD4T

The number of videos (#video) for each score of this dataset, as well as the minimum/maximum number of frames (#min/#max) for each task can be seen in Table 1.

For **gait** analysis, the patients were asked to walk 10 metres at a comfortable pace and then return to their starting point. For **hand movement**, the patients opened and closed each of their hands (separately) 10 times, as fully and as quickly as possible. For **finger tapping** each patient had to tap their index finger on their thumb 10 times quickly while spanning the amplest range possible. For **leg agility**, while seated, the patient was asked to raise their foot high and stomp on the ground repeatedly and quickly for 10 times.

Score		Normal (0)	Slight (1)	Mild (2)	Moderate (3)	Severe (4)
Gait	#video	196	158	64	8	0
	#min	325	580	421	664	-
	#max	980	1866	13428	10688	-
Finger tapping	#video	152	465	164	23	2
	#min	129	129	129	162	159
	#max	450	724	853	398	460
Hand movem.	#video	234	407	179	23	5
	#min	131	136	150	197	220
	#max	334	571	717	648	648
Leg agility	#video	407	376	54	11	3
	#min	129	135	155	273	345
	#max	513	427	686	504	435

Table 1. The PD4T dataset summary, categorized by severity scores. For each of the four motor tasks the table lists the total number of videos (#video), the minimum (#min) and maximum (#max) number of frames for the respective task.

2. Temporal Parsing Transformer [1]

The Temporal Parsing Transformer (TPT) is a recent SOTA AQA method based on transformers [1]. Unlike existing AQA methods that focus on holistic video representations for score regression, TPT decomposes the video into temporal segments (part-level representations) to extract features. Such a decomposition is critical to TPT's learning process to capture the possible phases of a typical AQA action, e.g. a diving action which contains several key parts, such as approach, take off, flight, etc.

We evaluate the performance of TPT[1] on our PD4T dataset with and without PECoP. To this end, we first split the input video into 5 overlapping clips, and feed each clip into our continually pretrained I3D backbone to get clip level feature representations. Then, TPT is used to convert these representations into temporal part-level representations. Finally, a part-aware contrastive regressor (following [2]) computes part-wise relative representations and fuses them to perform the final relative score regression.

As shown in Table 2, PECoP significantly boosts the performance of TPT across the various actions in PD4T. We note that, the performance of TPT is significantly lower than CoRe and USDL on PD4T tasks (See Table 4 in the main paper). We believe this may be attributed to the substantial degree of action repetition (e.g. in finger tapping or leg agility). In such cases, TPT's part-level representations, as opposed to a more holistic representation, cannot provide enough discriminative information for its learning process and hence TPT's part-level representations do not necessarily align well with some AQA tasks, such as those in PD4T.

Method	Gait	Finger tapping	Hand movem.	Leg agility	Avg. $\mathcal{S}$
TPT	77.80	36.05	47.80	46.27	51.98
TPT + PECoP	79.90	40.73	51.07	50.38	**55.52**

Table 2. Spearman Rank Correlation results on the PD4T dataset with TPT as the baseline.

[1]To train and evaluate TPT we used the code provided in https://github.com/baiyang4/aqa_tpt

References

[1] Yang Bai, Desen Zhou, Songyang Zhang, Jian Wang, Errui Ding, Yu Guan, Yang Long, and Jingdong Wang. Action quality assessment with temporal parsing transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 422–438. Springer, 2022. 1

[2] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7919–7928, 2021. 1