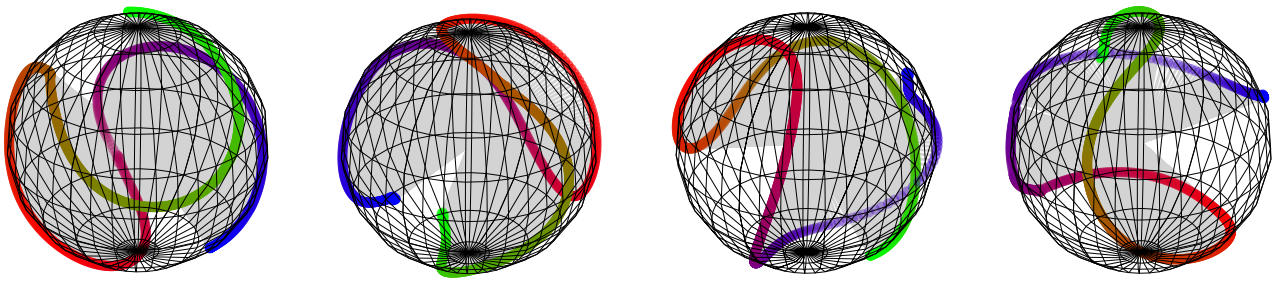
Coherent Temporal Synthesis for Incremental Action Segmentation Supplementary Material

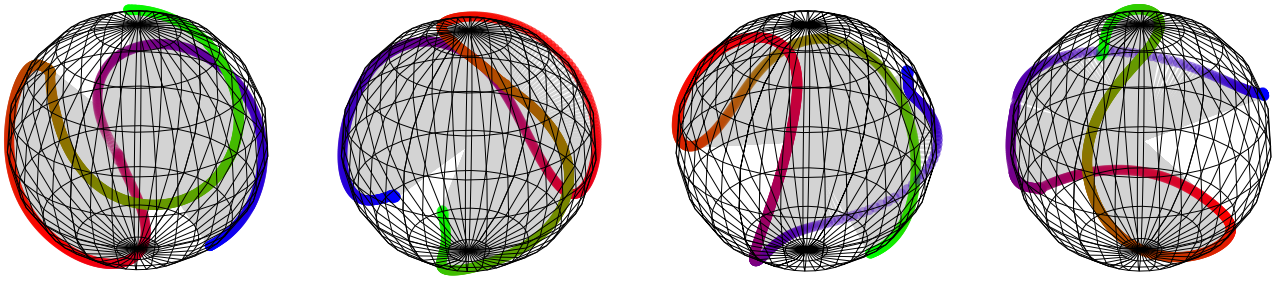
Guodong Ding, Hans Golong and Angela Yao
National University of Singapore
{dinggd, hgolong, ayao}@comp.nus.edu.sg

# Tasks	MSTCN [1]					ASFormer [3]				
	Acc	Edit	F1 @ {10, 25, 50}			Acc	Edit	F1 @ {10, 25, 50}		
Breakfast										
10	29.4±2.1	25.9±2.0	26.3±2.1	23.5±2.1	17.7±1.9	34.2±1.9	32.4±1.3	33.1±1.5	30.1±1.4	23.4±1.2
5	54.5±3.1	49.4±2.2	51.1±2.9	46.9±2.4	37.7±2.6	57.2±3.5	56.8±3.4	58.3±3.6	54.0±3.8	43.6±4.0
Breakfast Blurry										
10	38.5±2.8	43.3±2.5	44.9±2.9	39.5±2.9	29.7±2.9	44.2±2.7	51.2±2.5	53.0±2.2	47.6±3.2	36.8±3.7
YouTube Instructional										
5	30.2±6.7	25.0±3.9	21.9±2.6	18.5±2.3	11.1±1.8	25.2±4.0	20.9±3.0	20.1±2.3	17.5±1.5	11.4±1.7

Table A. The performance means and variations over multiple random seeds for the incremental temporal action segmentation. The variations are denoted in blue.



(1) '*pour milk*' in "*cereal*" (2) '*pour milk*' in "*cereal*" (3) '*pour milk*' in "*pancake*" (4) '*take knife*' in "*juice*"



(5) '*pour milk*' in "*cereal*" (6) '*pour milk*' in "*cereal*" (7) '*pour milk*' in "*pancake*" (8) '*take knife*' in "*juice*"

Figure A. More segment visualizations with T-SNE [2].

	Blurry	# Classes
Breakfast	✗	84
	✓	48
YouTube Instructional	✗	50

Table B. Total class numbers for two benchmarks used in our experiments.

Multiple Runs. We adopt the approach of prior incremental learning studies by initializing the task sequence with multiple random seeds. We utilized five specific random seeds throughout our experiments: $\{42, 123, 1000, 1993, 2023\}$. The performance results presented in the Main Paper reflect the average, while we provide the variations across runs in Tab. A.

Total Classes. In Tab. B, we list the total number of action classes present in our experiments. Breakfast comprises a total of 84 actions without permitting overlaps, whereas there are 48 actions when allowing for overlapping actions. YouTube Instructionals encompasses 50 actions that do not overlap.

Segment Visualization. We present additional generated segments in Fig. A. Specifically, Figs. A(5) and A(6) depict the action of '*pour milk*' generated by our TCA model for "*cereal*" using two different latent variables z, highlighting the diversity in trajectories. Fig. A(7) illustrates the same '*pour milk*' action in a distinct activity "*pancake*". Furthermore, Fig. A(8) shows a segment of '*take knife*' in "*juice*". All these visualized segments indicate the temporal coherence between the frame features.

Dim	Acc	Edit	F1 @ {10, 25, 50}		
128	40.5	45.1	46.3	40.9	31.5
256	41.8	45.0	47.0	41.5	32.0
512	40.8	45.9	46.9	42.1	32.4

Table C. Different latent space sizes for TCA on Breakfast.

Iter.	Acc	Edit	F1 @ {10, 25, 50}		
2,500	41.8	45.0	47.0	41.5	32.0
5,000	42.2	45.2	46.9	41.7	31.8
7,500	38.8	44.4	45.5	40.4	30.9

Table D. Training epochs for TCA.

TCA Latent Space Sizes. We varied the latent space sizes in our TCA model and presented the outcomes in Tab. C. The performance in incremental learning appears quite consistent across various latent space sizes. Notably, a larger dimension (512) only marginally improves performance (by less than 1%) compared to the smaller size (128).

TCA Training Epochs. Tab. D presents TCA's performance across various training epochs. Increasing the training epoch to 5,000 yields a slight performance improvement. However, as the epoch extends to 7,500, a 3% decline in Acc and 1% in segmental metrics occurs.

References

[1] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *CVPR*, 2019. 1

[2] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 1

[3] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. In *BMVC*, 2021. 1