Supplementary Materials of
Leveraging Spatio-Temporal Dependency for Skeleton-Based Action Recognition

1. Additional Experimental Results

Every experimental result of NTU-RGB+D 60 [3], 120 [2], Kinetics-Skeleton [1] and Northwestern-UCLA [4] datasets are shown in Tab. 1 and Tab. 2. By applying our ensemble method, each individual model with different dilation scaling factor σs and streams contributes to achieving the best performance.

Stream	factor σ	NTU-RGB+D 60		NTU-RGB+D 120	
		X-Sub (%)	X-View (%)	X-Sub (%)	X-Set (%)
Joint	1	91.0	96.0	86.2	87.9
	2	90.9	96.1	86.1	88.0
	3	90.9	96.2	86.0	87.7
Bone	1	91.1	95.6	87.2	88.6
	2	91.2	95.3	87.1	88.7
	3	91.3	95.5	87.0	88.9
Ensemble		**93.3**	**97.3**	**90.2**	**91.7**

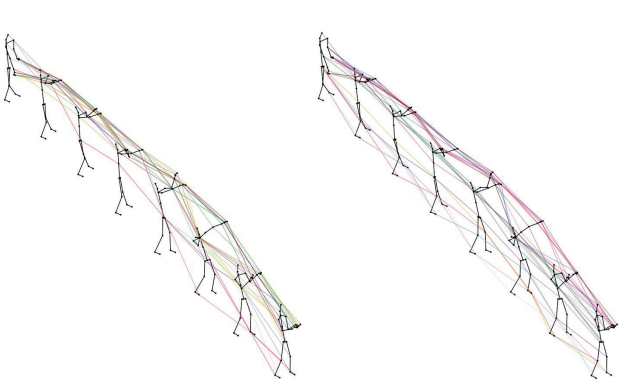
Table 1. Experimental results of NTU-RGB+D 60 and 120 datasets according to data streams and dilation scaling factor σ of our spatial module.

Stream	factor σ	Kinetics-Skeleton		Northwestern
		Top-1 (%)	Top-5 (%)	UCLA (%)
Joint	1	37.9	61.0	94.8
	2	37.5	60.8	95.7
	3	37.2	60.3	95.3
Bone	1	37.4	60.3	95.0
	2	37.5	60.5	94.0
	3	37.2	60.1	93.8
Ensemble		**41.2**	**64.2**	**97.4**

Table 2. Experimental results of Kinetics-Skeleton and Northwestern-UCLA datasets according to data streams and dilation scaling factor σ of our spatial module.

2. Additional Ablation Study

To verify the superiority of our STC module, we conduct several additional experiments for ablation study. Firstly, we compare the model with our curves to the model with straight lines to prove that increasing spatio-temporal receptive field via STC module is technically effective. The straight lines are implemented by restricting the value of k for k-NN to 1. As shown in Tab. 3, although the utilization of curves leads to higher performance of the network, even the models A and E that use straight lines outperforms the baseline model as the straight lines utilize all frames at once, resulting in a large temporal receptive field. In addition, we have mentioned in our main paper that choosing structurally identical node v_{t+1} during inter-frame k-NN hinders the model's ability to capture diverse curves. To prove it quantitatively and qualitatively, we compare the performance of the model including curves without v_{t+1} and the model including curves with v_{t+1}, and visualize them in Fig. 1 ("throw" class). As shown in Tab. 3, the model C and G respectively outperforms the model B and F. Furthermore, referring to Fig. 1 (a) and (b), it can be observed that the curves that do not include v_{t+1} tend to point towards the optimal nodes, namely the hands and arms, whereas the



(a) Curves without v_{t+1} (b) Curves with v_{t+1}

Figure 1. Visualizations for comparing the curves without v_{t+1} and the curves with v_{t+1}

Method	DK-GC	Curve Type			NTU-RGB+D 120	
		Curve w/o v_{t+1}	Curve with v_{t+1}	Straight Line	X-Sub (%)	X-Set (%)
Baseline (+ M)					84.8	86.6
A				✓	85.4	87.2
B			✓		85.7	87.5
C		✓			85.9	87.7
D	✓				85.8	87.3
E	✓			✓	85.9	87.7
F	✓		✓		**86.2**	87.8
G	✓	✓			**86.2**	**88.0**

Table 3. Comparison of the STC module variants. M: motion data

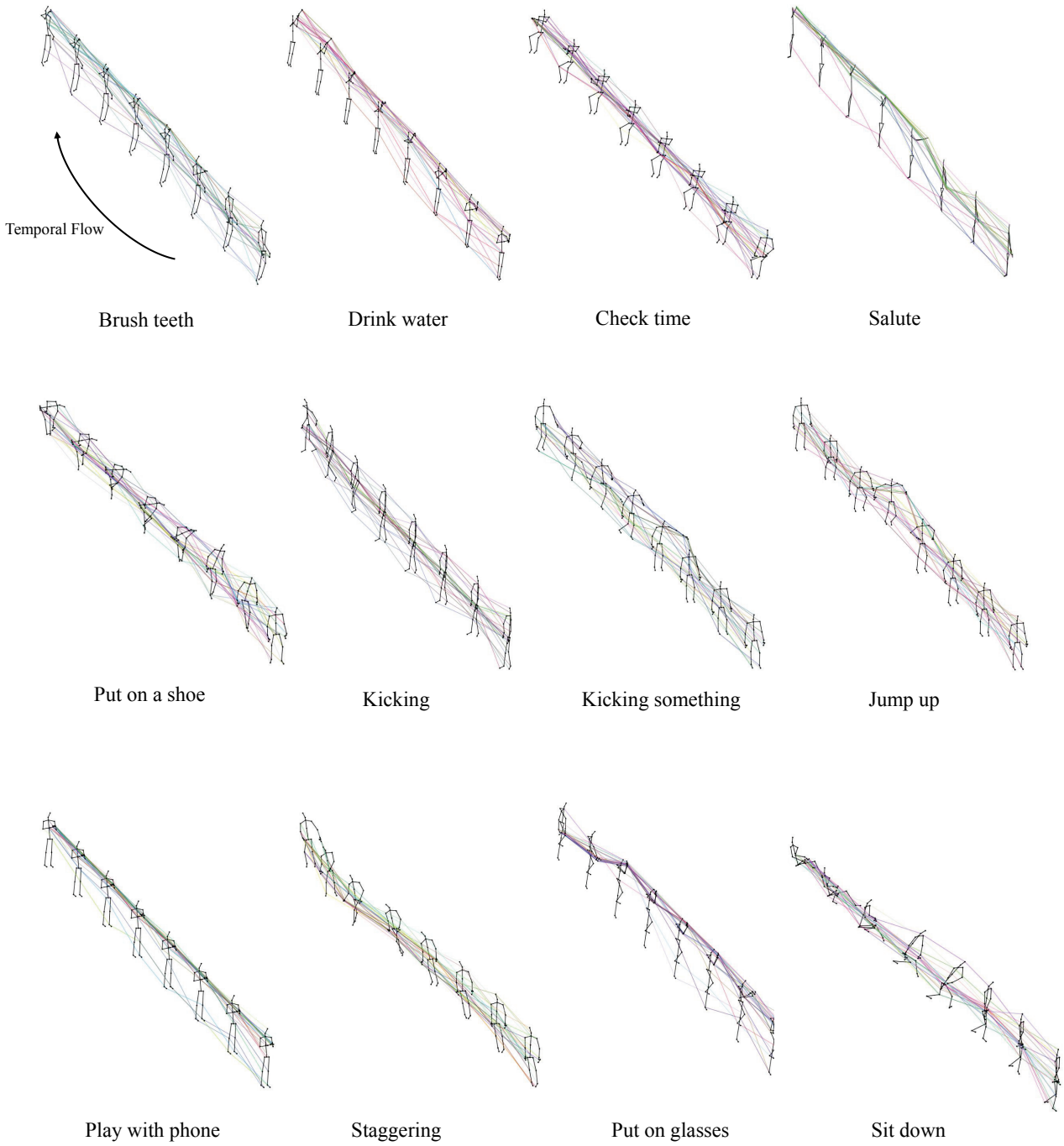


Figure 2. Visualization of trained curves for several data samples.

curves that do include v_{t+1} also point towards the hands and arms but also exhibit a tendency to point towards sub-optimal nodes (e.g., knees and hips).

3. Qualitative Results of Inter-Frame Curves

We propose the Spatio-Temporal Curve (STC) module to identify spatio-temporal dependencies of the human skeleton. Additional qualitative results of the module are shown

in Fig. 2. As mentioned in our main paper, the curves start from every node in the first frame and tend to proceed toward the primary joints for each sequence. Inspired by human visual recognition, it is reasonable to highlight hand gestures for the "Brush teeth", "Drink water", and "Salute" classes, and leg gesture for the "Put on the shoe", "Kicking", and "Kicking something" classes.

References

[1] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[2] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2019. 1

[3] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019, 2016. 1

[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. 1