Supplement Material

Lei Shi[1,2,3] Yifan Zhang[1,3*] Jian Cheng[1,3] Hanqing Lu[1,3]
[1]NLPR & AIRIA, Institute of Automation, Chinese Academy of Sciences
[2]Meituan
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
shilei53@meituan.com, {yfzhang, jcheng, luhq}@nlpr.ia.ac.cn

1. Dataset

We perform extensive experiments on three challenging datasets, namely, NTU-60, NTU-120 and SHREC.

NTU-60: NTU-60 [3] is the most widely used in-door-captured dataset for skeleton-based action recognition. It contains 56,578 action clips in 60 action classes. The clips are performed by 40 volunteers and is captured by 3 KinectV2 cameras with different views. This dataset provides 25 joints for each subject in the skeleton sequences. It recommends two benchmarks: cross-subject (CS) and cross-view (CV), where the subjects and cameras used in the training/test splits are different, respectively.

NTU-120: NTU-120 [2] is larger and more challenging compared with NTU-60. It contains 113,945 action clips in 120 action classes. The clips are performed by 106 volunteers in 32 camera setups. It recommends two benchmarks: cross-subject (CS) and cross-setup (CE). Cross-subject is the same with NTU-60. Cross-setup means using samples with odd setup IDs for training and others for testing.

SHREC: SHREC [1] is a widely used dataset for skeleton-based human hand gesture recognition. We use it to show the generalizability of the proposed method for different types of skeleton data. It contains 2800 gesture sequences performed 1 and 10 times by 28 participants in two ways: using one finger and the whole hand. This dataset provides 22 joints for hands in the skeleton sequences. It splits the sequences into 1960 train sequences and 840 test sequences. The length of sample gestures ranges from 20 to 50 frames. It has two benchmarks: recognizing 14 rough gesture categories (14G) and recognizing 28 fine-grained gesture categories (28G).

2. Implement Details

We provide three ($L = 3$) kinds of joint sets for policy network. The number of joints in these sets are $\{1, 9, 25\}$ for NTU-60/120 and $\{1, 11, 22\}$ for SHREC. Fig. 1 shows the initial stage of the three kinds of joint sets for NTU-60/120 and SHREC.

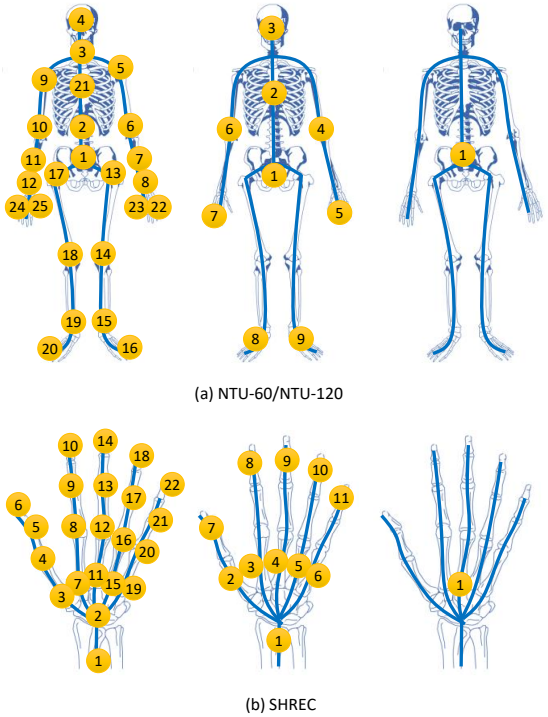

Figure 1. Illustration of the three kinds of joint sets for NTU-60/120 (a) and SHREC (b).

We also design two ($K = 2$) kinds of models (SM_0 and SM_1) with different sizes. The spatial module (SM) of the SGN [4] occupies most of the computations. It consists of three graph convolutional layers, where the input/output channels are 128/128, 128/256 and 256/256. We denote the original SM as SM_0. To provide a choice that has less GFLOPs, we propose the SM_1, which consists of only one graph convolutional layer. The input/output channels of the graph convolutional layer are set to 128/256 to keep consistent with other layers.

*Corresponding Author

For policy network, we use a single temporal convolutional layer with the kernel size set to 3. We also tried using LSTM and Self-Attention module for policy network, but the temporal convolutional layer shows better performance. Other details of the network architecture are the same with the original SGN.

When training, we first pretrain the single models with different number of joints. The training scheme is the same with the original SGN, except for that we freeze the transform matrix in the beginning 30 epochs. In detail, the initial learning rate is 0.001, and is divided by 10 at the 60_{th}, 90_{th} and 110_{th} epochs. Training is finished at the 120_{th} epochs. Adam optimizer is used with $\beta = [0.9, 0.999]$. Batch size is 64. Label smoothing is used with smooth rate 0.1. Weight decay is 0.0001. For data preprocessing, 20 frames are sampled for each action, which is randomly sampled for training and uniformly sampled for test.

3. GFLOPs v.s. Accuracy.

We plot the accuracy-efficiency trade-off curves for NTU-120 and SHREC on Figure 2 and Figure 3, respectively. They show consistent results with the curve of NTU-60 as shown in the main paper.

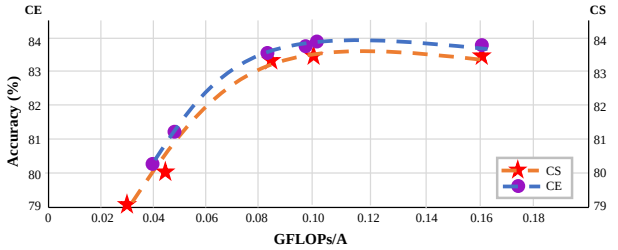

Figure 2. GFLOPs v.s. accuracy for AdaSGN on NTU-120 dataset.

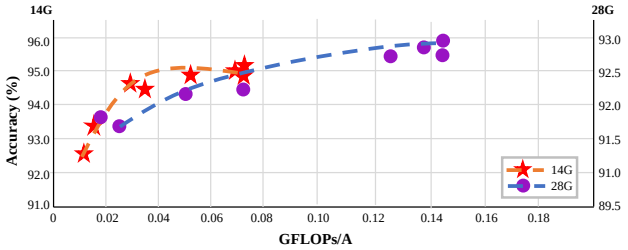

Figure 3. GFLOPs v.s. accuracy for AdaSGN on SHREC dataset.

4. Qualitative Results

Figure 4 shows more qualitative examples from NTU-120 (a, b) and SHREC (c, d). For action "selfie", the action mainly take places from frame 5 to frame 15. So the policy network uses all joints (25-joint) in these frames and ignore other frames (1-joint). For action "sitting down", the beginning frames are ignored. During the 5-20 frames, the policy network uses a part of the joints (9-joint). For gesture "grab", grabbing mainly takes places in frames from 10 to 15. The model costs more computations on these frames. For "swipe right" gesture, it is related to the trajectory of the hand. So the model transform the hand skeleton into a single point to model the global motion of the hand.

References

[1] Quentin De Smedt, Hazem Wannous, Jean-Philippe Vandeborre, Joris Guerry, Bertrand Le Saux, and David Filliat. SHREC'17 Track: 3D Hand Gesture Recognition Using a Depth and Skeletal Dataset. In I. Pratikakis, F. Dupont, and M. Ovsjanikov, editors, *Eurographics Workshop on 3D Object Retrieval*, pages 1–6, 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*, pages 1–1, 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016. 1

[4] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition. pages 1112–1121, 2020. 1

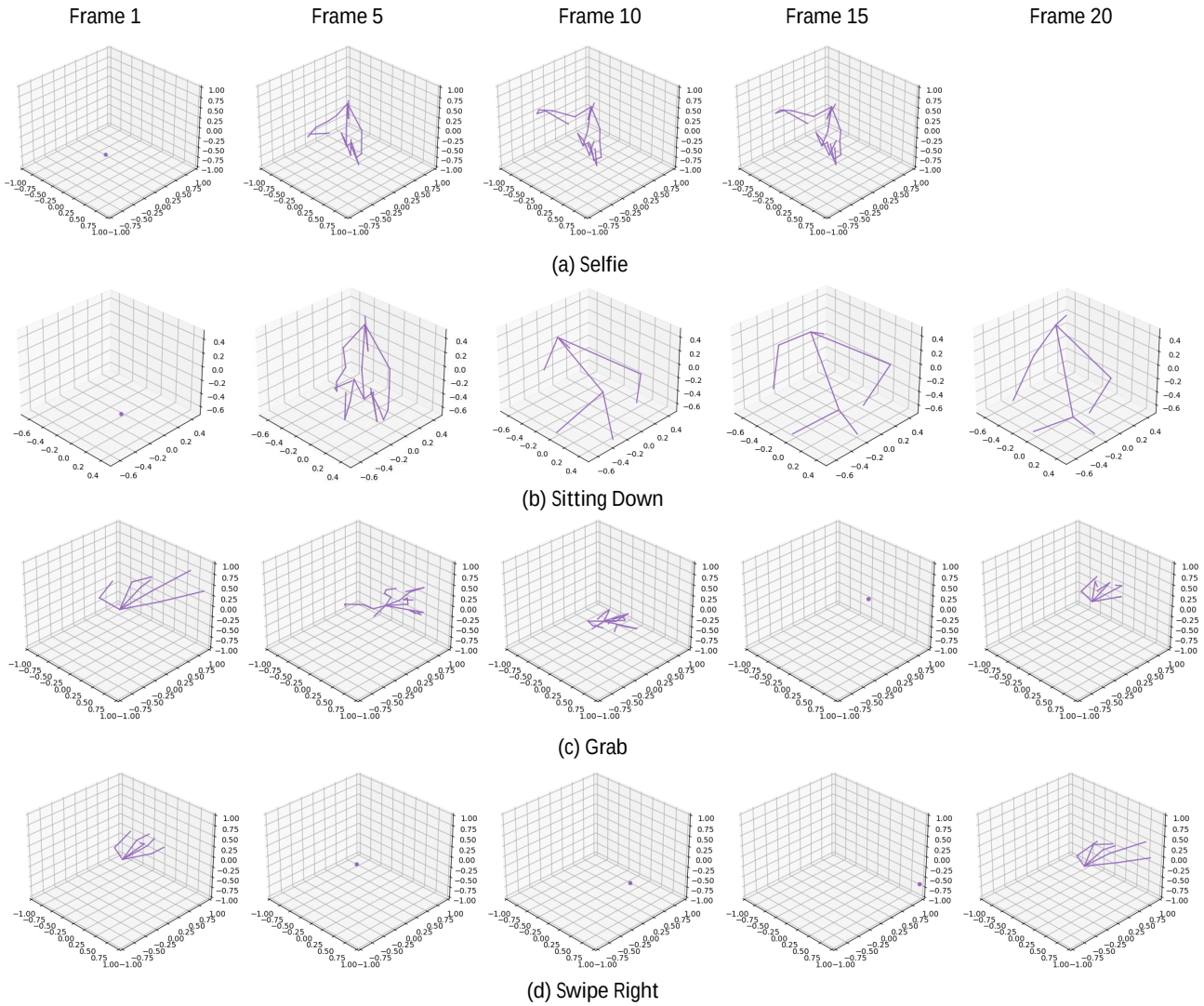

Figure 4. More qualitative examples from NTU-120 (a, b) and SHREC (c, d). Each sample has 20 frames and we show 5 of them evenly. Non-informative skeletons are transformed into less points based on the sample classes where other informative skeletons are kept with the original point numbers.