Exploiting Spatial-temporal Relationships for 3D Pose Estimation via Graph Convolutional Networks Supplementary Document

Yujun Cai[1], Liuhao Ge[1], Jun Liu[1], Jianfei Cai[1,2], Tat-Jen Cham[1], Junsong Yuan[3], Nadia Magnenat Thalmann[1]
[1]Nanyang Technological University, Singapore
[2]Monash University, Australia
[3]State University of New York at Buffalo University, Buffalo, NY, USA
{yujun001, ge0001ao, jliu029}@e.ntu.edu.sg
{asjfcai, astjcham}@ntu.edu.sg, jsyuan@buffalo.edu, nadiathalmann@ntu.edu.sg

In this supplementary document, we provide materials not included in the main paper due to space constraints. Firstly, Section 1 provides more details of our proposed network structures. Next, Section 2 elaborates on our quantitative results on Human3.6M using the MPJPE metric under protocol #1. Finally, Section 3 presents additional qualitative results for comparison.

1. Network Architecture

Figure 1 illustrates the detailed architectures of our proposed GCN unit and the hierarchical local-to-global network. We note that the local-to-global network takes consecutive 2D joint locations with the size of $T \times M_0 \times 2$ as input and the output is the consecutive 3D poses with the size of $T \times M_0 \times 3$. Here T is the input sequence length. M_i denotes the node number of the i-th graph resolution level for each frame, with $M_0 = 17$, $M_1 = 5$, $M_2 = 1$ for 3D body pose estimation and $M_0 = 21$, $M_1 = 6$, $M_2 = 1$ for 3D hand pose estimation, respectively.

For data-processing, the input 2D joint locations are normalized between -1 to 1 based on the size of the input image. We perform horizontal flip augmentations at train and test time. Since we do not predict the global position of the 3d prediction, we zero-centre the 3d poses around the hip joint for human pose estimation and palm joint for hand pose estimation (in line with previous work) .

2. Additional Quantitative Evaluation

2.1. 3D Pose Estimation from Ground Truth 2D joints

In Section 4.4 of the main manuscript, we provide results of 3D pose estimation with the input 2D poses detected from RGB images. For human pose estimation, some previous work additionally reported estimation results using the ground truth 2D coordinates as input. In this section, we follow the evaluation protocol #1 and present our estimation performance in comparison with the previously reported approaches on Human3.6M, where T represents the number of input frames. As shown in Table 1, our method obtains superior results to the competing methods using ground truth 2D joints as input, achieving an error of 37.2mm with 3 input frames.

Protocol #1	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Pavlakos, CVPR18 [5] ($T = 1$)	47.5	50.5	48.3	49.3	50.7	55.2	46.1	48.0	61.1	78.1	51.05	48.3	52.9	41.5	46.4	51.9
Martinez, ICCV'17[6] ($T = 1$)	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
Hossain, ECCV'18 [2] ($T = 5$)	35.7	39.3	44.6	43	47.2	54.0	38.3	37.5	51.6	61.3	46.5	41.4	47.3	34.2	39.4	44.1
Lee, ECCV18 [4] ($T = 3$)	34.6	39.7	37.2	40.9	45.6	50.5	42.0	39.4	47.3	48.1	39.5	38.0	**31.9**	41.5	37.2	40.9
Ours, ($T = 1$)	33.4	39.0	33.8	37.0	38.1	47.3	39.5	37.3	43.2	46.2	37.7	38.0	38.6	30.4	32.1	38.1
Ours, ($T = 3$)	**32.9**	**38.7**	**32.9**	**37.0**	**37.3**	**44.8**	**38.7**	**36.1**	**41.0**	**45.6**	**36.8**	**37.7**	37.7	**29.5**	**31.6**	**37.2**

Table 1. Comparison with the state-of-the-art methods for the Human3.6M under Protocol #1, using ground truth 2D joint locations as input. T denotes the number of input frames used in each method. The best score is marked in **bold**.

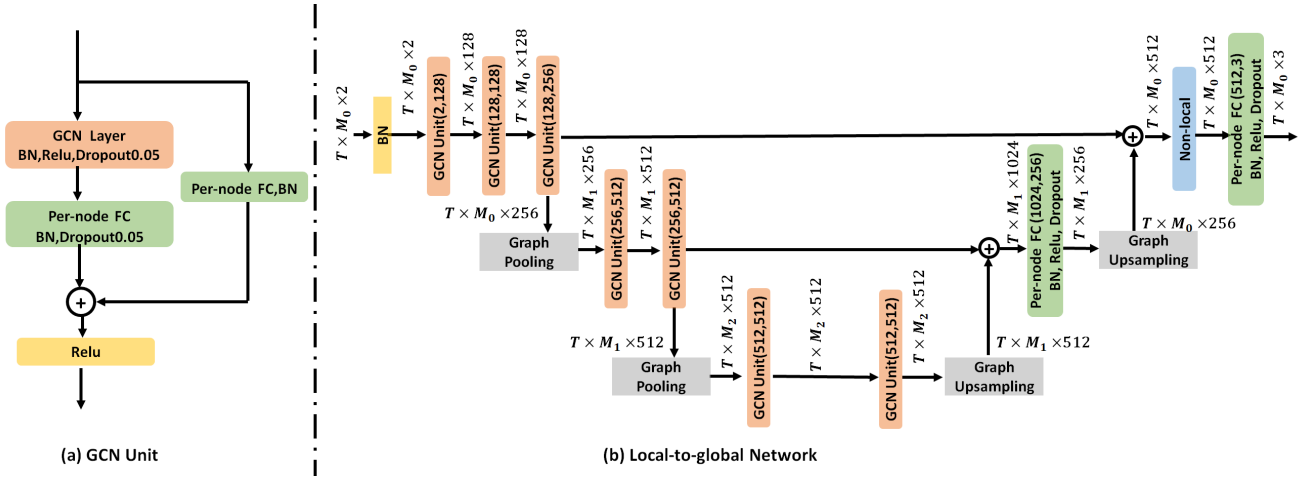

Figure 1. Details of our proposed network architectures. (a) Illustration of GCN Unit. (b) The hierarchical local-to-global network. Here 'BN' is short for batch normalization. T is the input sequence length. M_i denotes the node number of the i-th graph resolution level for each frame, with $M_0 = 17$, $M_1 = 5$, $M_2 = 1$ for 3D body pose estimation and $M_0 = 21$, $M_1 = 6$, $M_2 = 1$ for 3D hand pose estimation, respectively.

3. Additional Qualitative Evaluation

We provide additional qualitative results of the our proposed method under challenging scenarios with various viewpoints and severe self-occlusions. The included images are from Human3.6M dataset[3], STB dataset [7] and MPII dataset [1], as shown in Figure 2.

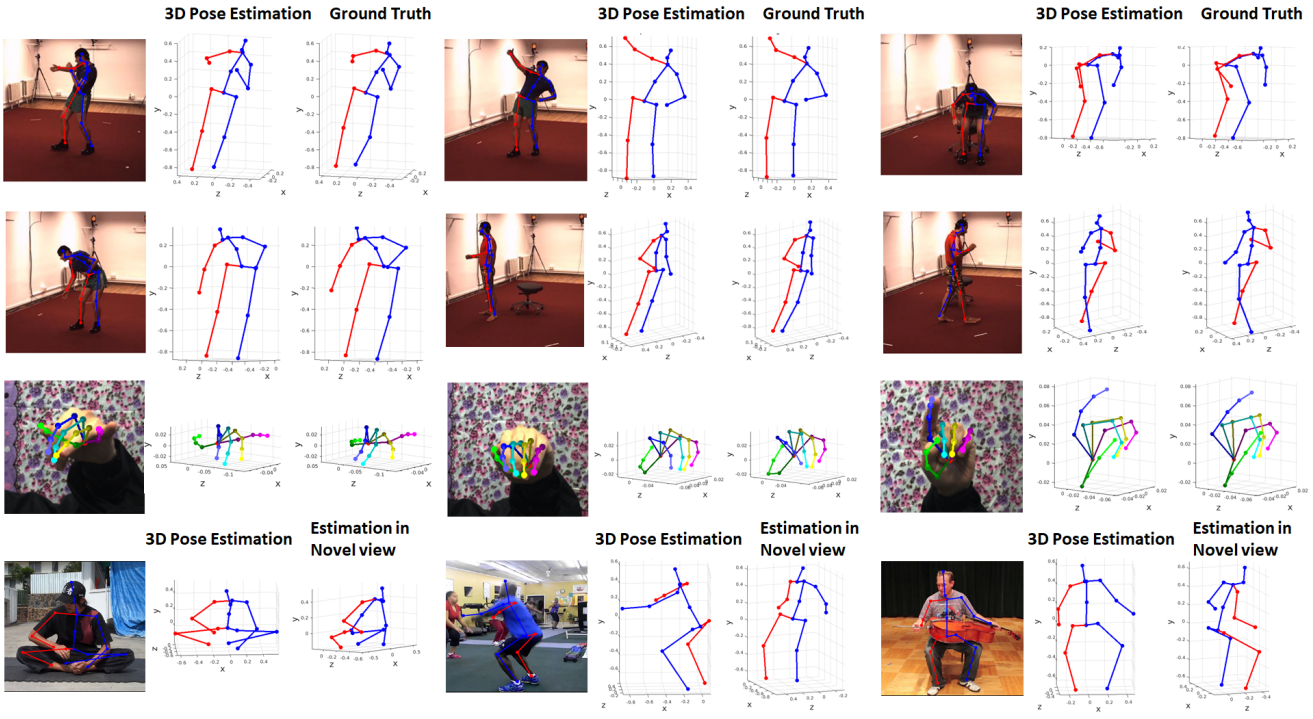

Figure 2. Additional quantitative results of our proposed method on Human3.6M, STB and MPII datasets. The detected 2D joint locations are overlaid with the RGB images. First and second rows: Examples from Human3.6M dataset[3]. Third row: Examples from STB dataset[7]. Forth row: Examples from MPII dataset[1].

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[2] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *European Conference on Computer Vision*, pages 69–86. Springer, 2018.

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

[4] Kyoungoh Lee, Inwoong Lee, and Sanghoon Lee. Propagating lstm: 3d pose estimation based on joint interdependency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–135, 2018.

[5] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7307–7316, 2018.

[6] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2602–2611, 2017.

[7] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. A hand pose tracking benchmark from stereo matching. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 982–986. IEEE, 2017.