Modulated Graph Convolutional Network for 3D Human Pose Estimation (*Supplementary Material*)

Zhiming Zou and Wei Tang*
University of Illinois at Chicago
{zzou6, tangw}@uic.edu

A1. Decouple Self-connections

Previous works [7, 8, 3] found that decoupling the transformations of self-connections and other edges can improve the 3D HPE performance, which we have also observed. Following Liu *et al*. [3], Eq. (2) and Eq. (3) are rewritten as:

$$\mathbf{h}_i' = \sigma(\mathbf{W}\mathbf{h}_i\tilde{a}_{ii} + \sum_{j\in\mathcal{N}_i}\mathbf{V}\mathbf{h}_j\tilde{a}_{ij}) \tag{10}$$

$$\mathbf{h}_i' = \sigma(\mathbf{W}_i\mathbf{h}_i\tilde{a}_{ii} + \sum_{j\in\mathcal{N}_i}\mathbf{V}_j\mathbf{h}_j\tilde{a}_{ij}) \tag{11}$$

where $\mathcal{N}_i$ is the set of neighboring nodes of node i excluding itself, $\mathbf{W} \in \mathbb{R}^{D'\times D}$ and $\mathbf{W}_i \in \mathbb{R}^{D'\times D}$ $(i = 1, \cdots, N)$ are weight matrices for self-connections, and $\mathbf{V} \in \mathbb{R}^{D'\times D}$ and $\mathbf{V}_j \in \mathbb{R}^{D'\times D}$ $(j = 1, \cdots, N)$ are weight matrices for edges between different nodes. Eq. (10) and Eq. (11) are the decoupling versions of the vanilla graph convolution and weight unsharing [3], respectively.

The decoupling version of our weight modulation, *i.e*., Eq. (4), can be formulated as below:

$$\mathbf{h}_i' = \sigma((\mathbf{m}_i \odot \mathbf{W})\mathbf{h}_i\tilde{a}_{ii} + \sum_{j\in\mathcal{N}_i}(\mathbf{m}_j \odot \mathbf{V})\mathbf{h}_j\tilde{a}_{ij}) \tag{12}$$

Note we have reused the symbols defined in the paper.

Putting together updated features of all nodes, Eq. (10) and Eq. (12) can be equivalently rewritten as compact forms:

$$\mathbf{H}' = \sigma(\mathbf{W}\mathbf{H}\tilde{\mathbf{A}}^{self} + \mathbf{V}\mathbf{H}\tilde{\mathbf{A}}^{other}) \tag{13}$$

$$\mathbf{H}' = \sigma((\mathbf{M} \odot (\mathbf{W}\mathbf{H}))\tilde{\mathbf{A}}^{self} + (\mathbf{M} \odot (\mathbf{V}\mathbf{H}))\tilde{\mathbf{A}}^{other}) \tag{14}$$

where $\tilde{\mathbf{A}}^{self}$ and $\tilde{\mathbf{A}}^{other}$ are respectively affinity matrices of self-connections and other edges, and they are normalized [2] separately. We have tried to use two different weight modulation matrices for the self and other connections but do not observe improvement.

*Corresponding author.

Method	Channels	Params	MPJPE	P-MPJPE
w/o weight modu. + w/o affinity modu.	128	0.27M	49.73	39.92
w/ weight modu. + w/o affinity modu.	124	0.27M	42.04	33.25
w/o weight modu. + w/ affinity modu.	128	0.27M	40.53	31.39
w/ weight modu. + w/ affinity modu.	124	0.27M	38.83	30.35
w/o weight modu. + [8]	128	0.27M	43.05	33.43
w/ weight modu. + [8]	124	0.27M	40.98	32.82

Table 1. Ablation study on affinity modulation and weight modulation. The units of MPJPE and P-MPJPE are millimeters (mm). 'w/o affinity modu.' means using a predefined skeleton graph.

A2. Visualization of Affinity Modulation

We visualize affinity matrices learned by different affinity modulation methods in Fig. 1. They are extracted from the first graph convolution layer. We can see that learning a modulation matrix added to the human skeleton graph can include some meaningful relations beyond the natural connections of body joints, *e.g*., left/right hip and left/right ankle in Fig. 1 (f). We visualize different modulation matrices in skeleton graph in Fig. 2. Note that we draw the undirected graph of all the modulation matrices for simplicity.

A3. More Ablation Study

We provide extra ablation study to illustrate the effectiveness of affinity modulation and weight modulation in Tab. 1. We also include results obtained by replacing our affinity modulation with the learnable graph in [8].

A4. More Comparison with State of the Art

In Tab. 2, we compare our Modulated GCN with [6]. Our method has comparable performance with their single-frame model. We would like to point out that the 2D pose correction strategy proposed in [6] is complementary to our method and can also be used to improve the performance of Modulated GCN.

As mentioned in the paper, we set the channels to 384 to handle the detection errors for 2D pose detections. The model size is 2.87M. By comparison, the model sizes of

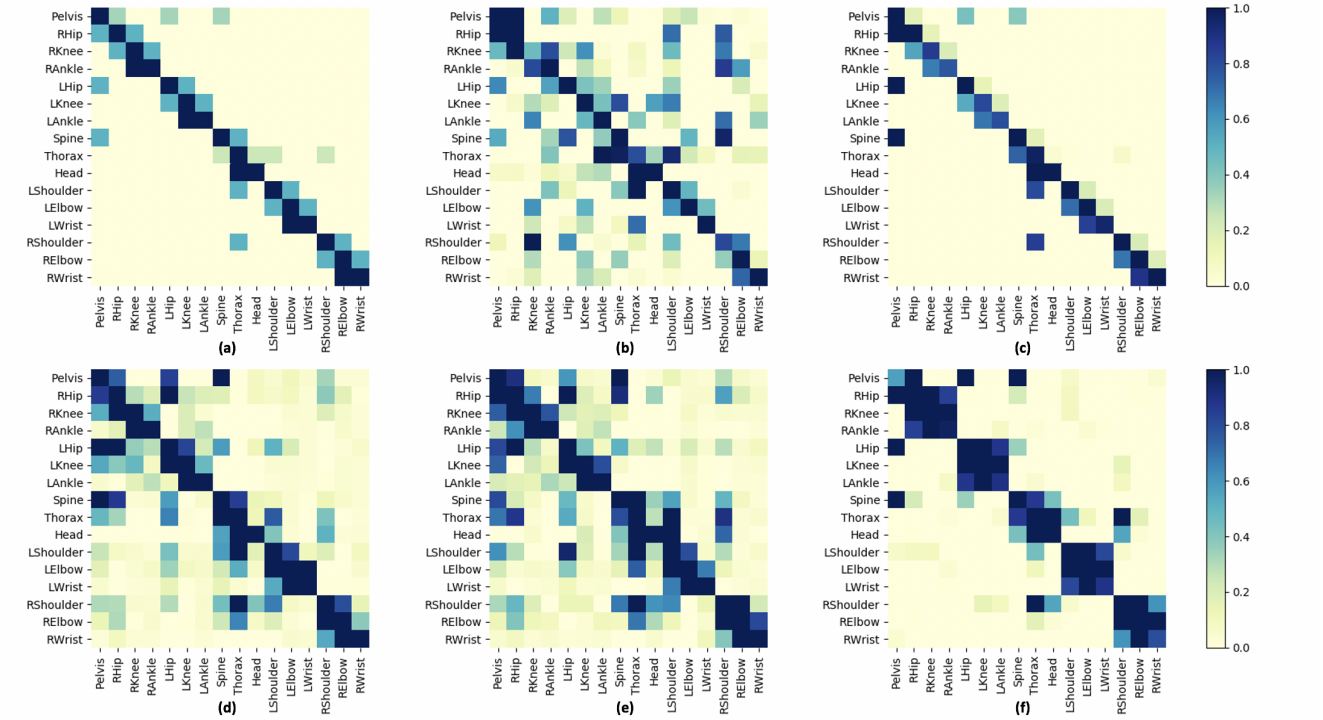

Figure 1. Visualization of affinity modulation: (a) $\mathbf{A}_{skeleton}$, (b) $\mathbf{A}_{no-skeleton}$, (c) $\mathbf{A}_{mul}$, (d) $\mathbf{A}_{mix}$, (e) $\mathbf{A}_{add}$, and (f) $\mathbf{A}_{add}$ with symmetry regularization.

[1], [5] (single-frame), [4] are respectively 2.92M, 4.29M and 4.29M.

A5. More Qualitative Results

Fig. 3 shows some extra visualization results obtained by our Modulated GCN on Human3.6M. We find that most of the failure cases occur when the 2D detector fails to predict accurate 2D human poses due to severe self-occlusion. However, our Modulated GCN can still generate plausible 3D poses with respect to the estimated 2D poses. Fig. 4 provides some qualitative results obtained by our Modulated GCN on the wild images. These image frames are extracted from different YouTube videos. Our model achieves satisfactory results on the unseen scenes. This indicates that our model generalizes well to unseen actions and datasets.

References

[1] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2262–2271, 2019.

[2] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[3] Kenkun Liu, Rongqi Ding, Zhiming Zou, Le Wang, and Wei Tang. A comprehensive study of weight sharing in graph networks for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[4] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017.

[5] Dario PavlloF, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 899–908, 2020.

[7] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[8] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3425–3435, 2019.

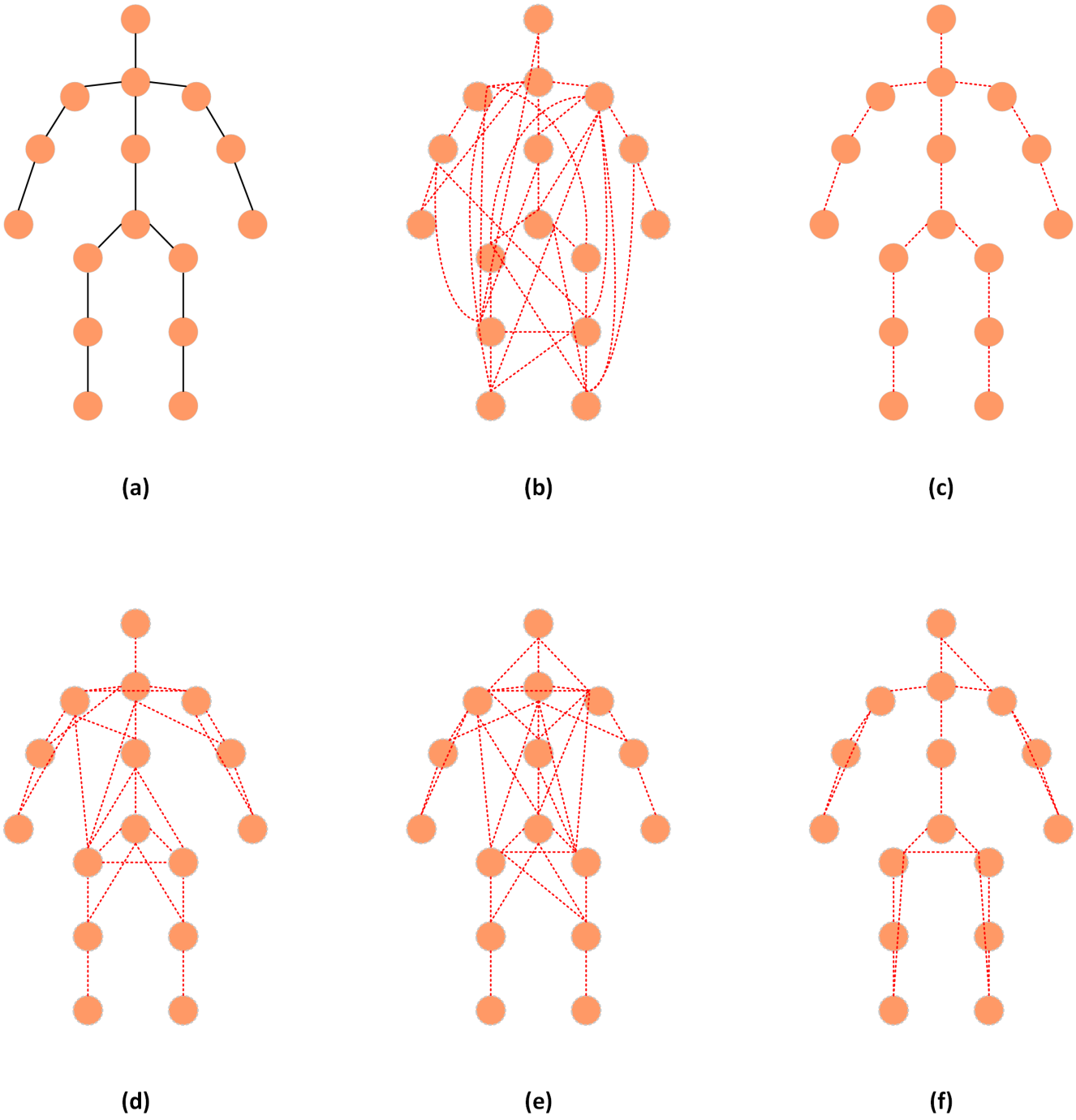

Figure 2. Visualization of affinity modulation in skeleton graph: (a) $\mathbf{A}_{skeleton}$, (b) $\mathbf{A}_{no-skeleton}$, (c) $\mathbf{A}_{mul}$, (d) $\mathbf{A}_{mix}$, (e) $\mathbf{A}_{add}$, and (f) $\mathbf{A}_{add}$ with symmetry regularization. The threshold is set as 0.5.

Method	Dire.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Xu *et al*. [6] (single-frame)	40.6	47.1	45.7	46.6	50.7	63.1	45.0	47.7	56.3	63.9	49.4	46.5	51.9	38.1	42.3	**49.2**
Ours	45.4	49.2	45.7	49.4	50.4	58.2	47.9	46.0	57.5	63.0	49.7	46.6	52.2	38.9	40.8	49.4

Table 2. Quantitative comparisons on Human3.6M under Protocol #1. Errors are in millimeters.

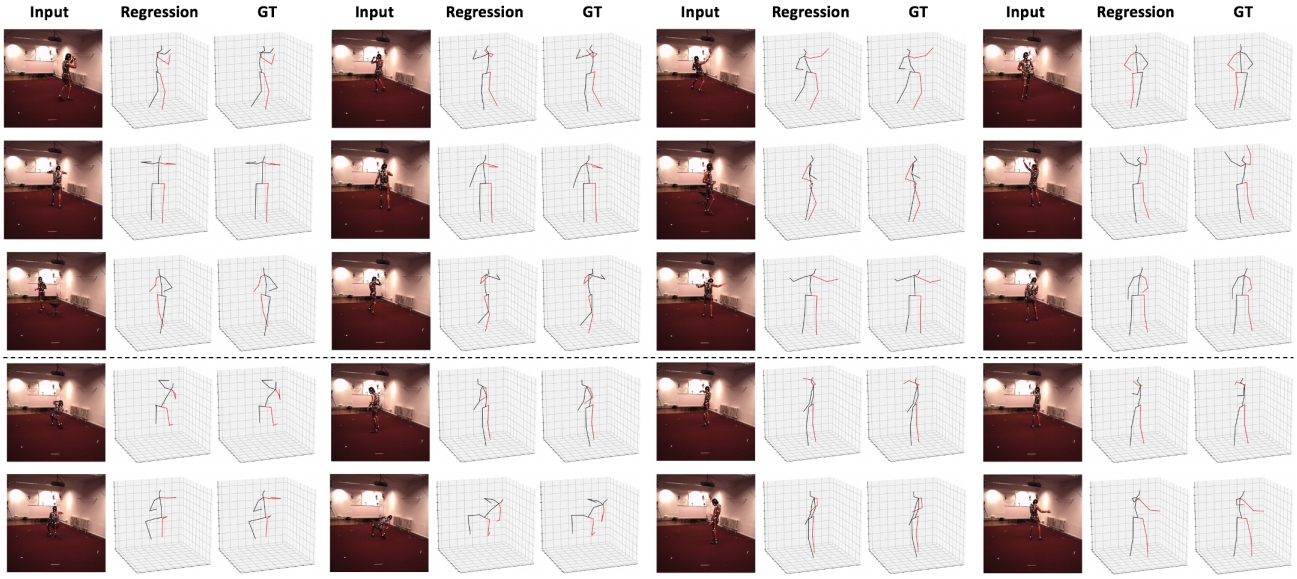

Figure 3. Qualitative results obtained by our Modulated GCN on the Human3.6M dataset. The last two rows are failure cases.

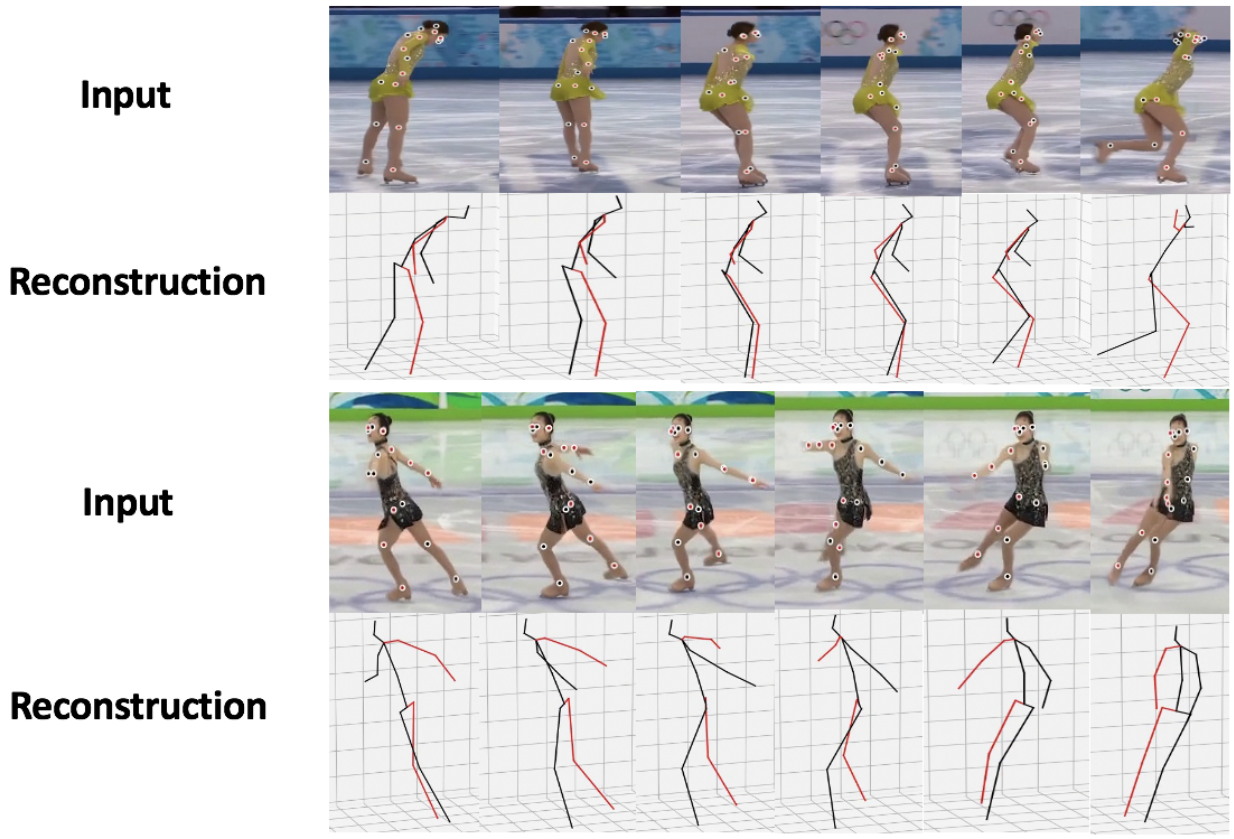

Figure 4. Qualitative results obtained by our Modulated GCN on the wild images