Spatio-Temporal MLP-Graph Network for 3D Human Pose Estimation

Tanvir Hassan
tanvirhassan970@gmail.com
A. Ben Hamza
hamza@ciise.concordia.ca
Concordia University
Montreal, QC, Canada

This supplementary material includes more detailed descriptions of the datasets, and additional experimental results.

1 Datasets and Implementation Details

Human3.6M is a large-scale dataset containing more than 3.6 million human poses, and includes 15 different human activities performed by 11 actors [3]. During training, we use 5 subjects (S1, S5, S6, S7, S8), and during testing, we use 2 subjects (S9, S11) from the dataset.

MPI-INF-3DHP contains 1.3 million frames and features 8 actors performing 8 actions, providing a wider range of poses [8]. It includes a test set with 6 subjects in both indoor and complex outdoor scenes, enabling the evaluation of the model's generalization ability to unseen environments.

More Implementaion Details. All experiments are conducted on a single NVIDIA GeForce RTX 3070 GPU with 8G memory, and our model is implemented in PyTorch. For the 2D ground truth, we set the batch size to 256, $L = 3$, $F = 128$, and $R = 256$. To prevent overfitting, we also add dropout with a factor of 0.2 after each graph weighted Jacobi layer.

2 Additional Experimental Results

Quantitative Results. Table 1 reports the results of our MLP-GraphWJ mixer model and various competing baselines when using 2D ground truth keypoints as input. The findings indicate that our model outperforms GraphMDN [9] on 12 out of 15 actions with an average error reduction of approximately 2.42% under Protocol #1. Moreover, our model shows better performance compared to MGCN [18], High-Order GCN [19], SemGCN [16], and Weight Unsharing [6] on average, while having a lower number of learnable parameters and inference time. These results highlight the effectiveness of our proposed method.

Qualitative Results. Figure 1 shows some additional visualization results of the proposed MLP-GraphWJ mixer model on the Human3.6M dataset. Our model demonstrates a high degree of accuracy in predicting hand poses, even in scenarios where joints overlap or occlusions occur, while MGCN [18] struggles to perform the same task effectively.

Table 1: Performance comparison of our model and baseline methods on Human3.6M under protocol #1 using the ground truth 2D pose as input. Boldface numbers indicate the best performance, whereas the underlined numbers indicate the second-best performance. (†) - uses temporal information.

	Action															
Protocol #1	Dire.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez *et al.* [0]	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
Pavlakos *et al.* [00]	47.5	50.5	48.3	49.3	50.7	55.2	46.1	48.0	61.1	78.1	51.1	48.3	52.9	41.5	46.4	51.9
Hossain *et al.* [0] (†)	35.7	39.3	44.6	43.0	47.2	54.0	38.3	37.5	51.6	61.3	46.5	41.4	47.3	34.2	39.	44.1
Cai *et al.* [0] (†)	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
Liu *et al.* [0]	36.8	40.3	33.0	36.3	37.5	45.0	39.7	34.9	40.3	47.7	37.4	38.5	38.6	29.6	32.0	37.8
Pavllo *et al.* [00] (†)	35.2	40.2	32.7	35.7	38.2	45.5	40.6	36.1	48.8	47.3	37.8	39.7	38.7	27.8	29.5	37.8
Zou *et al.* [00]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37.4
Oikarinen *et al.* [0]	33.9	39.9	33.0	35.4	36.8	44.4	38.9	33.0	41.0	50.0	36.4	38.3	37.8	28.2	31.5	37.2
Lee *et al.* [0]	34.6	39.6	31.3	34.7	33.9	40.3	39.5	32.2	**35.4**	43.5	34.0	35.0	36.9	29.7	31.4	35.6
Zhang *et al.* [00]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	35.3
Zhao *et al.* [00]	32.0	38.0	**30.0**	34.4	34.7	43.3	**35.2**	31.4	38.0	46.2	34.2	35.7	36.1	27.4	30.6	35.2
Zhan *et al.* [00] (†)	**31.2**	35.7	31.4	33.6	35.0	37.5	37.2	30.9	42.5	**41.3**	34.6	36.5	32.0	27.7	28.9	34.4
Ours (†)	31.6	**35.6**	31.5	**31.0**	**32.1**	**35.1**	36.3	**30.1**	38.8	41.4	**32.6**	**34.6**	**31.4**	**25.5**	**25.8**	**32.9**

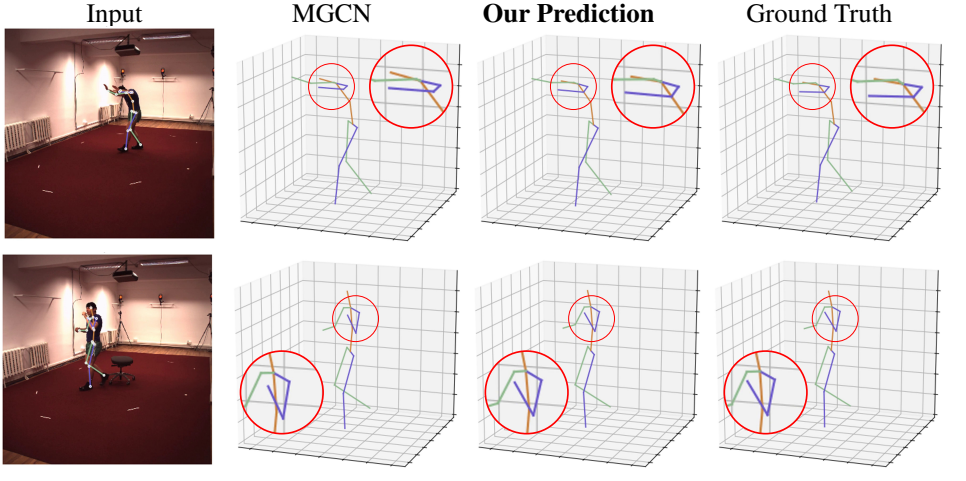

Figure 1: Qualitative comparison between our model and MGCN on Human3.6M actions. The red circle indicates the locations where our model yields better results.

Model Size Comparison. The proposed framework employs a weighted Jacobi (WJ) feature propagation rule obtained via graph filtering with implicit fairing. One of the key benefits of our model is that it presents a simple and competitive alternative to existing approaches that do not use self-attention mechanisms, while outperforming previous work and retaining a small model size, as illustrated in Figure 2. Moreover, our approach effectively merges temporal information within the feature channels, while incurring minimal computational cost in terms of sequence length.

Hyper-Parameter Sensitivity Analysis. We start by investigating the impact of the different hyper-parameters on model performance. Results are reported in Table 2. It can be

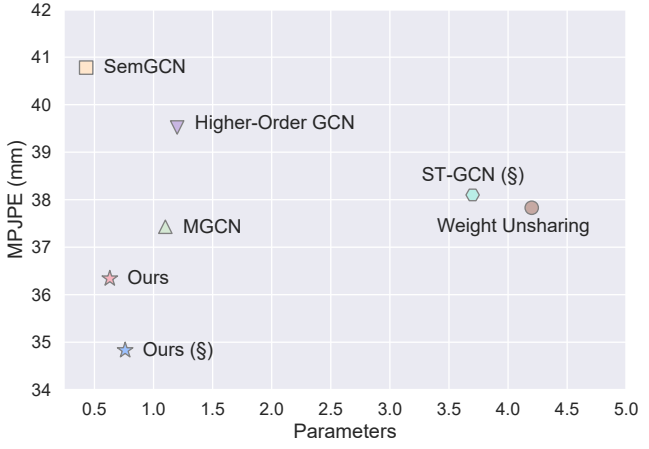

Figure 2: Performance and model size comparison between our model and state-of-the-art methods for 3D human pose estimation, including MGCN [18], SemGCN [16], High-Order GCN [19], ST-GCN [1], and Weight Unsharing [5]. Lower Mean Per Joint Position Error (MPJPE) values indicate better performance. Evaluation conducted on a single frame of Human3.6M [3] dataset with 2D joints as input. (§) - uses a pose refinement network.

observed that the expanding ratio of 2 ($F = 384, R = 768$) performs better than the commonly used ratio of 4 in vision Transformers and MLPs. The value of the skeleton embedding hidden dimension F affects the model ability to capture patterns. When increasing F from 128 to 384 and R from 256 to 768, the MPJPE decreases from 47.5mm to 45.3mm. However, the number of trainable parameters increases from 0.65M to 5.48M. The best results are obtained using $F = 384$, and $R = 768$. Using three MLP-GraphWJ mixer layers yields the best performance, while increasing or decreasing the number of layers negatively impacts performance.

Table 2: Ablation study on various configurations of our approach without pose refinement on Human3.6M under protocol#1 using detected 2D pose as input. L is the number of MLP-GraphWJ mixer layers, F is the hidden dimension of skeleton embedding and joints mixing MLP and R is the hidden dimension of GraphWJ mixing layer. The number of input frames is set to $T = 81$. Boldface numbers indicate the best performance.

L	F	R	Params. (M)	MPJPE (↓)
3	128	256	0.65	47.5
3	256	256	1.28	47.7
3	256	512	2.47	47.9
3	256	1024	4.86	47.3
3	384	384	2.80	46.8
3	384	768	5.48	**45.3**
3	384	1536	10.83	46.1
1	384	768	1.87	48.3
2	384	384	3.68	46.6
4	384	768	7.29	46.6

Comparison with GCN-based Methods. In order to bypass the influence of 2D pose detectors and gain further insight into the importance of our network architecture and graph propagation rule, we train our model on the Human3.6M dataset using 2D ground truth poses by maintaining the expanding ratio of 2 ($F = 128$, $R = 256$) and we report the results in Table 3. Our method demonstrates superior performance compared to recent state-of-art methods based on a single frame, despite utilizing fewer trainable parameters.

Table 3: Performance comparison of our model and baseline methods without pose refinement using ground-truth keypoints. Boldface numbers indicate the best performance.

Method	Filters	Params (M)	MPJPE (↓)	PA-MPJPE (↓)	Infer. Time
SemGCN [16]	128	0.43	40.78	31.46	.012s
High-Order GCN [19]	96	1.20	39.52	31.07	.013s
Weight Unsharing [6]	128	4.22	37.83	30.09	.032s
MGCN [18]	256	1.10	37.43	29.73	.008s
Ours	-	0.63	**36.34**	**28.97**	.005s

Improvements on Hard Poses. Hard poses, which are characterized by high prediction errors, are specific to the model being used. These poses often have certain inherent characteristics, such as overlapping and self-occlusion. The way in which such cases are dealt with, however, may vary across different models [12, 13, 16]. For instance, when a person is sitting down in a position with their legs crossed, estimating their 3D pose accurately can be difficult due to the complex interactions between different body parts. Our proposed method aims to address this challenge by learning to capture the complex relationships between the joints via the joints mixing MLP layer and GraphWJ mixing layer. As reported in the first table of the main paper, our method yields better performance on hard poses (e.g., Directions, Sitting Down, Photo, and Purchase) compared to recent GCN-based state-of-art methods [13, 16, 18]. In addition, we test our model on the top 5% hardest poses following [12, 13], yielding superior performance over the baselines, as shown in Figure 3.

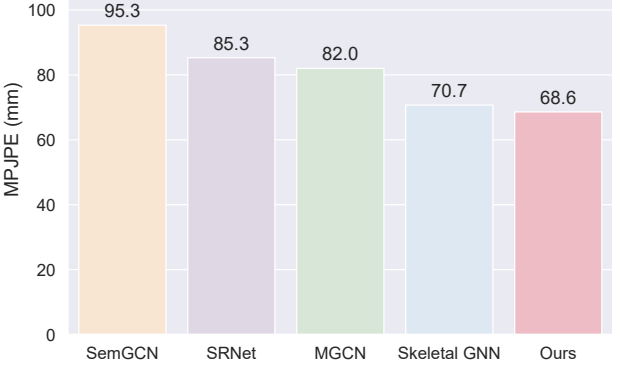

Figure 3: Comparison of our model and baselines on the 5% hardest poses under Protocol #1.

References

[1] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 2272–2281, 2019.

[2] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3D human pose estimation. In *Proc. European Conference on Computer Vision*, pages 68–84, 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, 2013.

[4] Jae Yung Lee and I Gil Kim. Multi-hop modulated graph convolutional networks for 3D human pose estimation. In *Proc. British Machine Vision Conference*, 2022.

[5] Kenkun Liu, Rongqi Ding, Zhiming Zou, Le Wang, and Wei Tang. Comprehensive study of weight sharing in graph networks for 3D human pose estimation. In *Proc. European Conference on Computer Vision*, 2020.

[6] 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 *Proc. European Conference on Computer Vision*, pages 318–334. Springer, 2020.

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

[8] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *Proc. International Conference on 3D Vision*, 2017.

[9] Tuomas Oikarinen, Daniel Hannah, and Sohrob Kazerounian. GraphMDN: Leveraging graph structure and deep learning to solve inverse problems. In *Proc. IEEE International Joint Conference on Neural Networks*, pages 1–9, 2021.

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

[11] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019.

[12] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Ching-Feng Lin. SRNet: Improving generalization in 3D human pose estimation with a split-and-recombine approach. In *Proc. European Conference on Computer Vision*, 2020.

[13] Ailing Zeng, Xiao Sun, Lei Yang, Nanxuan Zhao, Minhao Liu, and Qiang Xu. Learning skeletal graph neural networks for hard 3D pose estimation. In *Proc. IEEE International Conference on Computer Vision*, pages 11436–11445, 2021.

[14] Yu Zhan, Fenghai Li, Renliang Weng, and Wongun Choi. Ray3D: Ray-based 3D human pose estimation for monocular absolute 3D localization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 13116–13125, 2022.

[15] Zijian Zhang. Group graph convolutional networks for 3D human pose estimation. In *Proc. British Machine Vision Conference*, 2022.

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

[17] Weixi Zhao, Weiqiang Wang, and Yunjie Tian. GraFormer: Graph-oriented transformer for 3D pose estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 20438–20447, 2022.

[18] Zhiming Zou and Wei Tang. Modulated graph convolutional network for 3D human pose estimation. In *Proc. IEEE International Conference on Computer Vision*, pages 11477–11487, 2021.

[19] Zhiming Zou, Kenkun Liu, Le Wang, and Wei Tang. High-order graph convolutional networks for 3D human pose estimation. In *Proc. British Machine Vision Conference*, 2020.