A Comprehensive Study of Weight Sharing in Graph Networks for 3D Human Pose Estimation: Supplementary Material

Kenkun Liu[1⋆], Rongqi Ding[2⋆], Zhiming Zou[1], Le Wang[3], and Wei Tang[1⋆⋆]

[1] University of Illinois at Chicago, Chicago, IL, USA
{kliu44,zzou6,tangw}@uic.edu
[2] Northwestern University, Evanston, IL, USA
rongqiding2020@u.northwestern.edu
[3] Xi'an Jiaotong University, Xi'an, Shaanxi, P.R. China
lewang@xjtu.edu.cn

1 Training Curves and Validation Errors

Fig. 1 plots the loss and validation error of each method in the training phase (control model size to be 4.2 M). Compared with `full-sharing`, the training losses of other methods decrease faster. As the training goes on, the validation errors of `pre-aggregation`, `post-aggregation` and `no-sharing` are significantly lower than those of `full-sharing` and `convolution-style`. A larger model size of `full-sharing` makes its validation errors oscillate severely.

2 Reduce Complexity

We can implement the feature update of each node of all weight sharing methods in a unified fashion by first computing the feature transformations of neighboring nodes and then aggregating them. This makes their computational complexities the same given a fixed number of channels. In practice, we can reduce the complexities of `full-sharing`, `pre-aggregation` and `post-aggregation` from $D' \times D \times \sum_{i=1}^{N} |\hat{\mathcal{N}}_i|$ to $D' \times D \times N$ by precomputing and saving the feature transformation of each node or first aggregating the features from neighboring nodes before feature transformation. Here D', D, $|\hat{\mathcal{N}}_i|$ and N denote the number of output channels, the number of input channels, the number of neighboring nodes of node i and the total number of nodes, respectively.

3 Ablation Study on 2D Detections

We compare the five weight sharing methods, Martinez et al. [4] and Zhao et al. [10] given different 2D pose detections. Tab. 1 shows `pre-aggregation` achieves the overall best performance.

⋆ The first two authors contributed equally to this work.
⋆⋆ Corresponding author.

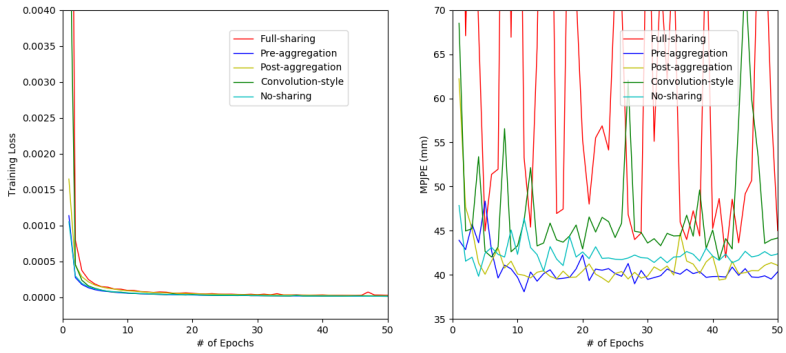

Fig. 1. Training curves (left) and validation errors (right) of different methods (control the model size to be 4.2 M)

Table 1. Quantitative comparisons on the Human 3.6M dataset using keypoints from 2D ground truth (GT), hourglass network (HG) [6] or CPN [1]. The MPJPE (P1) and P-MPJPE (P2) are measured in millimeters (mm)

Keypoints/Protocol	GT/P1	GT/P2	HG/P1	HG/P2	CPN/P1	CPN/P2
Martinez et al. [4]	44.40	35.25	63.48	48.15	-	-
Zhao et al. [10]	40.78	31.46	61.24	47.71	-	-
`Pre-agg`	**37.83**	**30.09**	**59.53**	**46.35**	**52.42**	**41.48**
`Post-agg`	38.92	31.33	63.40	48.92	56.15	44.01
`Conv-style`	41.19	32.20	60.52	46.66	53.24	41.62
`No-sharing`	39.62	30.93	59.62	**46.35**	53.44	42.39
`Full-sharing`	41.70	33.02	59.85	46.55	53.71	42.20

4 Results on MPI-INF-3DHP

The MPI-INF-3DHP dataset [5] includes not only indoor scenes but also outdoor scenes with 3D pose annotations. Following previous works [2, 11], we use its test set to evaluate the generalization capacity of our model (`pre-aggregation`) trained on the training set of Human3.6M. The performance is measured by the metrics 3D PCK and AUC. The results in Tab. 2 show our method outperforms other states of the art in both "**GS**" and "**noGS**", and is competitive to [11] in "**Outdoor**". Overall, our method achieves the best performance.

5 Failure Cases

Fig. 2 shows some failure cases of our best weight sharing method `pre-aggregation`. We find most of the failures occur on wrists and elbows. This is mainly caused

Table 2. Quantitative comparisons on the MPI-INF-3DHP dataset between our method (`pre-aggregation`) and state-of-the-art methods

	Training Data	GS	noGS	Outdoor	ALL(PCK)	ALL(AUC)
Martinez et al. [4]	H36m	49.8	42.5	31.2	42.5	17.0
Yang et al. [9]	H36m+MPII	-	-	-	69.0	32.0
Zhou et al. [12]	H36m+MPII	71.1	64.7	72.7	69.2	32.5
Pavlakos et al. [7]	H36m+MPII+LSP	76.5	63.1	77.5	71.9	35.3
Ci et al. [2]	H36m	74.8	70.8	77.3	74.0	36.7
Wang et al. [8]	H36m	-	-	-	71.9	35.8
Li et al. [3]	H36m+MPII	70.1	68.2	66.6	67.9	-
Zhou et al. [11]	H36m+MPII	75.6	71.3	**80.3**	75.3	38.0
Ours	H36m	**77.6**	**80.5**	80.1	**79.3**	**47.6**

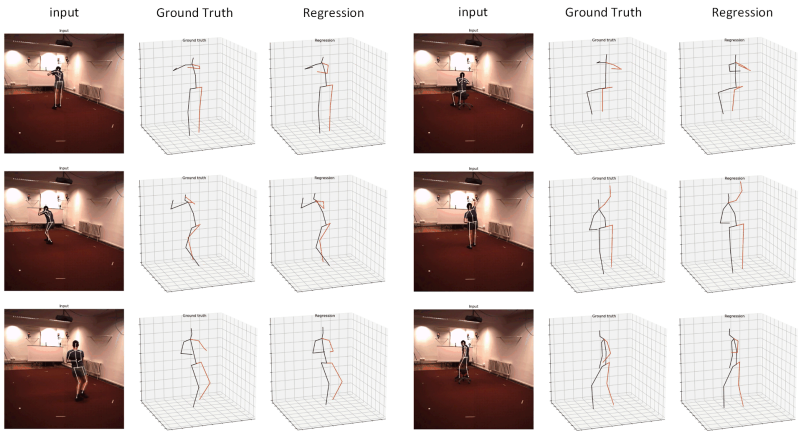

Fig. 2. Failure cases of `pre-aggregation`

by 2D detection errors when occlusion occurs. Though some predicted poses are not perfectly aligned with their respective ground truth, they still look plausible.

References

1. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7103–7112 (2018)
2. Ci, H., Wang, C., Ma, X., Wang, Y.: Optimizing network structure for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2262–2271 (2019)
3. Li, C., Lee, G.H.: Generating multiple hypotheses for 3d human pose estimation with mixture density network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9887–9895 (2019)
4. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2640–2649 (2017)
5. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 international conference on 3D vision (3DV). pp. 506–516. IEEE (2017)
6. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016)
7. Pavlakos, G., Zhou, X., Daniilidis, K.: Ordinal depth supervision for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7307–7316 (2018)
8. Wang, J., Huang, S., Wang, X., Tao, D.: Not all parts are created equal: 3d pose estimation by modeling bi-directional dependencies of body parts. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7771–7780 (2019)
9. Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., Wang, X.: 3d human pose estimation in the wild by adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5255–5264 (2018)
10. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3d human pose regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3425–3435 (2019)
11. Zhou, K., Han, X., Jiang, N., Jia, K., Lu, J.: Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2344–2353 (2019)
12. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 398–407 (2017)