—Supplementary Materials— SRNet: Improving Generalization in 3D Human Pose Estimation with a Split-and-Recombine Approach

Ailing Zeng[1,2⋆], Xiao Sun[2], Fuyang Huang[1], Minhao Liu[1], Qiang Xu[1], and Stephen Lin[2]

[1] The Chinese University of Hong Kong
[2] Microsoft Research Asia
alzeng, fyhuang, mhliu, qxu@cse.cuhk.edu.hk
xias, stevelin@microsoft.com

In this supplementary material, we present more implementation details of our model, additional experimental results and more qualitative results which are not shown in the main paper due to the space limitation. First, Section 1 gives more details on experiment settings. Second, Section 2 discuses different design choices for the combination operator $\circ$ in Equation 5 of the main paper. Third, Section 3 shows results of using both MPJPE and PA-MPJPE as the evaluation metrics in a comparison with state-of-the-art methods on the Human3.6M dataset. Next, Section 4 shows more results under the cross action protocol. Finally, Section 5 demonstrates additional qualitative results.

1 Implementation Details

Training Data. Following many previous works [14, 11, 6, 18, 20, 5], we show results of using two different kinds of 2D keypoints as input for our model in the experiments. They are 2D ground truth and 2D detections from an off-the-shelf 2D keypoint detector. Following those works [14, 3, 9], we use the smoothed CPN model [4] which finetuned on the Human3.6M dataset by an eight-layer residual fully-connected temporal model as our 2D keypoint detector, which is pretrained on the COCO dataset. No extra 2D data has been used for mixed training. In the ablation study (Section 5.1) of the main paper, we use 2D ground truth as input. When comparing with previous works in Section 5.2, both inputs are used and compared respectively. For data normalization, we use two methods. One is provided by [14] called the ***Basic*** normalization and another is provided by [5] called the ***Pixel*** normalization. Please refer to their code base [1, 2] for detailed implementation. By default, the ***Basic*** normalization is used in our ablation study. When comparing with the state-of-the-arts ([5] in *single pose* and [14] in *temporal pose*), we use the same data normalization method as each for a fair comparison. For data augmentation, we follow [14, 5] by using horizontal flip data augmentation at both training and test stages.

⋆ The work is done when Ailing Zeng is an intern at Microsoft Research Asia.

Training Setting. Amsgrad [17] is used as the optimizer. The initial learning rate is 0.001 and it decays by 5% after each epoch of training. 80 epochs are used in total. The total channel dimension of each connected/convolution layer is 1024. Batch Normalization [7] and Leaky ReLU [19] activation are applied to each connected/convolution layer. The final network consists of 8 stacked layers, and every two layers (except for the first and last ones) are wrapped with a residual connection as in [11, 14]. Batch size is 1024. L1 loss is used for training.

2 Design Choices for the Combination Operator

In Equation 5 of the main paper, we show how the low-dimensional global contexts can be brought back to the local group using a combination operator $\circ$. By default, the combination operator $\circ$ is implemented using multiplication in the main paper. Here, we empirically evaluate the design choices of using addition, multiplication and concatenation in Table 1. It is shown that both addition and multiplication obtain favorable results. They surpass the FC and SFS baselines, indicating their effectiveness in recombining the low-dimensional global contexts.

Method	FC	SFS	SR (add.)	SR (mult.)	SR (concat.)
MPJPE(mm)	46.8	39.4	36.4	36.6	38.3
Params.(M)	6.39	3.04	1.33	0.88	1.34

Table 1. Comparison on different design choices for the combination operator under the *Subject* protocol. MPJPE is used as the evaluation metric. 2D ground truth is used as input. The third row shows the number of learnable parameters of different models.

3 More Results on Human3.6M

In Tables 8, 9, and 10 of the main paper, we compare our model with previous works under different input settings (using 2D ground truth or detection, with or without temporal information). We summarise them in Table 2 and 3 with more detailed results on different actions. Our approach achieves the new state-of-the-art with either 2D keypoint detection or 2D ground truth as input. Specifically, we improve upon [5] from 36.3mm to 33.9mm (relative 6.6% improvement) with 2D ground truth input for single pose inputs. We improve upon [9] from 46.6mm to 44.8mm (relative 3.9% improvement) with 2D temporal keypoint detection input.

In Table 4 and 5, we compare with the previous works using the *PA-MPJPE* metric where available. Our approach achieves the new state-of-the-art with either 2D keypoint detection or 2D ground truth (denoted by $\bigtriangledown$) as input. Specifically, we improve upon [5] from 27.9mm to 24.3mm (relative 14.8% improvement) with 2D ground truth input. We improve upon [14] from 36.5mm to 34.9mm (relative 4.4% improvement) with 2D keypoint detection input.

Method	Direct	Discuss	Eat	Greet	Phone	Photo	Pose	Purcha.	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg.
Luvizon et al. [10]	63.8	64.0	56.9	64.8	62.1	70.4	59.8	60.1	71.6	91.7	60.9	65.1	51.3	63.2	55.4	64.1
Martinez et al. [11]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Park et al.[12]	49.4	54.3	51.6	55.0	61.0	73.3	53.7	50.0	68.5	88.7	58.6	56.8	57.8	46.2	48.6	58.6
Wang et al. [18]	47.4	56.4	49.4	55.7	58.0	67.3	**46.0**	46.0	67.7	102.4	57.0	57.3	41.1	61.4	40.7	58.0
Zhao et al. [20]	47.3	60.7	51.4	60.5	61.1	**49.9**	47.3	68.1	86.2	**55.0**	67.8	61.0	42.1	60.6	45.3	57.6
Ci et al. [5]	46.8	52.3	44.7	50.4	52.9	68.9	49.6	46.4	60.2	78.9	**51.2**	50.0	54.8	40.4	43.3	52.7
Pavllo et al. [14]	47.1	50.6	49.0	51.8	53.6	61.4	49.4	47.4	59.3	67.4	52.4	49.5	55.3	39.5	42.7	51.8
Cai et al. [3]	46.5	48.8	47.6	50.9	52.9	61.3	48.3	45.8	**59.2**	64.4	**51.2**	48.4	**53.5**	39.2	41.2	50.6
Ours	**44.5**	**48.2**	**47.1**	**47.8**	**51.2**	56.8	50.1	**45.6**	59.9	66.4	52.1	**45.3**	54.2	**39.1**	**40.3**	**49.9**
Martinez et al. [11]▽	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
Pham et al. [15]▽	36.6	43.2	38.1	40.8	44.4	51.8	43.7	38.4	50.8	52.0	42.1	42.2	44.0	32.3	35.9	42.4
Zhao et al. [20]▽	37.8	49.4	37.6	40.9	45.1	**41.4**	40.1	48.3	50.1	42.2	53.5	44.3	40.5	47.3	39.0	43.8
Wang et al. [18]▽	35.6	41.3	39.4	40.0	44.2	51.7	39.8	40.2	50.9	55.4	43.1	42.9	45.1	33.1	37.8	42.0
Ours-Basic ▽	35.9	36.7	29.3	34.5	36.0	42.8	37.7	31.7	40.1	44.3	35.8	37.2	36.2	33.7	34.0	36.4
Ci et al.-Pixel [5] ▽	36.3	38.8	29.7	37.8	34.6	42.5	39.8	32.5	**36.2**	**39.5**	34.4	38.4	38.2	31.3	34.2	36.3
Ours-Pixel ▽	**32.9**	**34.5**	**27.6**	**31.7**	**33.5**	42.5	**35.1**	**29.5**	38.9	45.9	**33.3**	**34.9**	**34.4**	**26.5**	**27.1**	**33.9**

Table 2. Detailed *single pose* comparison in terms of the mean per-joint position error (MPJPE) on Human3.6M. Below the double line are results from 2d ground truth inputs (indicated by ▽) to explore the upper bound of these methods. Best results in bold.

Method	Direct	Discuss	Eat	Greet	Phone	Photo	Pose	Purcha.	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg.
Hossain et al. [16]	48.4	50.77	57.2	55.2	63.1	72.6	53.0	51.7	66.1	80.9	59.0	57.3	62.4	46.6	49.6	58.3
Lee et al. [8]	**40.2**	49.2	47.8	52.6	50.1	75.0	50.2	43.0	55.8	73.9	54.1	55.6	58.2	43.3	43.3	52.8
Cai et al. [3]	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9	57.9	61.9	49.7	46.6	51.3	37.1	39.4	48.8
Pavllo et al. [14]	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
Lin et al. [9]	42.5	**44.8**	**42.6**	44.2	48.5	57.1	**42.6**	41.4	56.5	64.5	47.4	43.0	48.1	33.0	35.1	46.6
Ours	43.1	47.1	43.9	**41.6**	**45.8**	**49.6**	46.5	**40.0**	**53.4**	**61.1**	**46.1**	**42.6**	**46.6**	**31.5**	**32.6**	**44.8**
Hossain et al. [16] ▽	35.2	40.8	37.2	37.4	43.2	44.0	38.9	35.6	42.3	44.6	39.7	39.7	40.2	32.8	35.5	39.2
Lee et al. [8] ▽	32.1	36.6	34.3	37.8	44.5	49.9	40.9	36.2	44.1	45.6	35.3	35.9	37.6	30.3	35.5	38.4
Pavllo et al.-243f [14] ▽	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37.2
Ours-243f ▽	**34.8**	**32.1**	**28.5**	**30.7**	**31.4**	**36.9**	**35.6**	**30.5**	**38.9**	**40.5**	**32.5**	**31.0**	**29.9**	**22.5**	**24.5**	**32.0**

Table 3. Detailed *temporal pose* comparison in terms of the mean per-joint position error (MPJPE) on Human3.6M. Below the double line are results from 2d ground truth inputs (indicated by ▽) to explore the upper bound of these methods. Best results in bold.

Method	Direct	Discuss	Eat	Greet	Phone	Photo	Pose	Purcha.	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg.
Martinez et al. [11]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Fang et al.[6]	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
Park et al. [12]	38.3	42.5	41.5	43.3	47.5	53.0	39.3	37.1	54.1	64.3	46.0	42.0	44.8	34.7	38.7	45.0
Ci et al. [5]	36.9	41.6	38.0	41.0	41.9	51.1	38.2	37.6	49.1	62.1	43.1	39.9	43.5	32.2	37.0	42.2
Pavlakos et al. [13]	**34.7**	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
Pavllo et al.[14]	36.0	38.7	38.0	41.7	40.1	45.9	**37.1**	**35.4**	**46.8**	**53.4**	41.4	36.9	43.1	30.3	**34.8**	40.0
Ours	35.8	**39.2**	**36.6**	**36.9**	**39.8**	**45.1**	38.4	36.9	47.7	54.4	**38.6**	**36.3**	**39.4**	**30.3**	35.4	**39.4**
Ours-Basic $\bigtriangledown$	26.0	28.9	23.7	26.9	27.4	33.1	27.9	25.0	32.4	40.9	28.8	29.2	29.3	23.3	24.5	28.5
Ours-Pixel $\bigtriangledown$	24.1	28.6	24.2	26.6	26.3	35.1	27.7	24.5	32.8	39.1	27.8	28.0	29.6	22.3	23.0	28.0

Table 4. Comparison *single pose* results regarding PA-MPJPE after rigid transformation from the ground truth.$\bigtriangledown$ indicates the use of 2D ground truth poses as input. Best results in bold.

Method	Direct	Discuss	Eat	Greet	Phone	Photo	Pose	Purcha.	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg.
Lee et al.[8]	38.0	39.3	46.3	44.4	49.0	55.1	40.2	41.1	53.2	68.9	51.0	39.1	**33.9**	56.4	38.5	46.2
Hossain et al.[16]	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	39.4	44.1
Cai et al.[3]	35.7	37.8	36.9	40.7	39.6	45.2	37.4	34.5	46.9	**50.1**	40.5	36.1	41.0	29.6	33.2	39.0
Lin et al. [9]	32.5	35.3	**34.3**	36.2	37.8	43.0	**33.0**	32.2	45.7	51.8	38.4	32.8	37.5	25.8	28.9	36.8
Pavllo et al.-243f [14]	34.1	36.1	34.4	37.2	36.4	**42.2**	34.4	33.6	45.0	52.5	37.4	33.8	37.8	**25.6**	27.3	36.5
Ours-243f	**31.9**	**33.7**	34.7	**35.0**	**35.5**	42.8	36.4	**30.5**	**43.6**	51.3	**36.7**	**32.5**	**36.5**	27.5	**25.7**	**34.9**
Ours-243f $\bigtriangledown$	**23.7**	**25.2**	**22.9**	**23.1**	**24.0**	**28.7**	**25.0**	**22.1**	**31.8**	**32.8**	**24.8**	**23.5**	**23.4**	**17.0**	**18.3**	**24.3**

Table 5. Comparison *temporal pose* results regarding PA-MPJPE after rigid transformation from the ground truth. $243f$ means inputs contain 243 frame poses. $\bigtriangledown$ indicates the use of 2D ground truth poses as input. Best results in bold.

4 Cross Action Results Using 2D Ground Truth Input

In Table 7 of the main paper, we compare our cross-action results with [5] under the same data settings. Here, we provide more results of using 2d ground truth as input under the cross-action protocol. The FCN baseline [11] (with our implementation) and our SRNet are compared. Both MPJPE and PA-MPJPE (with $\times$) are used as the evaluation metrics. Both ***Basic*** and ***Pixel*** [5] normalization results of our method are reported.

In Table 6, our method gains improvements in terms of MPJPE from 80.6mm to 64.3mm, by 16.3mm (relatively 20.2%). For PA-MPJPE, the improvement is from 60.5mm to 49.4mm, by 11.1mm (relatively 18.3%).

5 Additional Qualitative Results

Besides the aforementioned quantitative results, we also present some qualitative results. First, we visualize some hard poses, which are also rare in the subject protocol evaluation, in Figure 1. Under this protocol, our method can predict well even on challenging poses such as kowtow, side-lying and legs lifting. Next, Figure 2 demonstrates some unseen poses in the cross-action protocol to verify our method's generalization ability. Finally, Figure 3 shows some qualitative

Method	Direct	Discuss	Eat	Greet	Phone	Photo	Pose	Purcha.	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg.
FCN-Pixel [11]	117.0	67.4	62.6	93.0	59.5	72.8	66.7	80.0	71.2	71.6	58.6	75.2	73.3	114.9	125.0	80.6
Ours-Basic	91.1	54.8	59.0	71.2	50.9	61.5	65.0	71.4	76.6	74.0	50.3	64.8	58.1	78.0	85.8	67.5
Ours-Pixel	**86.2**	**53.0**	**55.0**	**70.5**	**47.9**	**57.9**	**63.1**	**68.4**	**71.2**	**72.9**	**47.5**	**59.4**	**56.3**	**70.8**	**83.8**	**64.3**
FCN-Pixel[11] ×	91.9	55.3	51.8	75.2	49.3	60.6	57.3	64.7	62.2	60.6	49.5	62.7	61.3	95.4	99.8	60.5
Ours-Basic ×	65.9	42.4	46.3	54.5	39.8	46.6	50.6	55.8	58.4	57.4	39.3	49.6	45.0	56.7	61.8	51.3
Ours-Pixel×	**61.7**	**42.0**	**44.2**	**53.1**	**38.5**	**45.2**	**49.5**	**53.6**	**55.5**	**55.5**	**37.9**	**46.4**	**43.8**	**54.7**	**59.7**	**49.4**

Table 6. *Cross Action* comparison to the FCN baseline with *2D ground truth input* on Human3.6M in terms of mean per-joint position error (MPJPE) and PA-MPJPE (denoted by ×).

results with training only on the Human3.6M dataset and testing on unseen poses and unseen camera angles. Nevertheless, our method is still able to reconstruct many plausible 3D poses well.

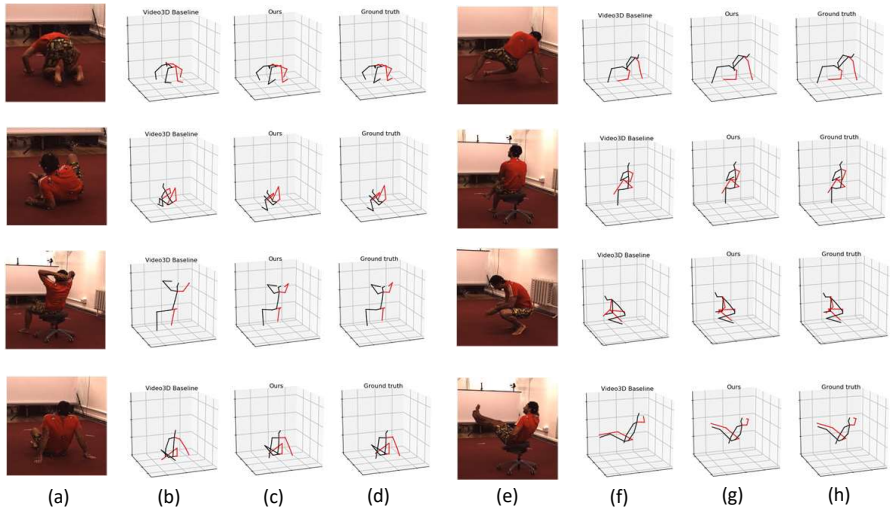

Fig. 1. Visualization results trained with the subject protocol settings on the Human3.6M dataset. (a), (e) are the original test images. (b), (f) show the 3D pose predictions of temporal 3D pose baseline [14]. (c), (g) are the 3D pose predictions of our method. (d), (h) are the 3D ground truth poses.

Fig. 2. Visualization results for the cross-action protocol. (a), (b) are two kinds of original training actions. (c), (d) show the 3D predicted results by FCN [11], our method, and the 3D ground truth poses on two kinds of test actions. When training action is "greet", poses like in (a), test on the action "sit" to get those predictions in (c). Similarly, when training action is "take photos" in (b), test on the action "sit down" to show the differences between the FCN and our method in (d).

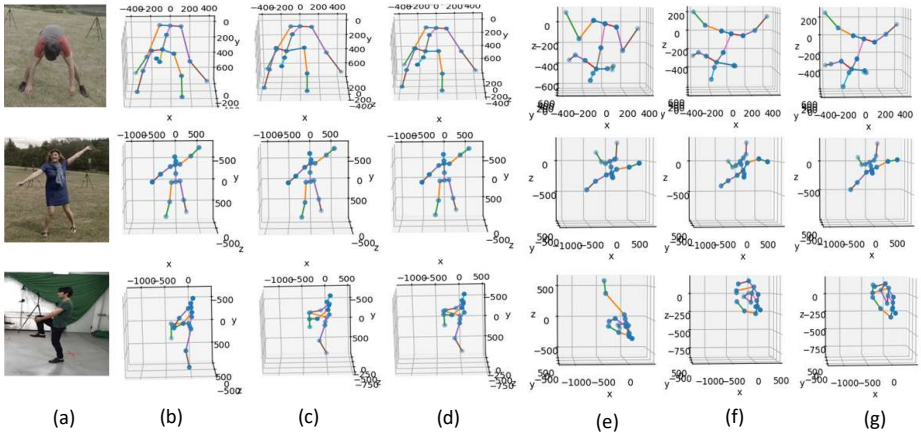

Fig. 3. Visualization results for the MPI-INF-3DHP dataset. (a) are the original images. (b), (e) show the 3D predicted results by [11] from the front viewpoint and the top viewpoint. (c), (f) show the prediction poses of our method, and (d), (g) are the 3D ground truth poses from the front viewpoint and the top viewpoint, separately.

References

1. Code for 3d human pose estimation in video with temporal convolutions and semi-supervised training. https://github.com/facebookresearch/VideoPose3D
2. Code for optimizing network structure for 3d human pose estimation. https://chunyuwang.netlify.app/
3. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.J., Yuan, J., Thalmann, N.M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2272–2281 (2019)
4. 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)
5. 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)
6. Fang, H.S., Xu, Y., Wang, W., Liu, X., Zhu, S.C.: Learning pose grammar to encode human body configuration for 3d pose estimation. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
7. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
8. Lee, K., Lee, I., Lee, S.: Propagating lstm: 3d pose estimation based on joint interdependency. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 119–135 (2018)
9. Lin, J., Lee, G.H.: Trajectory space factorization for deep video-based 3d human pose estimation. arXiv preprint arXiv:1908.08289 (2019)
10. Luvizon, D.C., Picard, D., Tabia, H.: 2d/3d pose estimation and action recognition using multitask deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5137–5146 (2018)
11. 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)
12. Park, S., Kwak, N.: 3d human pose estimation with relational networks. arXiv preprint arXiv:1805.08961 (2018)
13. 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)
14. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7753–7762 (2019)
15. Pham, H.H., Salmane, H., Khoudour, L., Crouzil, A., Zegers, P., Velastin, S.A.: A unified deep framework for joint 3d pose estimation and action recognition from a single rgb camera. arXiv preprint arXiv:1907.06968 (2019)
16. Rayat Imtiaz Hossain, M., Little, J.J.: Exploiting temporal information for 3d human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 68–84 (2018)
17. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of adam and beyond. arXiv preprint arXiv:1904.09237 (2019)

18. Wang, L., Chen, Y., Guo, Z., Qian, K., Lin, M., Li, H., Ren, J.S.: Generalizing monocular 3d human pose estimation in the wild. arXiv preprint arXiv:1904.05512 (2019)
19. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 (2015)
20. 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)