HEMlets Pose: Learning Part-Centric Heatmap Triplets for Accurate 3D Human Pose Estimation Supplementary Material

Kun Zhou[1], Xiaoguang Han[2], Nianjuan Jiang[1], Kui Jia[3], and Jiangbo Lu[1*]
[1]Shenzhen Cloudream Technology Co., Ltd. [2]The Chinese University of Hong Kong (Shenzhen)
[3]South China University of Technology [*]Corresponding email: jiangbo.lu@gmail.com

This supplementary material presents more results, both visually and numerically, which were not included in the main manuscript.

1. Qualitative Ablation Studies

In this section, we show the qualitative results for the ablation studies in the following three aspects.

1.1. Alternative intermediate supervision

There are several variants to train our HEMlets model: a) only the 3D joint loss $\mathcal{L}_{\lambda}^{\mathrm{3D}}$ is used (denoted as "Baseline"), b) both $\mathcal{L}_{\lambda}^{\mathrm{3D}}$ and 2D joint loss $\mathcal{L}^{2D}$ are used ("w/ 2D joint loss"), c) both $\mathcal{L}_{\lambda}^{\mathrm{3D}}$ and our HEMlets loss $\mathcal{L}^{\mathrm{HEM}}$ are used ("w/ HEM loss"), d) all of the three losses are used ("Full loss"). The qualitative results of 12 examples (sampled from Human3.6M [1]) are shown in Fig. 1. As can be seen from Fig. 1, the model trained with "Full loss" achieves the best visual performance.

1.2. Variants of HEMlets

As presented in the main manuscript, there are two primary variants of the proposed HEMlets representation, i.e., "5s-HEM" and "2s-HEM". The qualitative results of 8 examples, sampled from Human3.6M [1], are shown in Fig. 2. As shown, the proposed HEMlets generates the pose estimation results of better visual quality than the others.

1.3. Fine-tuning with additional datasets

As presented in the main manuscript, there are two recent datasets [4, 5] that provide relative depth ordering annotations. Numerical comparisons for augmenting datasets [4, 5] have been reported in our main manuscript (Sec. 4.3). In this supplementary material, we give some visual results on MPI-INF-3DHP [3] for the finetuned models when using different additional datasets, which are shown in Fig. 3. We find the model finetuned with FBI [5] produces better predictions than the ones trained additionally with Ordinal [4].

2. More Qualitative Results

This section provides more qualitative results on Human3.6M [1] in Fig. 4, HumanEva-I [6] in Fig. 5 and Leeds Sports Pose (LSP) [2] in Fig. 6.

3. Video Results for Additional Evaluation

We also provide two videos for visual inspection.

"Human36Mcomparison.mp4" This video shows both visual and numerical comparisons of our HEMlets method with the state-of-the-art [7], on one of the most challenging actions in Human3.6M [1]. Specifically, the video clip of "SittingDown" from the test sequence of Human3.6M is used (SittingDown.60457274 in S11). For fair comparison, the model of [7] is obtained after re-training with the same hardware settings as ours (the parameters are set to follow the details reported in [7]). The bottom draws the per-frame prediction errors of these two methods.

"MPIINF3DHPeval.mp4" This video demonstrates the visual results generated by our method on one example sequence of MPI-INF-3DHP [3]. The results seen from two different perspectives are shown.

References

[1] 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.

[2] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference (BMVC)*, 2010.

[3] 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 *3D Vision (3DV)*, pages 506–516. IEEE, 2017.

[4] 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 (CVPR)*, pages 7307–7316, 2018.

[5] Yulong Shi, Xiaoguang Han, Nianjuan Jiang, Kun Zhou, Kui Jia, and Jiangbo Lu. FBI-pose: Towards bridging the gap between 2d images and 3d human poses using forward-or-backward information. *arXiv preprint arXiv:1806.09241*, 2018.

[6] Leonid Sigal, Alexandru O Balan, and Michael J Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1-2):4, 2010.

[7] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545. Springer, 2018.

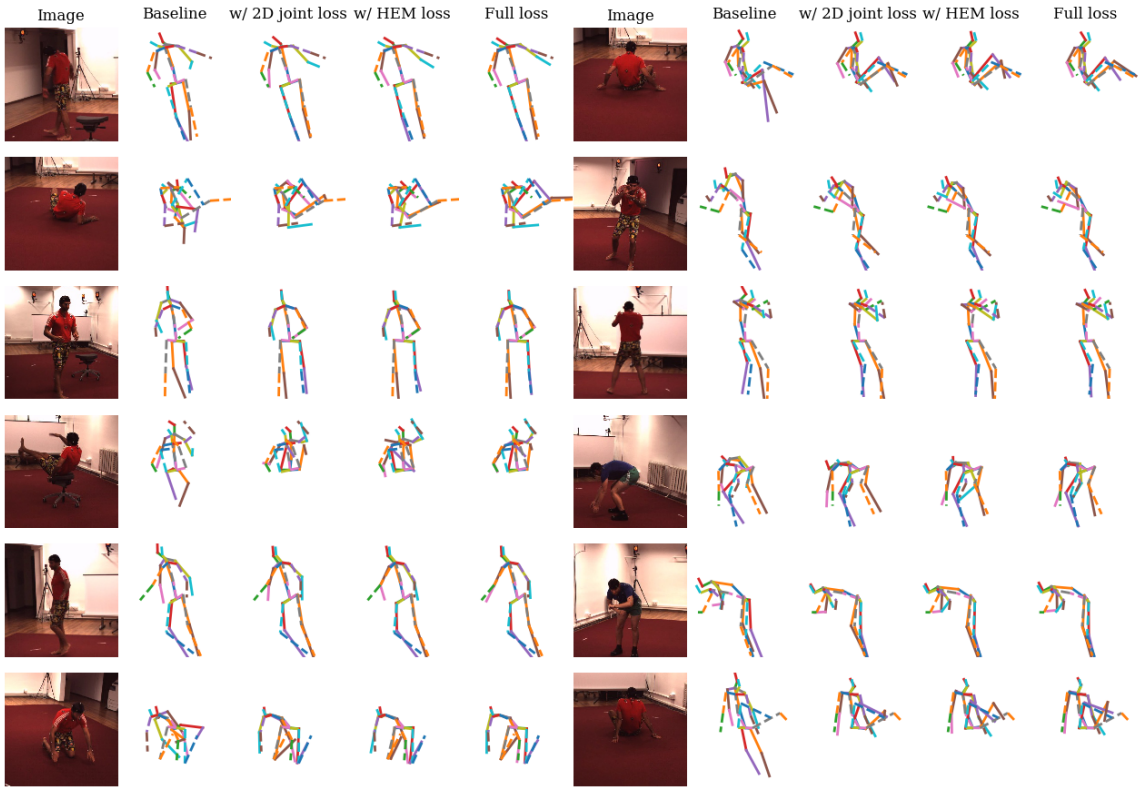

Figure 1. The qualitative results on several sampled examples from Human3.6M [1], based on alternative intermediate supervisions. The groundtruth pose is shown in dashed line.

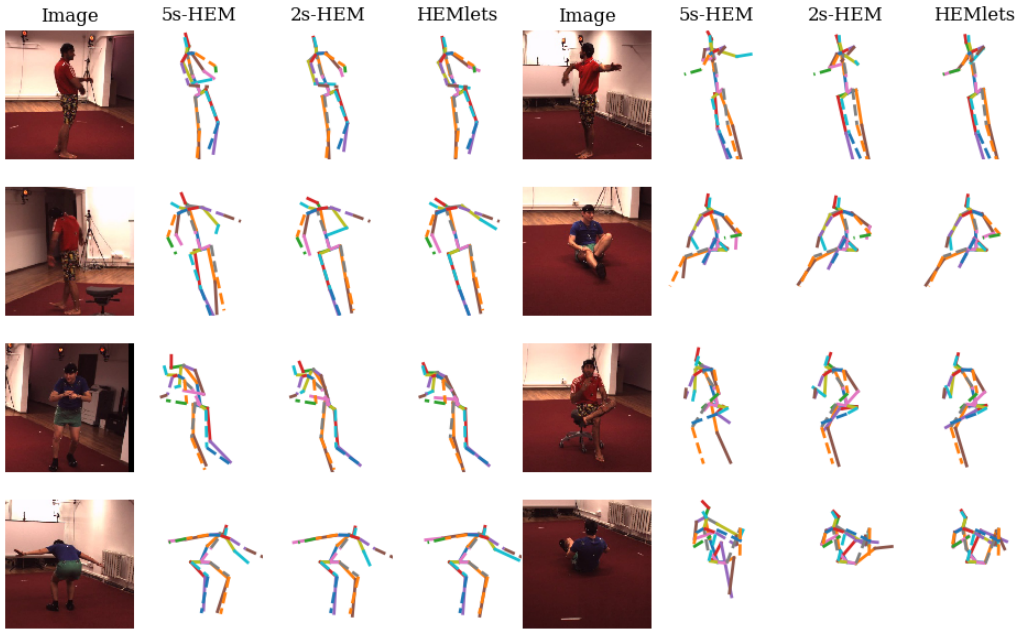

Figure 2. We sample 8 examples from Human3.6M [1]. For each example, the results of "5s-HEM" , "2s-HEM" and "HEMlets" are shown. The groundtruth pose is shown in dashed line.

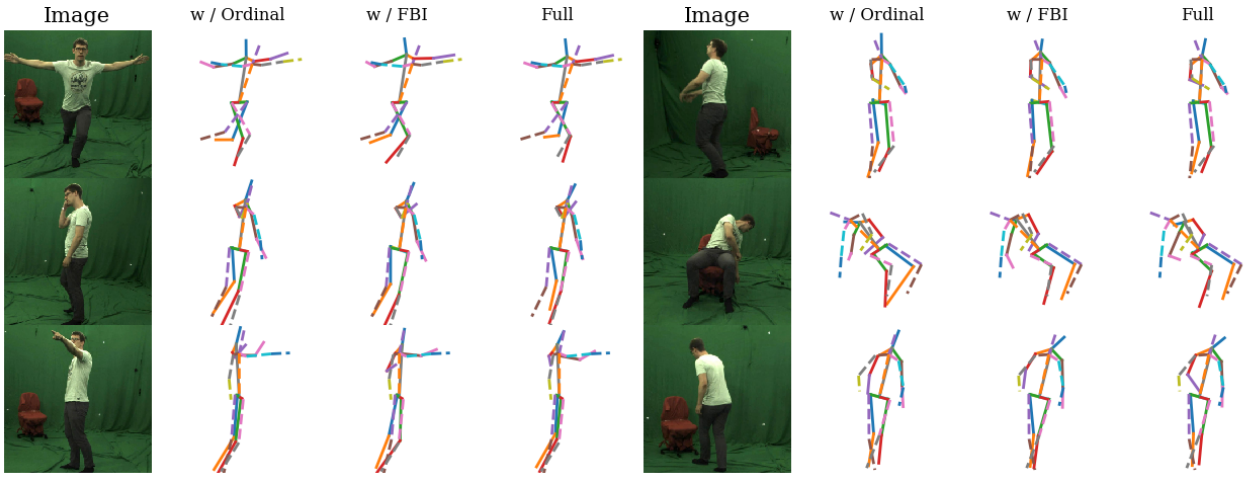

Figure 3. The qualitative results for 6 sampled examples of MPI-INF-3DHP [3], using different additional datasets. For each example, we present the input RGB image, the 3D human pose predicted by three different models. The groundtruth pose is shown in dashed line.

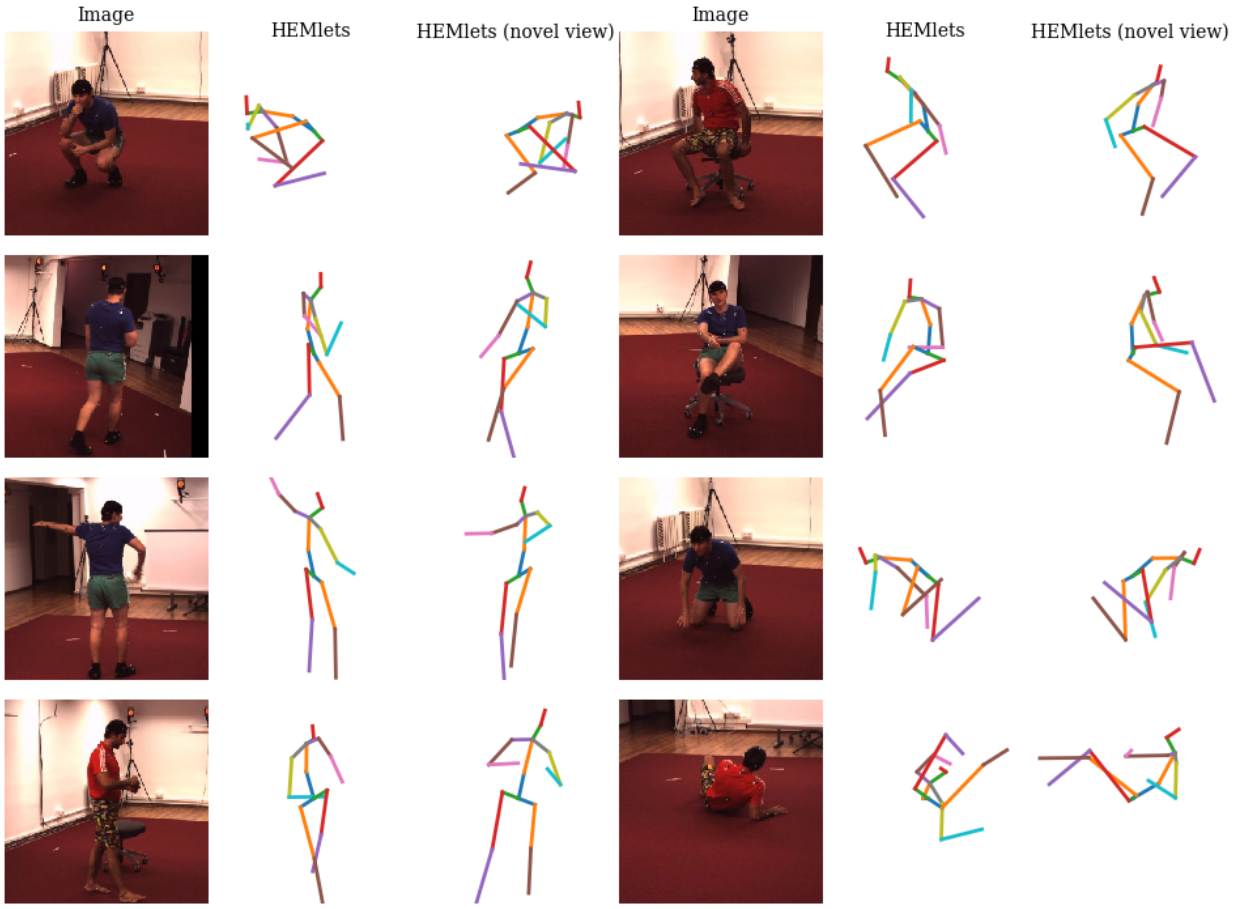

Figure 4. The qualitative results of the presented approach on Human3.6M [1]. The examples are randomly picked from the test set of Human3.6M [1]. For each example, two different views of the predicted 3D human pose are shown.

Figure 5. The qualitative results of the presented approach on HumanEva-I [6]. The three column blocks correspond to the three validation sequences “S1”, “S2” and “S3”, respectively. The examples are randomly picked from those three sequences. For each example, two different views of the predicted 3D human pose are shown.

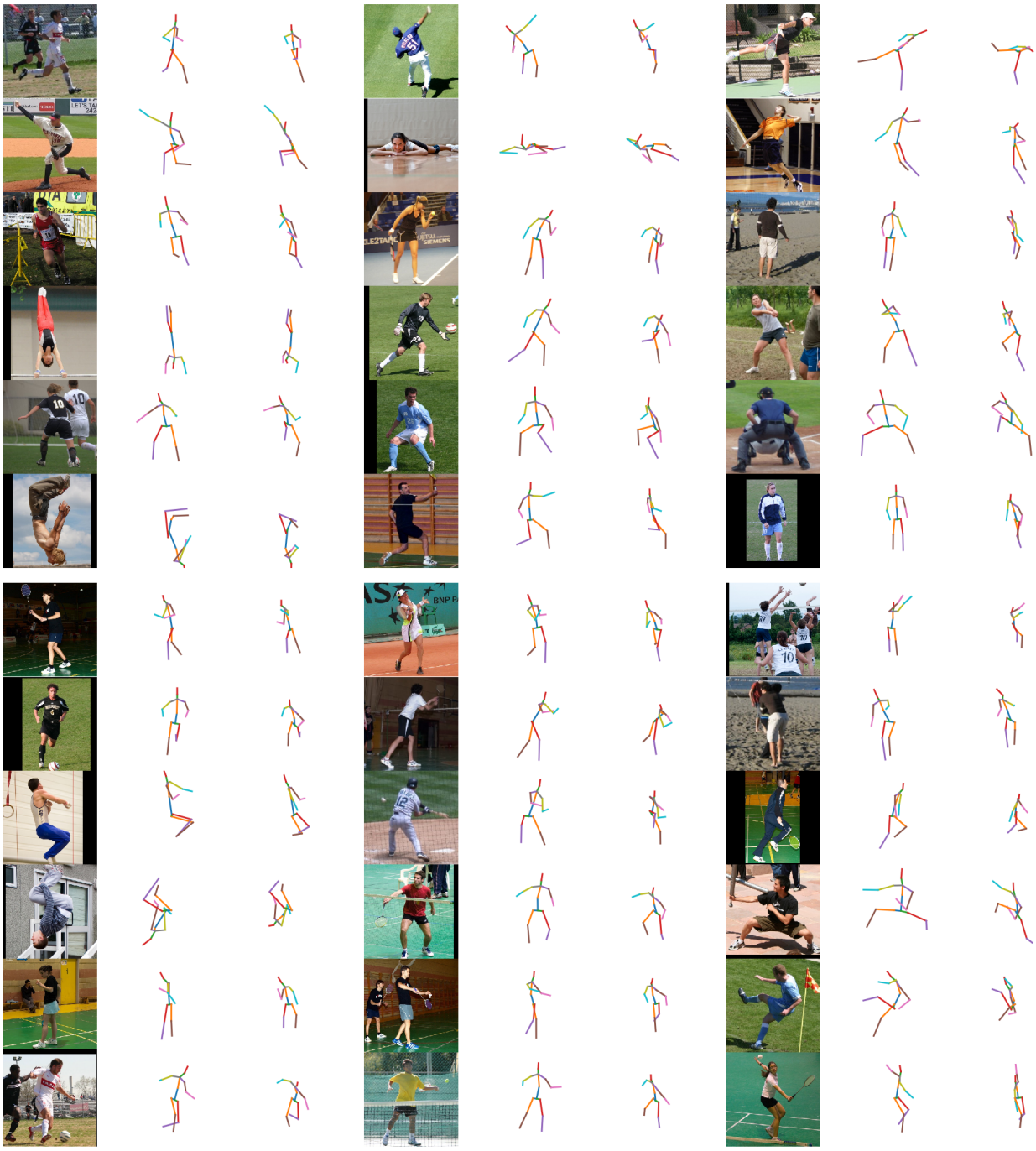

Figure 6. The visual results predicted by our method on Leeds Sports Pose (LSP) [2]. They demonstrate the great generalization ability of the proposed approach.