Supplemental Material: Hierarchical Kinematic Human Mesh Recovery

Georgios Georgakis*[1,2], Ren Li*[1], Srikrishna Karanam[1], Terrence Chen[1], Jana Košecká[2], and Ziyan Wu[1]

[1] United Imaging Intelligence, Cambridge MA, USA
[2] George Mason University, Fairfax VA, USA
{first.last}@united-imaging.com

1 Architecture and Training Details

- Encoder: We use the standard ResNet50, giving a 2048-dimensional feature vector.
- Chains Q_c: Each Q_c is implemented with a set of fully connected layers. The ψ embedding module E_c comprises two fully connected units (with ReLU activations and dropout in training) with 32-dimensional outputs each. The input dimensionality of E_c varies according to the chain. This is 2070 for the root chain, 2085 for the arm chains, 2082 for the leg chains, and 2076 for the head chain. Finally, each $\Delta\theta_i$ is realized with one single fully connected layer with a 3-dimensional output.
- Number of inner iterations: 4 (so 2 forward-backward cycles).
- Number of outer iterations $T = 3$.
- Shape estimating neural network: Three fully connected units (with ReLU activations and dropout during training) with output units of 512, 128, and 10 respectively.
- Camera estimating neural network: Three fully connected units (with ReLU activations and dropout during training) with output units of 512, 128, and 3 respectively.
- VAE encoder: Input dimensionality of 69 (23 joint pose parameters); two fully connected units (with Leaky ReLU activation, batch normalization, and dropout in training) with output units of 512 and 32 each, so the latent space dimensionality is 32.
- VAE decoder: Input dimensionality of 32 (latent space vector); two fully connected units (with Leaky ReLU activation, batch normalization, and dropout in training) with output units of 512 and 69 each, so the output dimensionality is 69.
- HKMR training parameters. We set the loss weights $\lambda_{smpl} = 1$, $\lambda_{2D} = 1$, $\lambda_{3D} = 1$, and $\lambda_{KL} = 0.001$. We use a batch size of 128, a learning rate of $3e-4$ and the Adam optimizer in training. In each batch, half of the

* Georgios Georgakis and Ren Li are joint first authors and contributed equally to this work done during their internships with United Imaging Intelligence, Cambridge MA, USA. Corresponding author: Srikrishna Karanam.

data samples have only 2D annotations and the other half have 2D and 3D annotations.

- HKMR_{MF} training parameters: We follow the same configuration as SPIN [1].

2 Additional results

In Figure 1, we show some qualitative renderings of the estimated shape and pose with our method (HKMR) and compare it with other baselines HMR [2] and CMR [3]. We note improved mesh fits with our method. In Figure 2, we show complete (degree of occlusion greater than 1) occlusion robustness bar plots, comparing our proposed method with SPIN. These results correspond to the experiment of Table 3 (left) in the main paper, where we only showed results with degree of occlusion set to 1. As can be noted from these results, our method outperforms SPIN, giving lower MPJPE across both protocols and various degrees of occlusion.

In Figure 3, we isolate the impact of occluding just the root chain joints. This is particularly important since our method first estimates the root parameters followed by other chain parameters conditioned on the root parameters. The results in Figure 3 correspond to Figure 6 (and the accompanying discussion) in the main paper, but with errors computed only across the four root chain joints (instead of all the joints). As can be noted from the results, our proposed method is more robust to occlusions across the root joints when compared to HMR [2] and CMR [3].

	Protocol #1		Protocol #2	
	MPJPE	Recon.	MPJPE	Recon.
SMPLify[4]	-	-	-	82.3
Pavlakos et al.[5]	-	-	-	75.9
NBF[6]	-	-	-	59.9
HMR[2]	87.97	58.1	88.0	56.8
Arnab et al.[7]	-	-	77.8	54.3
HoloPose[8]	-	-	64.28	50.6
CMR[3]	74.7	51.9	71.9	50.1
DaNet[9]	-	-	61.5	48.6
DenseRaC[10]	76.8	-	-	48.0
SPIN[1]	65.60	**44.1**	62.23	**41.1**
HKMR	71.08	52.1	67.74	50.1
HKMR_{MF}	**64.02**	45.9	**59.62**	43.2

Table 1. Comparison with the state of the art on Human3.6M.

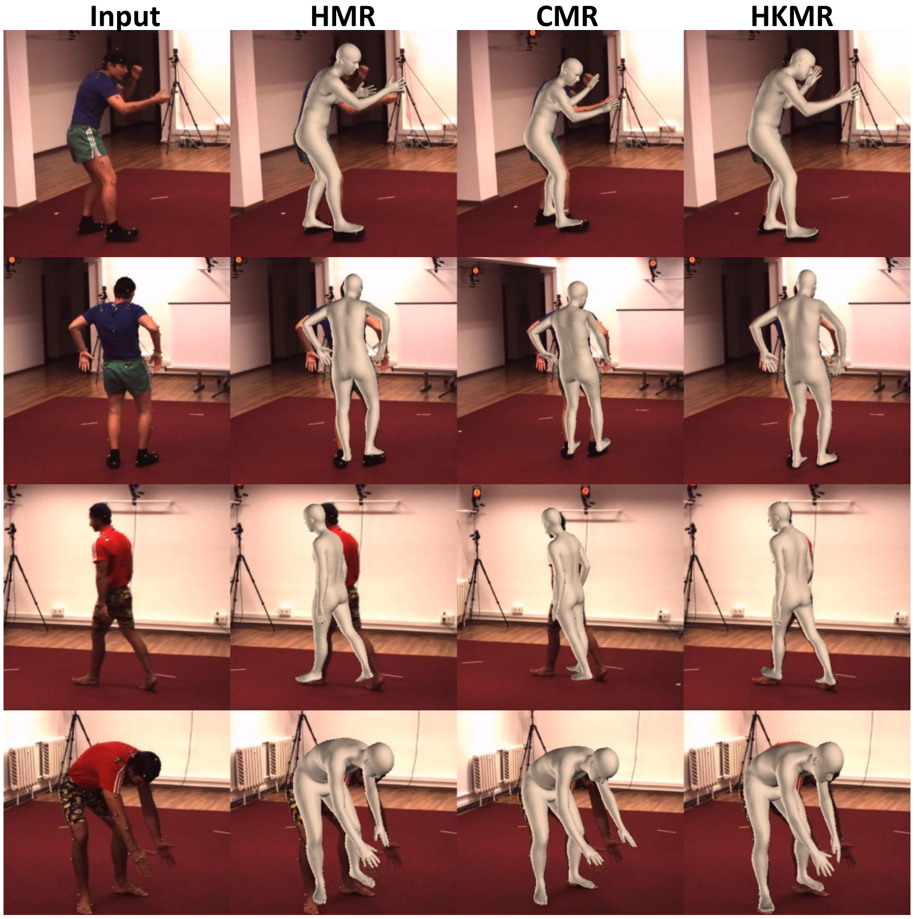

Fig. 1. Qualitative baseline architecture comparison: HMR vs. CMR vs. HKMR on the Human3.6M dataset.

In Table 1, we show both MPJPE and reconstruction error ("Recon." in table), i.e., MPJPE with Procrustes (rigid) post-processing, results on both protocols of the Human3.6M dataset, comparing our performance to the state of the art. These results correspond to Table 4 (right) in the main paper but with additional columns for the reconstruction error.

In Figure 4, Figure 5 and Figure 6, we show additional qualitative results of our proposed method on the LSP, MPII validation, and COCO validation datasets respectively.

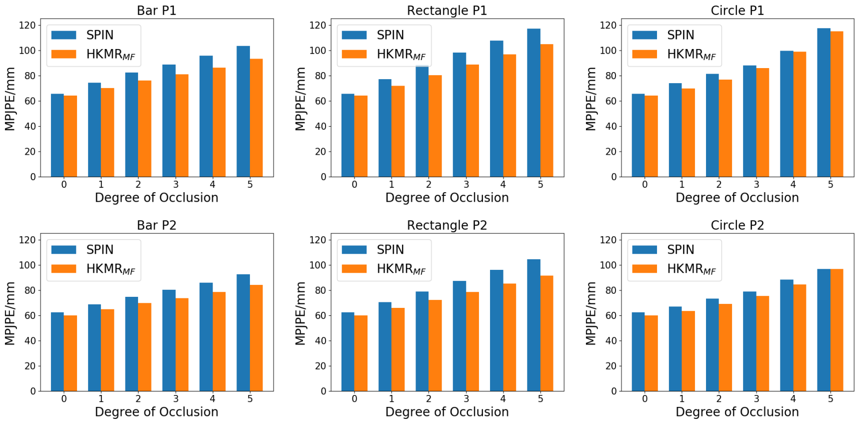

Fig. 2. Robustness to (single) occlusions: SPIN, our proposed method HKMR$_{MF}$.

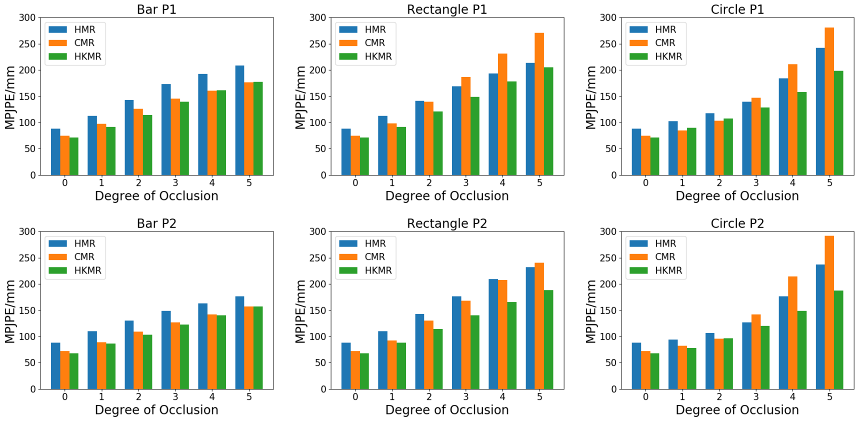

Fig. 3. Robustness to (root) occlusions: HMR, CMR and our proposed method HKMR.

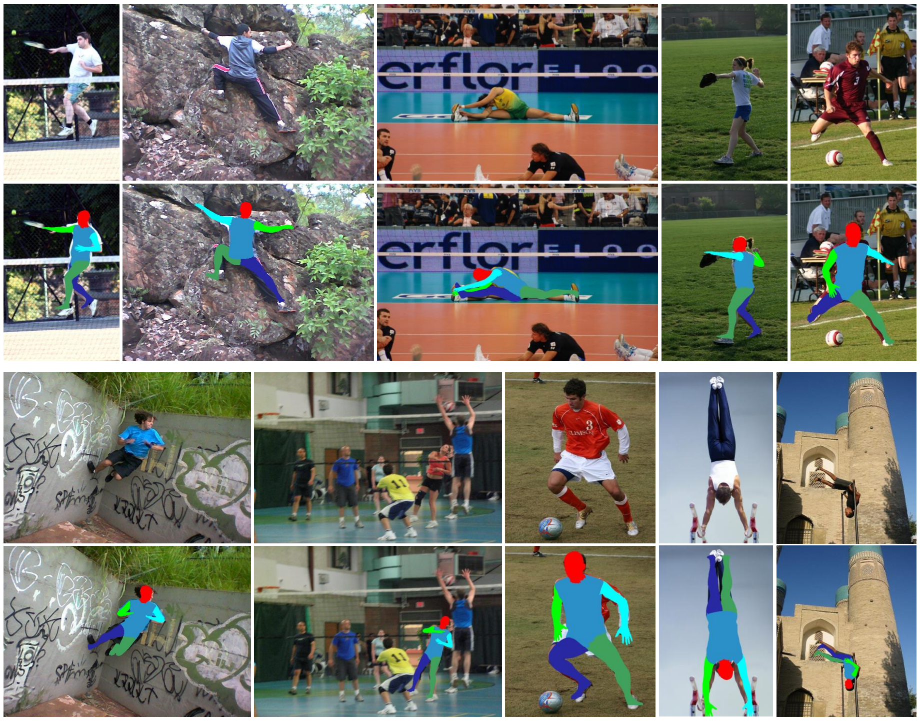

Fig. 4. Qualitative results on the LSP dataset.

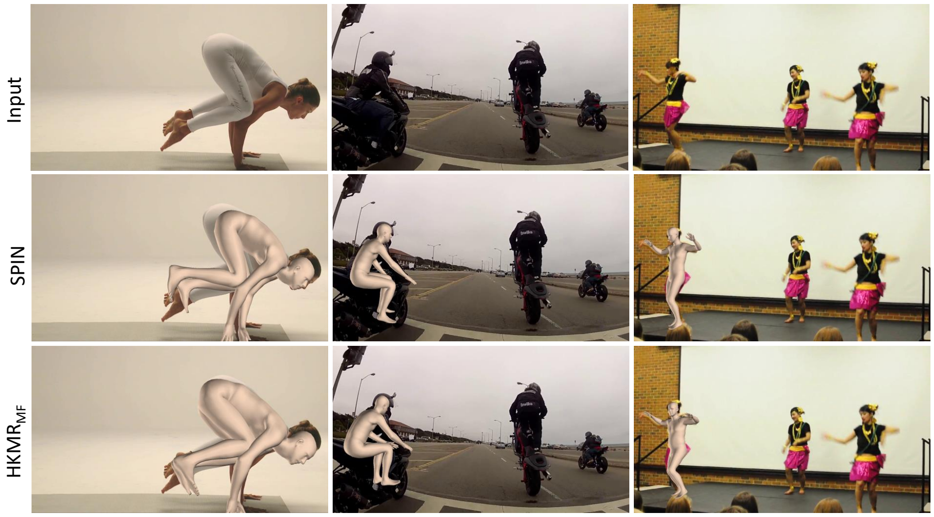

Fig. 5. Qualitative results on the MPII validation dataset with invisible joints.

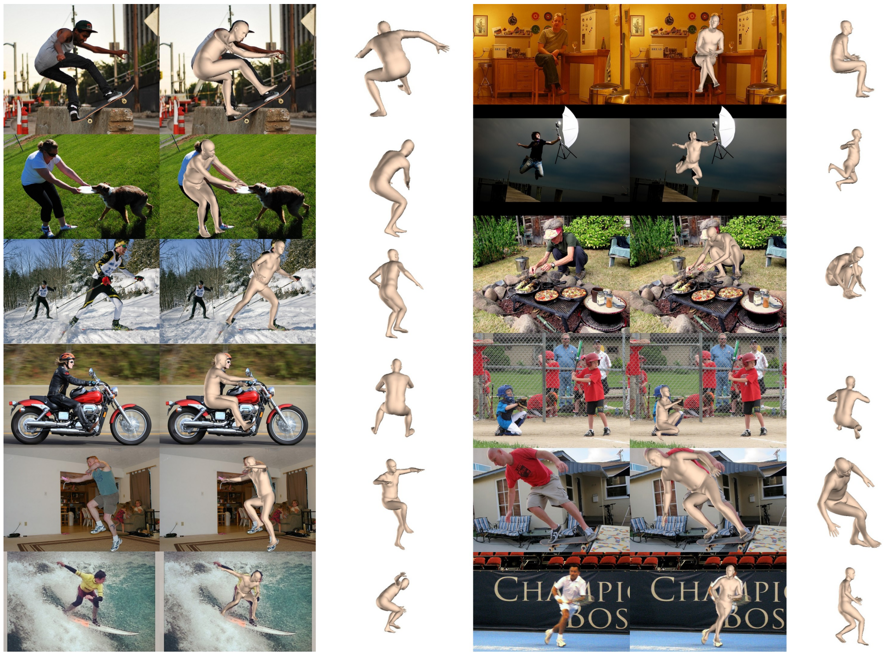

Fig. 6. Qualitative results with the side view on the COCO validation datatset.

References

1. Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019.
2. Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018.
3. Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019.
4. Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016.
5. Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, 2018.
6. Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, 2018.
7. Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *CVPR*, 2019.
8. Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3d human reconstruction in-the-wild. In *CVPR*, 2019.
9. Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. DaNet: Decompose-and-aggregate network for 3d human shape and pose estimation. In *ACM MM*, 2019.

10. Yuanlu Xu, Song-Chun Zhu, and Tony Tung. DenseRaC: Joint 3d pose and shape estimation by dense render-and-compare. In *ICCV*, 2019.