3D Human Mesh Regression with Dense Correspondence **Supplementary Material**

Wang Zeng[1], Wanli Ouyang[2], Ping Luo[3], Wentao Liu[4], and Xiaogang Wang[1,4]

[1]The Chinese University of Hong Kong [2]The University of Sydney [3]The University of Hong Kong [4]SenseTime Research
{zengwang@link, xgwang@ee}.cuhk.edu.hk, wanli.ouyang@sydney.edu.au, pluo@cs.hku.hk, liuwentao@sensetime.com

This supplementary material provides details not included in the main manuscript because of the space constrain. In Section 1, we present the training data used by different methods mentioned in the paper. In Section 3, we provide details about the weight map used in the computation of $\mathcal{L}_{map}$ in the main manuscript. In Section 2, we show some qualitative results on the SURREAL [18] test set and present qualitative comparison between our method and other state-of-the-art mesh-based methods.

1. Training data

As mentioned in the Section 4.2 of the main manuscript, the mesh-based methods we mentioned utilize different training data and the results are not directly comparable. In this section, we provides more details about the training data of these methods. We first introduce the related datasets bellow.

LSP-extended: LSP-extended [6] is a 2D human pose benchmarks containing 10,000 images with challenging human poses. For every image, 14 visible joint locations are annotated.

MPII: MPII [1] is a large scale 2D human pose dataset composed of over 25K images with annotated 2D joint locations. The MPII dataset contains over 40K people and covers 410 human activities.

MS COCO: For MS COCO [12], only the part of keypoints detection task is used, which contains over 150,00 people and 1.7 million annotated 2D keypoints.

MPI-INF-3DHP: MPI-INF-3DHP [14] is a recent 3D human pose estimation dataset captured by using a multi-view setup and synthetic data augmentation. For each image, ground-truth 3D keypoints locations are provided.

MOCA: MOCA [20] is a recent synthetic dataset including 2 million synthetic images with corresponding ground-truth 3D human body shapes and poses.

In Table 1, we present the training data used by each method when evaluated on the Human3.6M [4] test set. Pavlakos *et al*. [16] uses no training data from Human3.6M and trains 3D prior net using data from CMU MoCap [3], while NBF [15] only uses training data from Human3.6M. HMR [7], SPIN [9] and DenseRac [20] all utilize extra training data from 2D human pose benchmarks. SPIN and DenseRac additionally includes training data from the MPI-INF-3DHP dataset [14]. In addition, DenseRac makes use of synthetic data from MOCA [20]. Our method follows the setting of CMR [10], and uses training data from Human3.6M and UP-3D [11] without extra data from 2D human pose benchmarks. Our framework outperforms CMR with a large margin on the Human3.6M test set (the MPJPE of CMR and our method are 50.1 mm and 39.3 mm respectively).

Although SPIN and our method have similar performance on Human3.6 test set, the contributions are totally different. The impressive performance of SPIN can be attributed to its effective utilization of training data from 2D human pose benchmarks. However, our method focuses on the dense correspondence between 3D mesh and image, as well as the utilization of local image features. Therefore, SPIN and our method are complementary.

2. Qualitative results

In this section, we present some qualitative results of our method. Figure 1 shows some qualitative results of our method on the SURREAL [18] test set. Our method is able to reconstruct 3D human bodies with various shapes and poses.

Figure 2 shows some qualitative results of our method and other state-of-the-art methods on the test set of Human3.6M [4]. The state-of-the-art model-free method (*i.e*. CMR [10]) and model-based method (*i.e*. SPIN [9]) all estimate the full human body based on the global image feature extracted by CNN and may fail to reconstruct details which

Datasets	Pavlakos *etc*. [16]	NBF[15]	HMR [7]	SPIN [9]	DenseRac [20]	CMR [10]	Ours
Human3.6M [4]		✓	✓	✓	✓	✓	✓
LSP [5]	✓		✓	✓	✓		
LSP-extended [6]	✓		✓	✓			
MPII [1]	✓		✓	✓	✓		
MS COCO [12]			✓	✓	✓		
MPI-INF-3DHP [14]				✓	✓		
MOCA [20]					✓		
UP-3D [11]						✓	✓

Table 1. The training data used by different methods when evaluated on the Human3.6M [4] test set. Our approach uses the same training data with CMR [10].

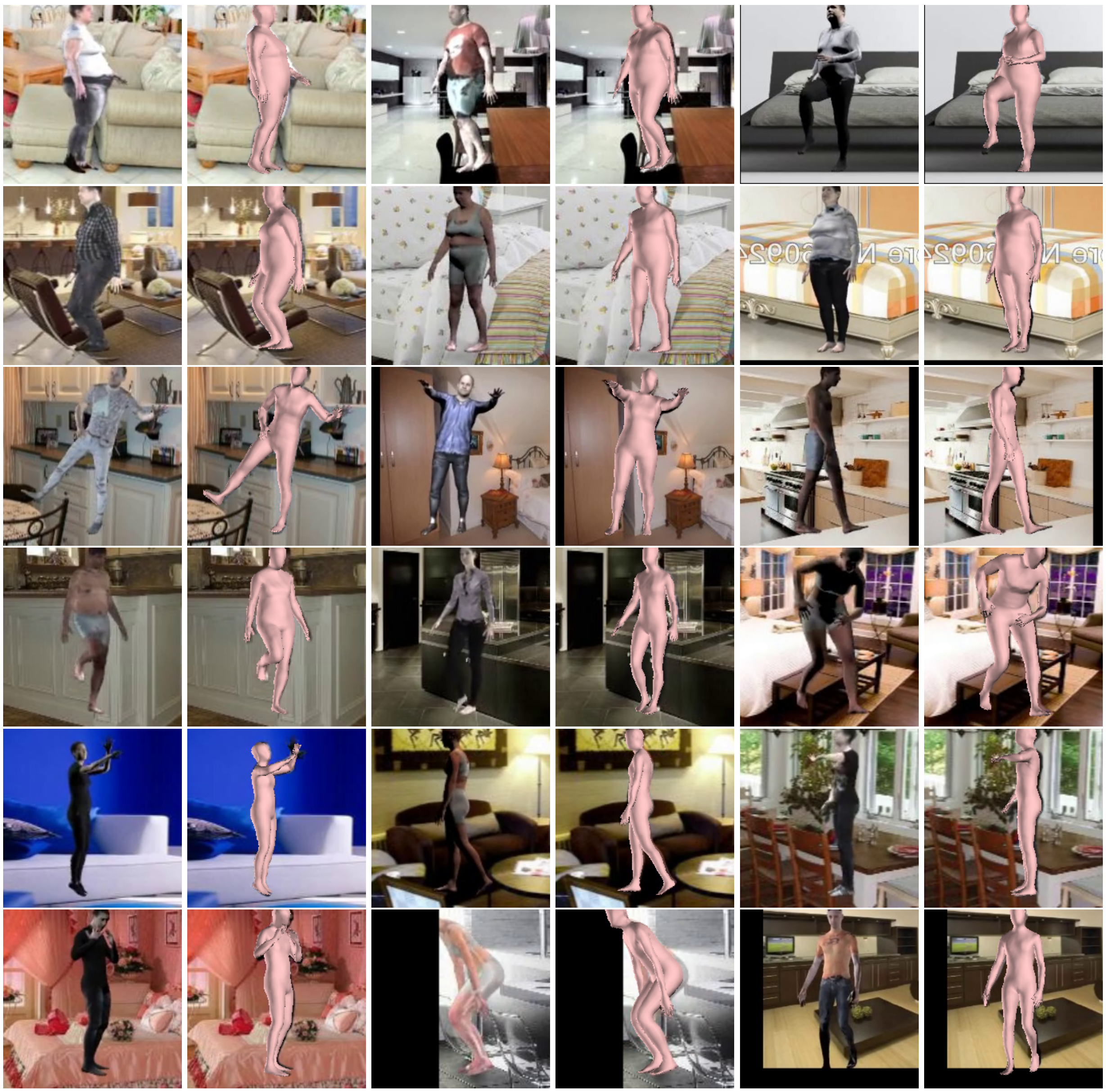


Figure 1. Qualitative results of our approach on the SURREAL [18] test set.

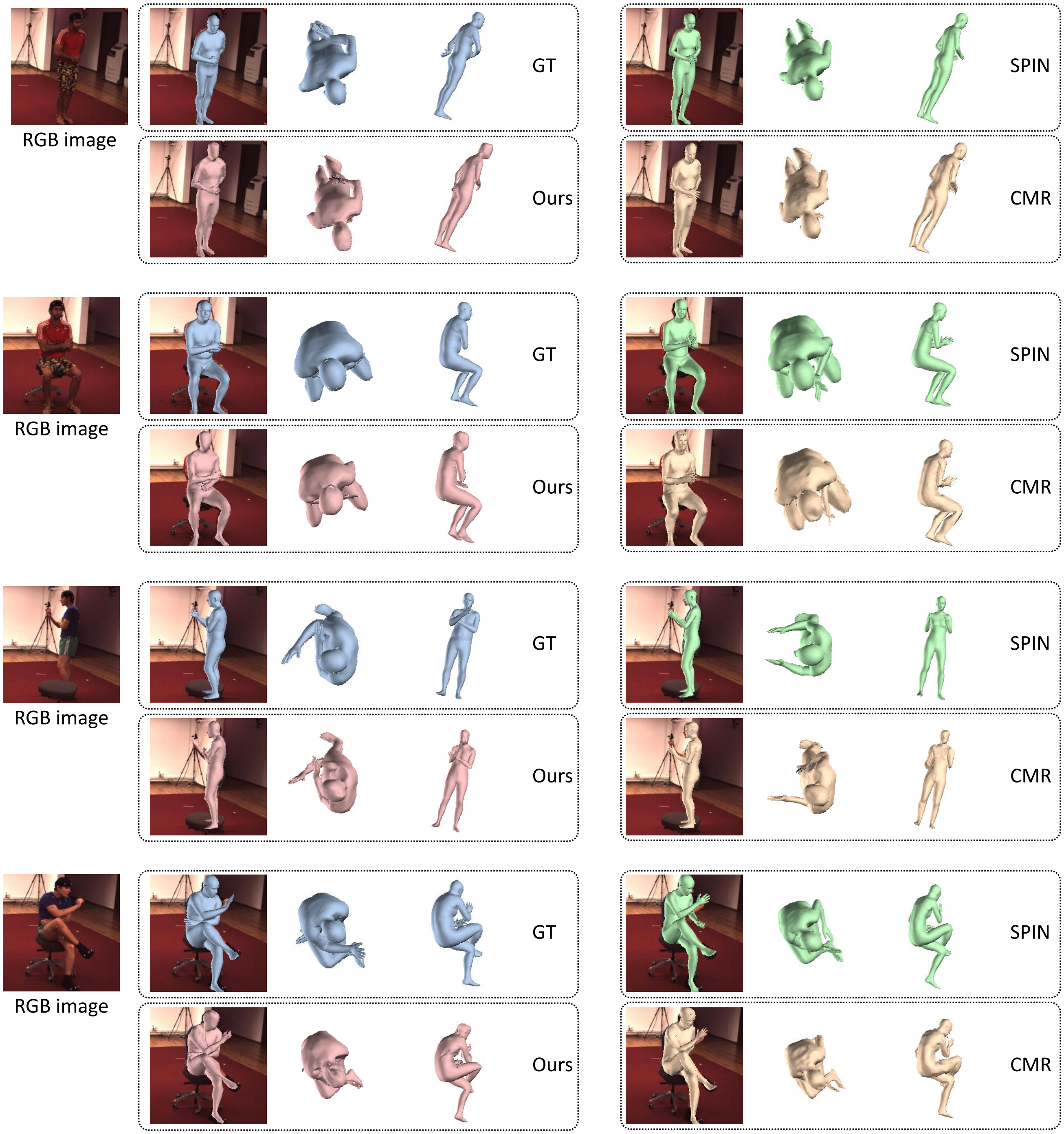
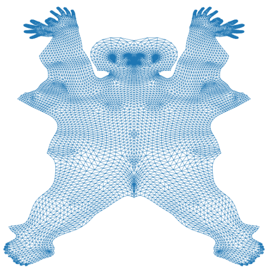


Figure 2. Comparison between our method and other state-of-the-art 3D mesh-based methods. CMR [10] and SPIN [9] may fail to reconstruct details which are not distinct on the image, while our method is able to reconstruct these details well.

are not distinct on the image. However, our method can utilize local image features with the explicitly established correspondence between mesh and image, and is able to reconstruct these details better.

3. Weight map

This section introduces the weight map for the loss term between regressed location map and ground-truth location map (*i.e.* $\mathcal{L}_{map}$). We assign larger weights to the parts away



(a) Mesh surface in UV space

(b) Weight map

Figure 3. Illustration of the weight map used for $\mathcal{L}_{map}$. Surface parts away from torso are assigned with larger weights.

from the torso, whose locations are with larger variance and more difficult for the network to estimate.

Specifically, we generate the weight map using the mesh part segmentation provided by SMPL model [13]. Different weights are assigned to different surface parts to get the weight map. Denote the weights for torso, neck&head, arms&legs, hands&foots respectively as λ_t, $\lambda_{n\&h}$, $\lambda_{a\&l}$, $\lambda_{h\&f}$. We set $\lambda_t : \lambda_{n\&h} : \lambda_{a\&l} : \lambda_{hs\&f}$ as $1:2:5:25$. Figure 3 shows the normalized weight map.

4. Evaluation on 3DPW dataset

3DPW [19] dataset is a recent outdoor 3D human body estimation benchmark. It provides 3D human pose and shape ground truth captured with IMU sensors. In our work, we only use its test set for evaluation.

Method	MPJPE-PA
HMR	81.3
CMR	70.2
[8]	72.6
[2]	72.2
[17]	69.9
Ours-A	**68.5**
Ours-B	61.7
SPIN	**59.2**

Table 2. Comparison with the state-of-the-art methods on 3DPW. SPIN and Ours-B utilize fitted SMPL parameters from SPIN for training, while other methods do not. Without using fitted SMPL parameters, our framework outperforms the methods using only global features. Utilizing fitted SMPL parameters further improves the performance to be competitive with the state-of-the-art method.

In order to investigate the generalization capability of our method, we evaluate our method on 3DPW test set. We use extra data from COCO [12], LSP [5] and MPII [1] as weak supervision to scale up our model (Ours-A) for fair comparison with prior works. We also train our model (Ours-B) with part of the fitted SMPL parameters from SPIN. The results are presented in Table 2. Without using fitted SMPL parameters, our model outperforms the methods using only global features. Utilizing fitted SMPL parameters further improves the performance to be competitive with the state-of-the-art. It is worth notice that we did not include the training data of LSP-extended and MPI-INF-3DHP as SPIN. Combining our method with the in-the-loop optimization process in SPIN may bring further performance improvement.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019.

[3] CMU. Graphics lab motion capture database. `http://mocap.cs.cmu.edu`, 2000.

[4] 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.

[5] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12.

[6] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[7] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[8] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019.

[9] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. *arXiv preprint arXiv:1909.12828*, 2019.

[10] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019.

[11] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6050–6059, 2017.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[13] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):248, 2015.

[14] 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 *2017 International Conference on 3D Vision (3DV)*, pages 506–516. IEEE, 2017.

[15] 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 *2018 International Conference on 3D Vision (3DV)*, pages 484–494. IEEE, 2018.

[16] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018.

[17] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5349–5358, 2019.

[18] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017.

[19] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018.

[20] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7760–7770, 2019.