Supplementary of 3D Human Pose and Shape Estimation Through Collaborative Learning and Multi-view Model-fitting

Anonymous WACV submission

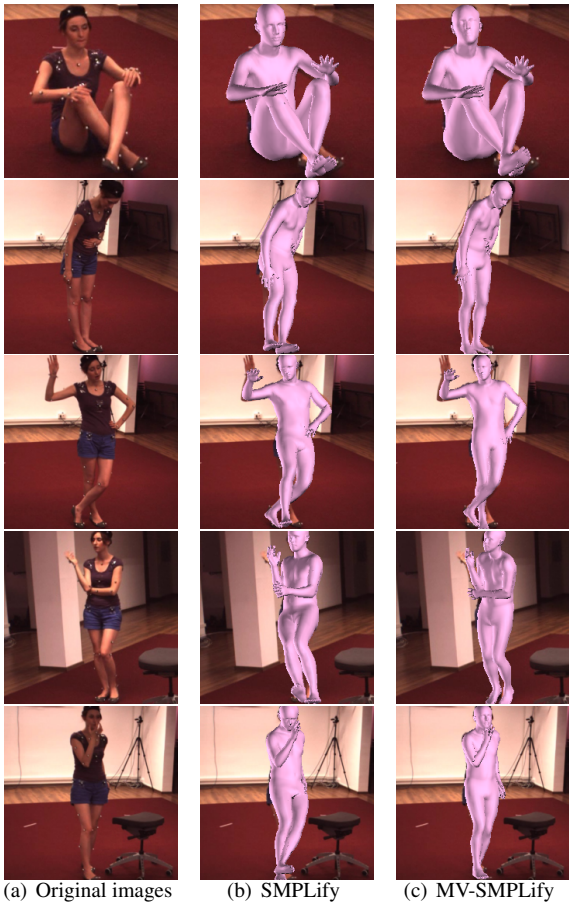
Paper ID 839

In the supplementary we provide more ablation studies on the MV-SMPLify part of our methods. The first part compares the performance of MV-SMPLify and classic SMPLifiy [1] on human pose and shape estimation. Then, more qualitative results from the Human3.6M [2], MPI-INF-3DHP [4] and 3DPW [5] of our method are given to show the performance of our method in more depth.

1. Comparison to SMPLify

As shown in the paper, MV-SMPLify is used to obtain the optimized SMPL models. In this section, we first compare the performance of MV-SMPLify and SMPLify[1] to demonstrate the advantage of using multi-view images. Taking 100×4 images from S1 in Human3.6M as an example, these images were fed into the CNN with the pre-trained parameters in [3]. Using the output of the CNN as initialization, we optimized the energy functions of the MV-SMPLify and SMPLify to get optimized pose and shape, respectively. Some examples from the 100×4 images are shown in Fig. 1. The second column in Fig. 1 shows the results of SMPLify, while the third column shows the result from the MV-SMPLify. We can see that the results from the MV-SMPLify better fit the ground truth and reduce the ambiguity of limbs in 3D space. Especially for the feet and body orientations, MV-SMPLify has more robust performance than the single image SMPLify. We also compute the reconstruction error, PCK and AUC of 3D joint points of the 100×4 images. The results are shown in Table 1 and Figure 2. We can see from Table 1 that MV-SMPLify can achieve higher PCK and AUC, while the reconstruction error is lower than when using a single image. Figure 2 gives the curve of PCK with different thresholds and it also shows that MV-SMPLify had higher AUC and PCK with 150 mm as threshold. Therefore, MV-SMPLify is more stable and reliable for our method and hence provides better supervision for training the CNN.

In addition, Figure 3 shows the comparison of the regressed SMPL model of CNN and optimized SMPL model obtained by MV-SMPLify. In the figure, the pink models



(a) Original images (b) SMPLify (c) MV-SMPLify

Figure 1. The results from SMPLify [1] and MV-SMPLify. From left to right: original image, SMPLify [1] and MV-SMPLify.

	PCK ↑	AUC ↑	Rec. Error ↓
SMPLify [1]	93.9	54.9	70.0
MV-SMPLify	**97.4**	**60.7**	**59.2**

Table 1. Comparison of the results from using single images and multi-view images, respectively.

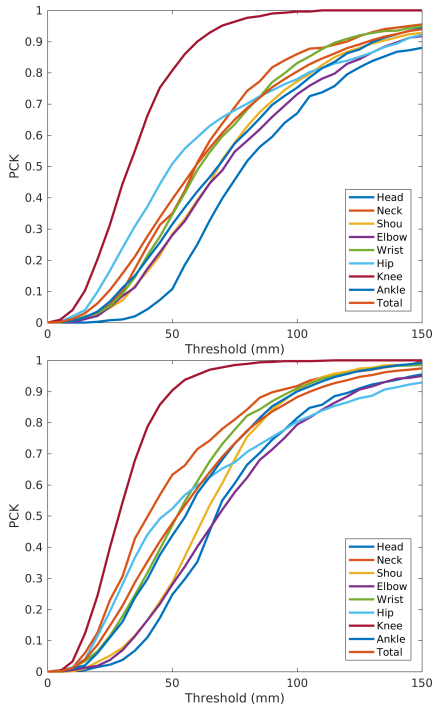


Figure 2. The AUC of SMPLify and MV-SMPLify for different joints. Top SMPLify and bottom MV-SMPLify.

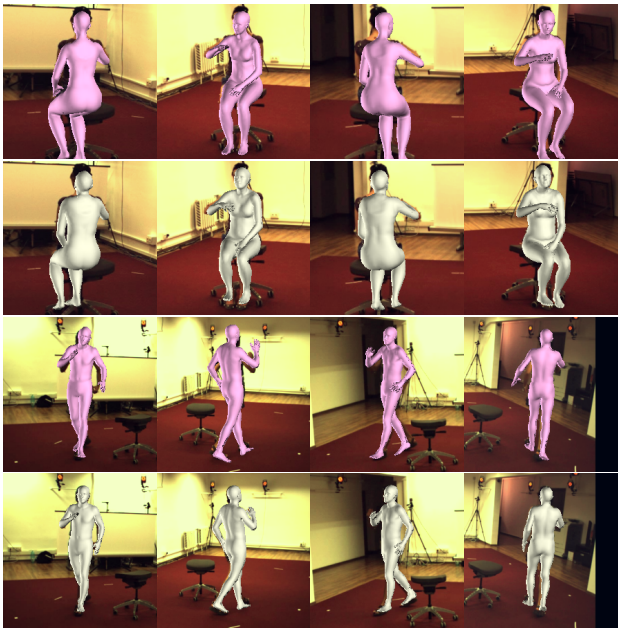


Figure 3. Comparison between the regressed and optimized SMPL model. The pink models are the results after regression. The white models are the the results after MV-SMPLify.

are the results of the CNN and the white models are the results after MV-SMPLify for the multi-view images. We can see that the results after MV-SMPLify are better, looking at the limbs of the optimized SMPL model, that are closer to the ground truth, especially for the results of the 3rd and 4th rows. This also demonstrates that it is advantageous to use the results of MV-SMPLify to supervise the training of the network, obtaining better estimation of the pose and shape.

2. Qualitative results

Extra results of our method from the Human3.6M [2], MPI-INF-3DHP [4] and 3DPW [5] are shown in Figure 4. These images show various poses and are captured under both indoor and outdoor scenarios. The first three rows are from Human3.6M and the middle three rows are from MPI-INF-3DHP. The last three rows are from 3DPW. The original image and the 3D model of our method (from different views) are given for each image. We can see that our method achieves promising 3D pose and shape estimation on the these images. Even for the 3DPW which only is used for testing, the estimated 3D models of our method are also satisfying. This figure demonstrates the effectiveness of our method.

References

[1] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *2014 European Conference on Computer Vision (ECCV)*, 2014. 1

[2] 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 Trans. Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 1, 2

[3] 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 *2019 International Conference on Computer Vision (ICCV)*, 2019. 1

[4] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. VNect: Real-time 3D human pose estimation with a single RGB camera. *ACM Transactions on Graphics*, 36(4), July 2017. 1, 2

[5] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 1, 2

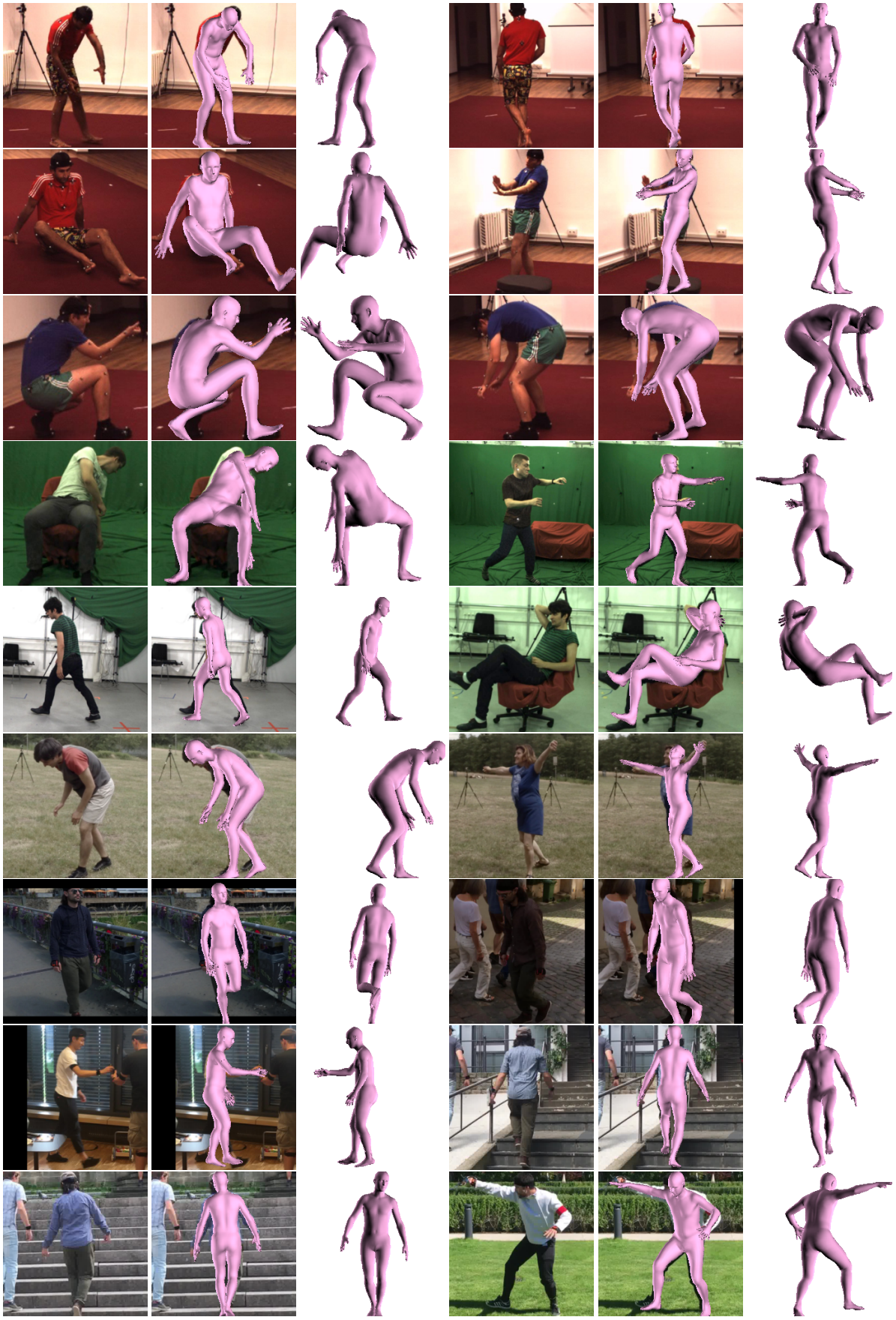


Figure 4. The results of our method on the three datasets. The first three rows are from Human3.6M, the middle three rows are from MPI-INF-3DHP and the last three rows are from 3DPW. For each example, the original image, the 3D model and the 3D model from another view are given.