

Appendix A. Extra Test Results

Table 6 and 8 shows the test results before Procrustes Alignment in MPI-INF_3DHP validation set and Human3.6M, respectively. The same conclusion about overfitting and multi-view improvement as the main text can also be drawn from these data.

Table 7 shows the result in MPI-INF_3DHP test dataset. Since there is only one view fed into the model, the results are similar.

Method	PCK/AUC/MPJPE w/ syn. training	PCK/AUC/MPJPE w/o syn. training
HMR [19]	66/33/141	71/36/129
Ours (single)	69/32/139	68/33/138
Ours (multi)	72/34/128	72/35/126

Table 6: Results on MPI-INF_3DHP, validation set, before Procrustes alignment.

Method	PCK/AUC/MPJPE w/ syn. training	PCK/AUC/MPJPE w/o syn. training
HMR [19]	65/30/139	65/29/137
HMR (PA)	84/47/91	85/48/89
Ours	65/29/142	66/29/137
Ours (PA)	85/49/89	86/49/89

Table 7: Results on MPI-INF_3DHP, test set. The results of [19] are tested on cropped images by Mask-RCNN [14] so the values have minor difference than their reported ones. Only single view is available in this dataset.

Appendix B. Additional Results on Real-World Images

As shown in Fig. 5, given similar joint estimation results, our model captures more image features that indicate the shape of the human body and thereby gives much better results in terms of human shape. We can distinguish between fat (Column 1-5) and slim (Column 6-8) persons, and between male and female. On the other hand, the output shapes from HMR are almost the same, which is around the mean shape value. By incorporating the shape-aware synthetic dataset, our method largely improves the recovery when the input human body does not have an average shape. We also tested with real-world multi-view images vs. single-view HMR. We feed the front view of the subject to HMR but input all views into our model. As shown in Fig. 6, the front view does not provide complete information of the subject pose, resulting in large pose errors on the limbs. By sharing information from more views (most importantly side views in this case), our model can effectively reduce the ambiguity from the camera projection and thereby provide good pose estimations across all views.

Appendix C. Comparison on Human3.6M with Single-View Methods

Table 8 shows the comparison with single-view results. As mentioned in the main text, the reason we don't have much better accuracy before rigid alignment is that:

- Our method does not assume known camera, resulting in an unknown scaling difference to the real-world co-

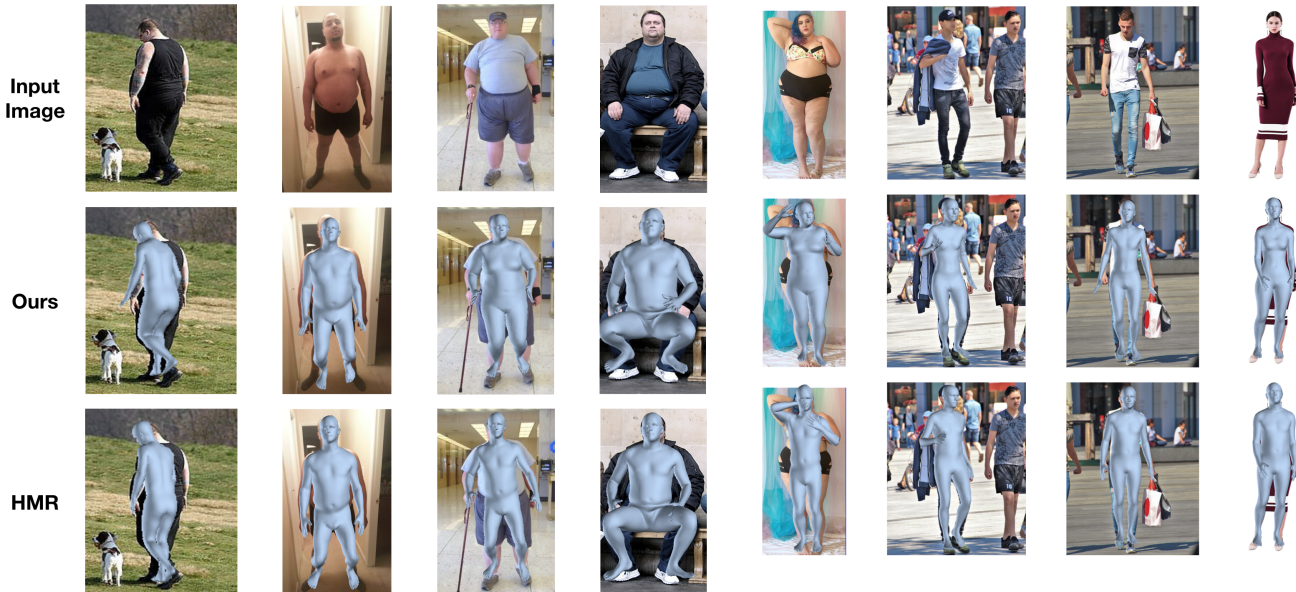


Figure 5: Results on images with varying pose and shape. The top row is the input image. The middle row shows our recovery results, and the bottom row shows the results from HMR [19]. Ours achieves better shape recovery results.

ordinates. After the Procrustes alignment, we achieved similar (and better with multi-view) performance.

- Our solution is constrained in a subspace. Other methods output joint positions directly so they have more DOF and can be more accurate. However, our output is more comprehensive, as it contains the entire human mesh in addition to joints and the result can be articulated and animated directly.

Compared to Kolotouros *et al.* [20], our model is trained on a much more diverse dataset (*e.g.* MS-COCO), which means that the accuracy may not be minimized on the specific subset (Human 3.6M).

Appendix D. Results Without Training on Synthetic Data

We further tested another variant of our model, which is trained without synthetic data (Fig. 7). It achieves better joint estimation, but the recovered human body does not seem to be visually correct, especially at the end-effectors. This is because the joint-only supervision does not impose any constraints on the orientations of the end-effectors, resulting in an arbitrary guess. The HMR model [19] avoids this by adding a discriminator, which however could have negative impact on shape estimations, as discussed in Sec. 4.4. Our synthetic dataset provides a supervision to not only the joint positions but also the rotations, hence the

Method	MPJPE	PA-MPJPE
Tome <i>et al.</i> [43]	88.39	-
Rogez <i>et al.</i> [38]	87.7	71.6
Mehta <i>et al.</i> [26]	80.5	-
Pavlakos <i>et al.</i> [31]	71.9	51.23
Mehta <i>et al.</i> [25]	68.6	-
Sun <i>et al.</i> [40]	59.1	-
Zhou <i>et al.</i> [55]	107.26	-
Debra <i>et al.</i> [9]	55.5	-
*Kolotouros <i>et al.</i> [20]	74.7	51.9
*Omran <i>et al.</i> [30]	-	59.9
*Pavlakos <i>et al.</i> [33]	-	75.9
*HMR [19]	87.97	58.1
*Ours (single-view)	88.34	58.55
*Ours (multi-view)	79.85	45.13

Table 8: Results on Human3.6M. Our method results in smaller reconstruction errors compared to HMR [19]. * indicates methods that output both 3D joints *and* shapes.

model will learn a prior at the end-effectors, demonstrating more natural results.

Appendix E. Detailed Errors on Real World Evaluation

The error percentages of each measure are shown in Table 9. Since the length of the arm and leg can be seen

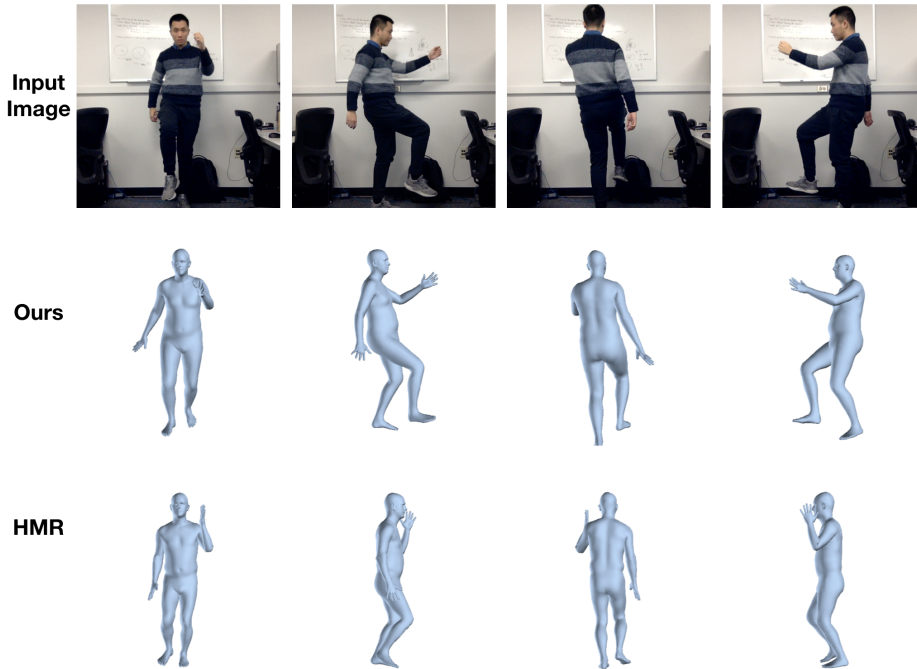


Figure 6: Results on real-world multi-view images. The top row is the input image. The middle row shows our recovery results, and the bottom row shows the results from HMR [19]. HMR is only given the front view as input. Ours achieves better pose recovery results due to more view angles.

error %	Regular				Dimmed				Partly Occluded			
input	Standing		Sitting		Standing		Sitting		Standing		Sitting	
# of views	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi
neck	1.12	12.19	0.048	3.53	0.58	11.31	0.39	2.55	0.45	11.28	22.11	6.11
arm	4.76	4.22	8.03	7.33	6.21	4.95	8.10	6.89	5.20	3.82	7.20	6.70
leg	6.65	4.66	2.94	3.46	5.18	3.92	2.83	3.64	2.53	3.54	4.94	4.24
chest	4.59	7.72	8.40	3.1	6.20	7.20	8.13	3.19	19.80	1.57	30.04	13.72
waist	2.42	12.80	5.46	0.70	3.73	11.98	5.01	0.0084	13.78	8.52	30.05	10.61
hip	8.88	0.62	11.88	5.83	11.36	0.12	11.78	5.50	15.08	1.65	15.95	7.54
error %	Regular				Dimmed				Partly Occluded			
input	Standing		Sitting		Standing		Sitting		Standing		Sitting	
method	HMR	BodyNet	HMR	BodyNet	HMR	BodyNet	HMR	BodyNet	HMR	BodyNet	HMR	BodyNet
neck	10.4	2.9	4.8	26.3	8.4	1.6	4.6	26.2	9.2	3.9	5.7	6.8
arm	6.1	21.3	9.8	25.6	8.6	22.8	9.7	23.6	8.1	19.5	9.7	9.6
leg	7.9	6.3	1.8	4.4	4.3	6.6	1.8	3.3	5.1	6.2	2.1	3.0
chest	11.2	26.3	11.7	51.9	11.7	24.9	11.6	41.3	11.9	24.9	11.6	21.3
waist	9.4	9.0	8.7	42.7	9.4	7.7	8.5	33.7	9.7	8.3	8.4	11.4
hip	1.25	19.2	7.8	79.8	3.5	18.8	7.7	80	2.9	17	5.5	36.9

Table 9: Percentages of errors in common measurements of the human body under various lighting conditions using single-view vs. multi-view images. The multi-view model performs significantly better in estimating measurements of chest, waist, and hip, and is more robust, given variations in lighting and partial occlusion.

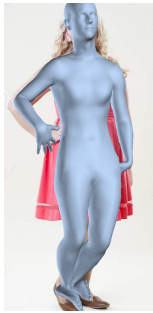


Figure 7: Our model trained without synthetic data.

clearly in the front view, both inputs provide a reasonably good estimation. However, given more views, our model can significantly reduce the error on other measurements, especially on those of chest, waist, and hip. We found that image illuminance has a negligible effect on the recovery result, which is due to the translation invariance of the convolutional layers. Occlusion has a notable impact on the recovery using only a single-view image, given only one view of the human body. However, by incorporating more views using our network model, the estimation can be considerably improved, indicating that the model using multi-view images is more robust to occlusion than with a single-view image as input.

Appendix F. Evaluation on *3D People in the Wild*.

We have conducted the evaluation on *3D People in the Wild* dataset. As shown in Table 10, although the dataset consists of single view images of only a few subjects with nearly standard shapes, our model achieved better accuracy

Method	Mean Joint Err.	Mean Vertex Err. (GT Pose)
HMR	93.77	21.71
Alldieck <i>et al.</i> [1]	169.61	47.07
Ours	96.86	20.96

Table 10: Evaluation on an unseen single-view dataset: *3D People in the Wild*. Values are mean joint error for pose and mean vertex error with ground-truth pose. We have smaller error than Alldieck *et al.*

over HMR, while Alldieck *et al.* did not generalize well. The metric we used is mean joint error for pose, and mean vertex error with ground-truth pose for shape.

Appendix G. Running Time

The previous work [19] trained 55 epochs for 5 days, while ours trained 20 epochs for 1 day. We list the training time here for reference, but it is actually not comparable since the batch size, epoch size and GPU type are not the same. In our environment, the inference time of HMR [19] is 2 microseconds while ours takes 7.5 (per view). This is because our network has a deeper structure to account for multiple views.