Self-Supervised Human Depth Estimation from Monocular Videos Supplementary Material

Feitong Tan[1,*] Hao Zhu[2,*] Zhaopeng Cui[3] Siyu Zhu[4] Marc Pollefeys[3] Ping Tan[1]
[1] Simon Fraser University [2] Nanjing University
[3] ETH Zürich [4] Alibaba AI Labs

In this supplementary material, we provide more information of our ablation study as well as additional qualitative results on the data in the wild.

1. Ablation Study

We provide more thorough ablation study in this section, including the performance of using poses captured by DoubleFusion [3] and qualitative comparison.

Using Poses Estimated by [3]. In the training of ReconNet, we replace the SMPL models estimated by TrackNet with those captured by DoubleFusion [3] and keep the other setting as same as Ours(M+SSIM$_{cs}$). DoubleFusion utilizes depth sensor to estimate the SMPL models, which can provide more accurate and stable SMPL coefficients than the predictions from neural networks with RGB images as input. We find the more accurate SMPL models (from DoubleFusion) lead to better performance of ReconNet, as shown in Table 1.

Table 1. Ablation study on Tang *et al.*'s test set. Please see text for more details.

Methods	Accuracy			MAE
	1.0cm	2.0cm	4.0cm	
Ours(Base Shape)	29.18	56.75	81.67	2.657
Ours(Baseline)	28.14	55.57	79.46	2.828
Ours(M)	28.52	56.35	80.67	2.714
Ours(M+SSIM$_{cs}$)	**31.47**	59.08	82.13	2.609
Ours(Captured Pose)	31.29	**59.36**	**83.28**	**2.544**

Finetuning the TrackNet. The TrackNet used in the experiments is finetuned from a original HMR model with our collected data. SMPL model estimation is not a solved problem yet. The original HMR model is trained using manually labeled 2D joints (LSP, COCO, MPII dataset) and 3D joints captured by MOCAP system (Human3.6M dataset). The manually labeled 2D joints lack depth information, and the data from MOCAP system contains only 8 actors who have to wear MOCAP markers. All these limitations lead to inaccurate SMPL prediction, as shown in Fig. 1 where the red arrows highlight the inaccurate results of HMR. Finetuning with additional data can improve its performance and help the self-supervised training of the ReconNet.

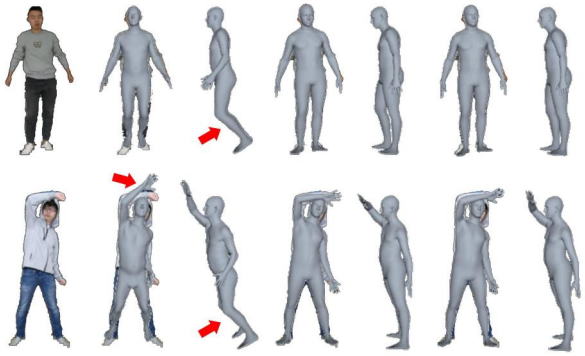

Figure 1. Intermediate SMPL models generated by the original HMR model, our fine-tuned one, and DoubleFusion. It is not a new experiment but an illustration.

Qualitative Results. To visually demonstrate the influence of the accuracy of SMPL models for our self-supervised learning and the effectiveness of our designed module for robust photometric loss, we show the quantitative comparison in Figure 2 and Figure 3. Comparing Ours(Captured Pose) and Ours(M+SSIM$_{cs}$), we can find both of them can recover small wrinkles especially for the main body, but Ours(Captured Pose) can capture more detailed geometry in trousers on examples (a), (b), (e), (h). Because the limbs are more difficult to track than the main body by the TrackNet, with more accurate SMPL models, our ReconNet can be trained with more stable photometric consistency loss in the limbs part. Without SSIM$_{cs}$ loss, Ours(M) prefers to generate smoother results, and we can find there are some artifacts in examples (c), (d), (f). The baseline results are very noisy and generate lots of wrong wrinkles. Without validation masks and SSIM$_{cs}$ loss, the training seems to become unstable.

*These authors contributed equally to this work.

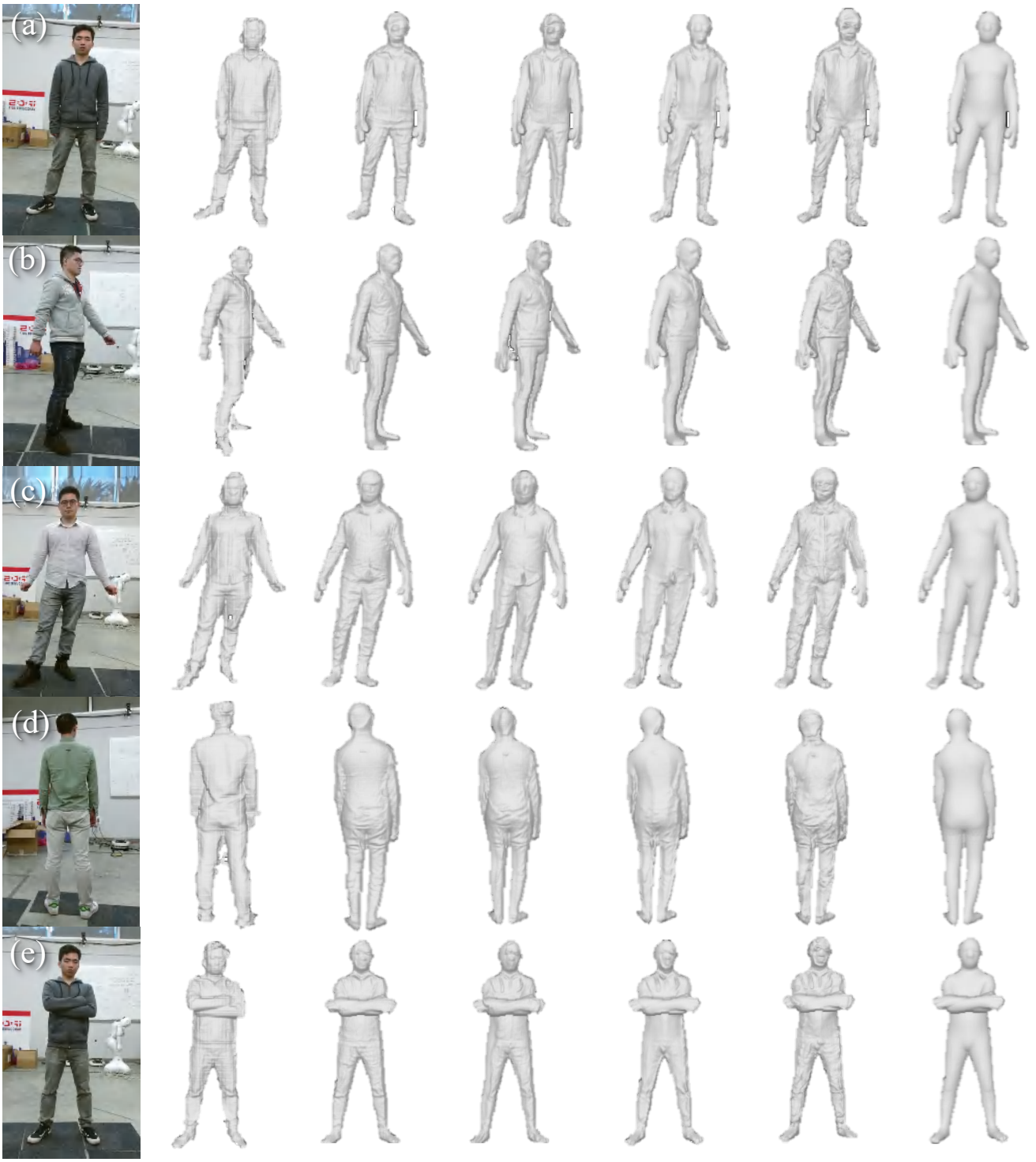

Figure 2. Comparison on Tang *et al.*'s testing datasets. From left to right, they are source image, ground truth, Ours(Captured), Ours(M+SSIM$_{cs}$), Ours(M), Ours(Baseline) and Ours(Base Shape).

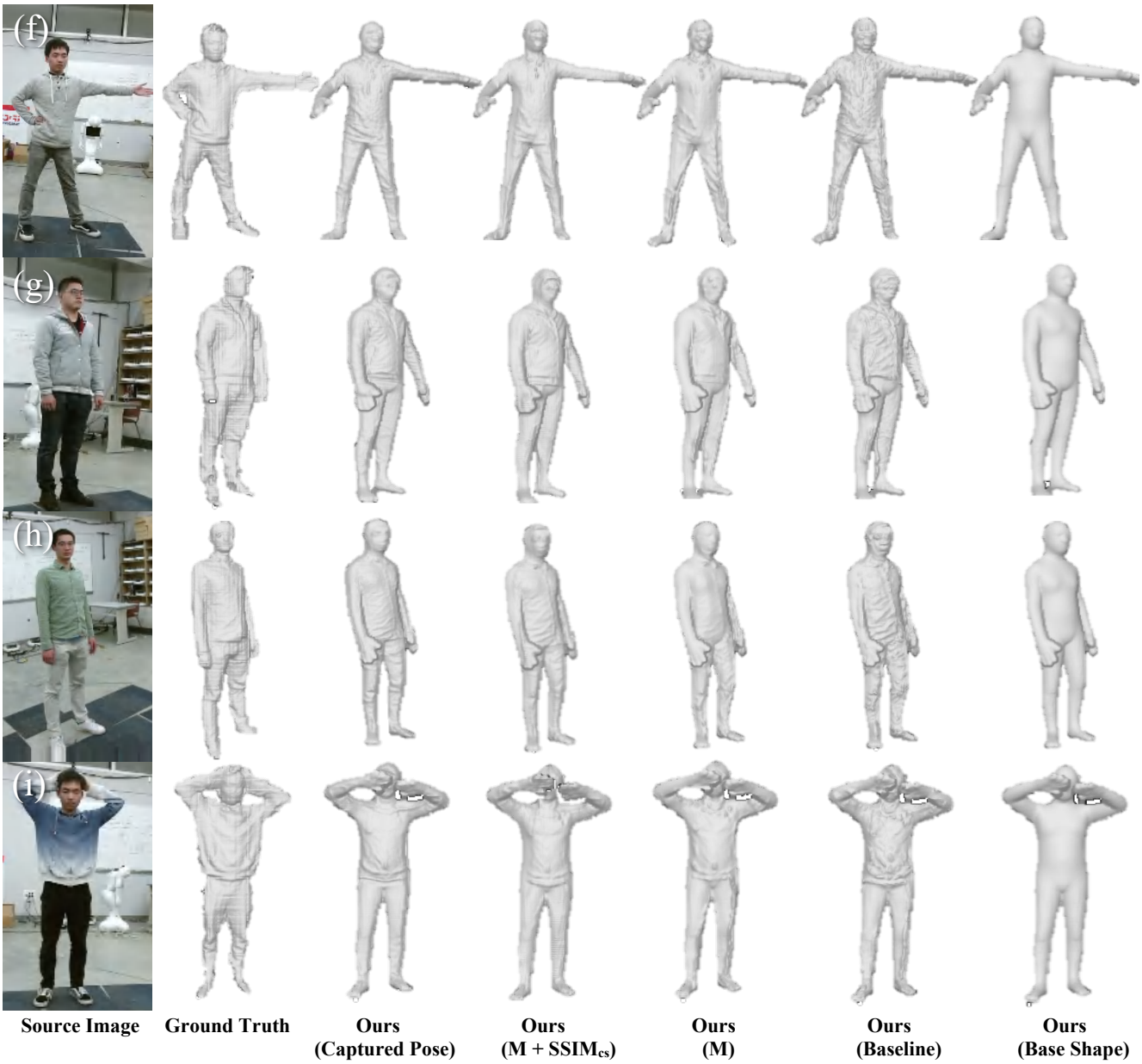

Figure 3. Comparison on Tang *et al.*'s testing datasets. From left to right, they are source image, ground truth, Ours(Captured), Ours(M+SSIM$_{cs}$), Ours(M), Ours(Baseline) and Ours(Base Shape).

2. Qualitative Results on Images in the Wild

We also compare our method with Tang *et al.*'s method [2] and HMD [4] on the images from the COCO dataset [1] and Internet, and the results are shown in Figure 4. From these results, we can see that our method can recover faithful details with stable layout in all wild images. By comparison, Tang *et al.*'s method predicts incomplete limbs in (f), (g) and (n), and doesn't recover plausible details. HMD suffers from implausible details such as wrinkles in (c), (e) and (j). The unstable performance of Tang *et al.*'s method and HMD indicates their generalization ability is poorer than our method. For some cases such as (c), (d) and (m), we find the top parts are better estimated than the bottoms parts. The main reasons can be concluded as follows: 1. The legs are usually less textured; 2. The SMPL model estimation is more precise at the main trunk (e.g. chest, waist, and shoulders) than the limbs, which leads to less shape details recovered at limbs; 3. Some Internet videos have an overhead

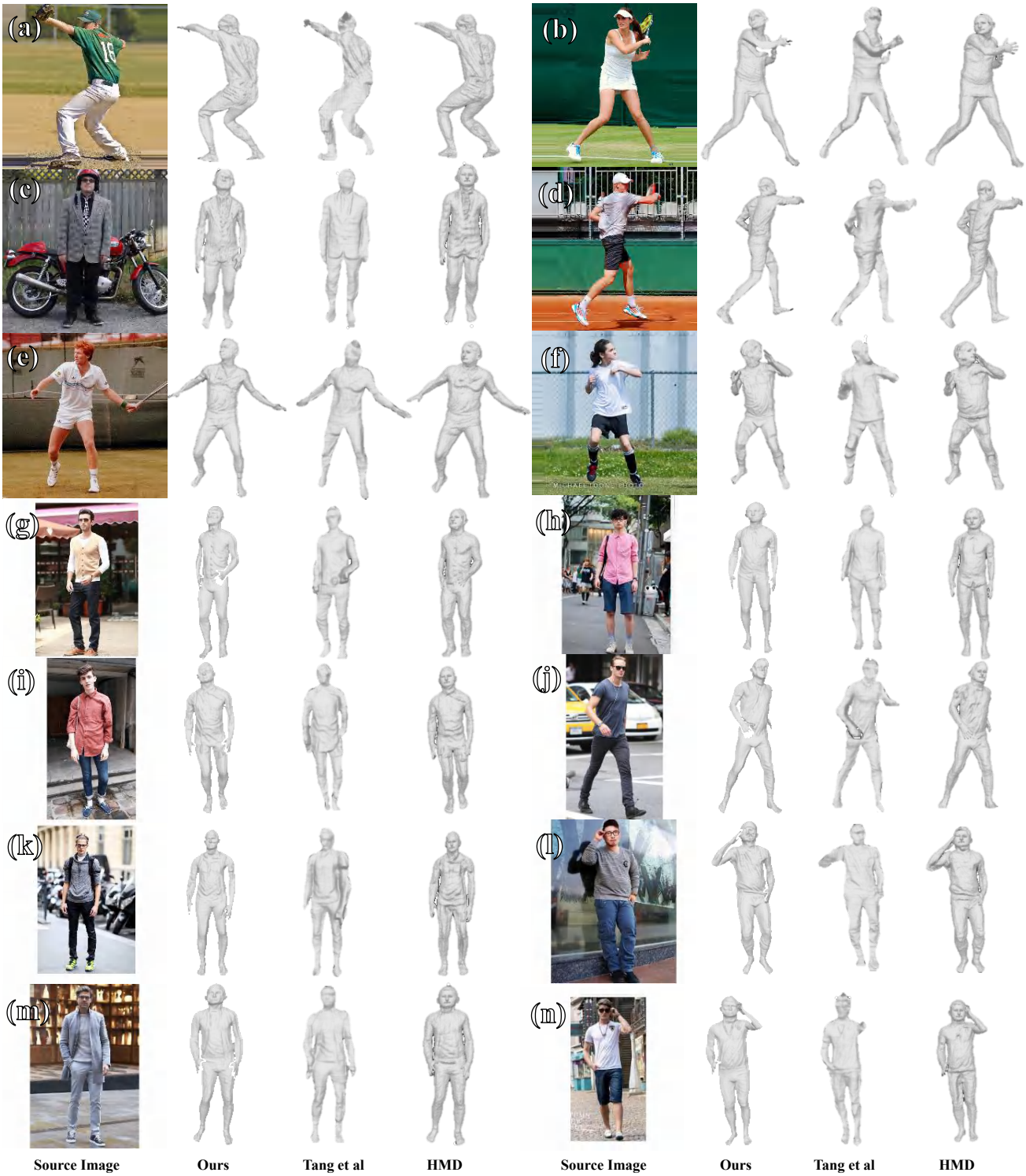

Figure 4. Comparison on the images in the wild. From left to right, they are source image, ours results, the results of Tang *et al*. [2] and the results of HMD [4]. (a) - (f) are from the COCO dataset [1], and (g) - (n) are from Internet.

lighting which makes the photometric consistency insufficient for the bottom part.

3. Results on Videos in the Wild

Our per-frame recovered depth on the videos in the wild are shown in the supplementary video. Our method predicts stable details and wrinkles for most of the wild videos, while the minor flaw is the unstable base shape prediction in the last sequence. By comparison, Tang *et al*.'s method [2] fails to predict complete limbs in many complex poses, and also recovers poorer details.

References

[1] 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 *Proc. of European Conference on Computer Vision*, 2014. 3, 4

[2] Sicong Tang, Feitong Tan, Kelvin Cheng, Zhaoyang Li, Siyu Zhu, and Ping Tan. A neural network for detailed human depth estimation from a single image. In *Proc. of International Conference on Computer Vision*, 2019. 3, 4, 5

[3] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 1

[4] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 3, 4