Tex2Shape: Detailed Full Human Body Geometry From a Single Image – Supplemental Material –

Thiemo Alldieck[1,2] Gerard Pons-Moll[2] Christian Theobalt[2] Marcus Magnor[1]

[1]Computer Graphics Lab, TU Braunschweig, Germany
[2]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
{alldieck,magnor}@cg.cs.tu-bs.de {gpons,theobalt}@mpi-inf.mpg.de

We show here additional experiments to understand the influence of illumination on our model and its robustness to varying camera intrinsics. We evaluate the β-regression network and perform an ablation of the UV map resolution. Finally, we present more qualitative results.

1. Influence of Illumination

As already emphasized in the main paper, shading is potentially a strong cue for our model. In the following, we evaluate the illumination augmentation during training and the robustness of our model to varying illumination.

In order to evaluate the effect of the illumination augmentation during training, we re-trained our model with constant ambient illumination. This means we render the scans using the textures only. While being scanned, the subjects have been exposed to uniform lighting. However, shading is still present in wrinkles and smaller structures. This means, we cannot factor out shading effects completely. Nevertheless, in Fig. 4 we can see more consistent details for our final method, especially for the faces.

Our model should produce the same or at least a very similar result when applied on two different photos of the same person in the same clothing but under varying illumination. To validate illumination invariance of our model, we took 9 photos of two subjects while rotating the light-source around the subject. In Fig. 1 we show the different photos and a heat-map illustrating areas with high standard deviation. We see a consistent picture with varying details only in areas of likely fabric movement.

2. Influence of Camera Intrinsics

Camera intrinsics are mostly unknown at test time, especially for in-the-wild photos. The focal length is an important camera parameter, which can affect the results of our method. We have trained our model with a fixed focal length. To study the robustness of our method against varying focal length, we render our test set in A-poses with different focal length and distance to the camera. We keep the ratio between distance and focal length fixed, creating a *Vertigo Effect*. In Fig. 2, we report the mean vertex-to-vertex error of the naked SMPL model under varying focal length. Although the lowest error is obtained for the focal length assumed during training, different focal lengths increase the error only slightly, which demonstrates the robustness of our model.

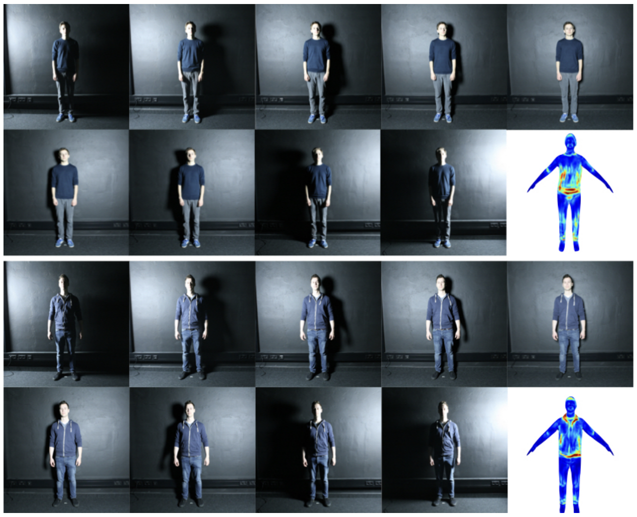

Figure 1. Displacement reconstruction consistency under varying illumination. The heatmap illustrates the vector norm of per surface point standard deviation (dark-red means $\geq$ 4cm).

3. Numerical Comparison with HMR

In order to evaluate the β-regression network, we compare our naked results without added displacements against HMR [3]. Since we do not estimate pose it has to be factored out before comparison. To this end, we follow the established procedure in [2] and adjust pose and scale

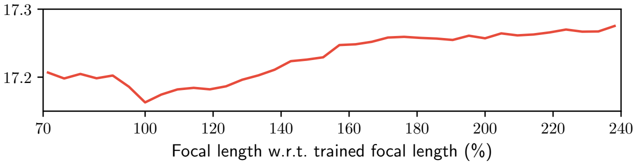

Figure 2. Mean SMPL vertex-to-vertex error in mm (without added displacements) over the test-set for varying focal length.

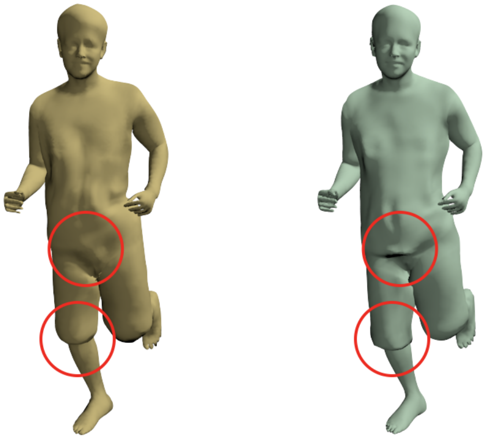

Figure 3. Comparison of two variants of our network: Using 256×256px resolution (left) decreased the quality only sightly, when compared to the original resolution of 512×512px (right).

of the results of both methods to match the ground truth scans. On our test-set, our method using DensePose mapping achieves a mean bi-directional vertex to surface error of 10.57 ± 10.68mm compared to the clothed scans. HMR achieves 16.28 ± 17.05mm. Our method can better estimate the body shapes. This is likely linked to the fact, that our method directly uses dense image-space detections, while HMR correlates surface with bone-lengths. With added displacements, our method achieves 5.19 ± 6.36mm. All results are up to scale.

4. UV Resolution Ablation

To evaluate our choice of the UV resolution (512×512px), we train a variant of the network with 256×256px maps. The results look surprisingly good. A close inspection of the results reveals missing details and smoothed edges. An example is shown in Fig. 3. However, this experiment demonstrates that Tex2Shape can be trained with lower resolution without largely decreased quality.

5. Additional Qualitative Results

In Fig. 5, we show more in-the-wild results of our method on *MonoPerfCap* [4] and *PeopleSnapshot* [1] datasets.

References

[1] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3D people models. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018. 2, 3

[2] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *IEEE International Conf. on Computer Vision*, pages 2300–2308, 2015. 1

[3] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018. 1

[4] W. Xu, A. Chatterjee, M. Zollhoefer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics*, 2018. 2, 3

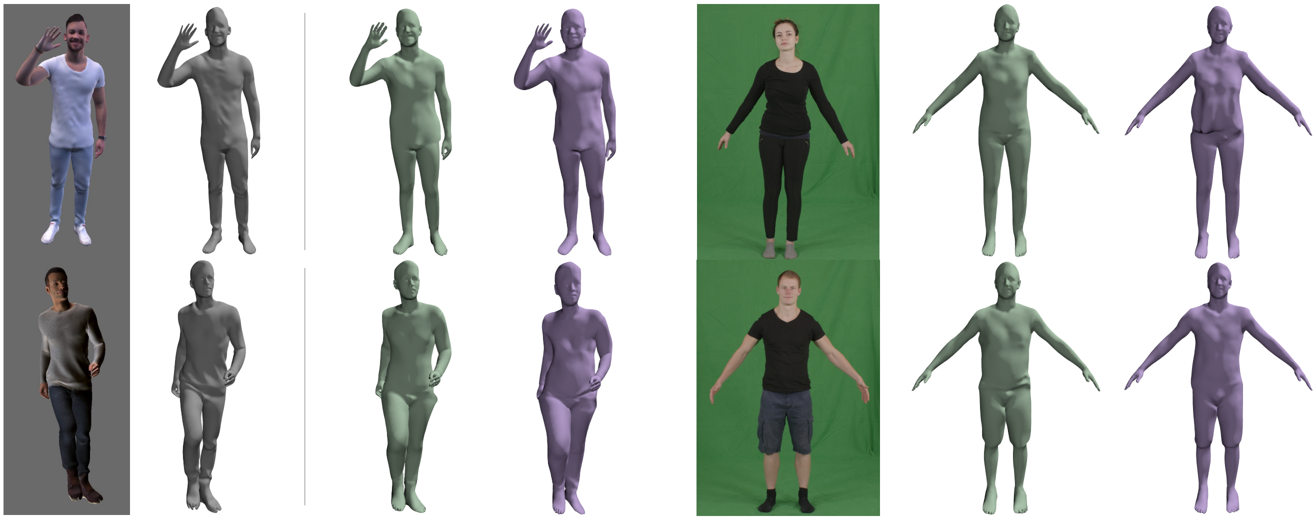

Figure 4. Our method (green) compared to our method trained without illumination augmentation (purple) and ground truth (grey). Looking closely, we notice worse performance specially on the face region, and artifacts for the method without illumination augmentation. Notice for example the example on the bottom left, the face, legs shape, and chest region is more accurately reconstructed when using augmentation (green).

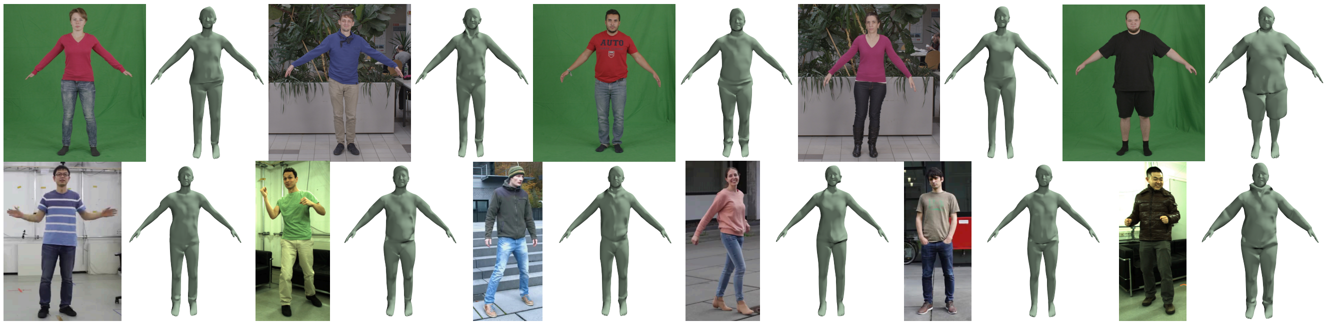

Figure 5. 3D reconstruction results on two in-the-wild datasets: PeopleSnapshot [1] (1st row) and MonoPerfCap [4] (2nd row).