Photorealistic Monocular 3D Reconstruction of Humans Wearing Clothing
– *Supplementary Material* –

Thiemo Alldieck Mihai Zanfir Cristian Sminchisescu

Google Research

{alldieck,mihaiz,sminchisescu}@google.com

In this supplementary material, we detail our implementation by listing the values of all hyper-parameters. Further, we report inference times, demonstrate how we can repose our reconstructions, conduct further comparisons, and show additional results.

1. Implementation Details

In this section, we detail our used hyper-parameters and provide timings for mesh reconstruction via Marching Cubes [6].

1.1. Hyper-parameters

When training the network, we minimize a weighted combination of all defined losses:

$$\mathcal{L} = \mathcal{L}_g + \lambda_e \mathcal{L}_e + \lambda_l \mathcal{L}_l + \mathcal{L}_a + \lambda_r \mathcal{L}_r + \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s + \lambda_{\text{VGG}} \mathcal{L}_{\text{VGG}}. \quad (1)$$

Further, we have defined the weights λ_{g_1}, λ_{g_2}, λ_{a_1}, and λ_{a_2} inside the definitions of $\mathcal{L}_g$ and $\mathcal{L}_a$. During all experiments, we have used the following empirically determined configuration:
$\lambda_e = 0.1$, $\lambda_l = 0.2$, $\lambda_r = 1.0$, $\lambda_c = 1.0$, $\lambda_s = 50.0$, $\lambda_{\text{VGG}} = 1.0$, $\lambda_{g_2} = 1.0$, $\lambda_{a_1} = 0.5$, $\lambda_{a_2} = 0.3$
Additionally we found it beneficial to linearly increase the surface loss weight λ_{g_1} from 1.0 to 15.0 over the duration of 100k interactions.

1.2. Inference timings

To create a mesh we run Marching Cubes over the distance field defined by f. We first approximate the bounding box of the surface by probing at coarse resolution and use Octree sampling to progressively increase the resolution as we get closer to the surface. This allows us to extract meshes with high resolution without large computational overhead. We query f in batches of 64^3 samples up to the desired resolution. The reconstruction of a mesh in a 256^3 grid takes on average 1.21s using a single NVIDIA Tesla V100. Reconstructing a very dense mesh in a 512^3 grid takes on average 5.72s. Hereby, a single batch of 64^3 samples takes 142.1ms. In both cases, we query the features once which takes 243ms. In practise, we also query f a second time for color at the computed vertex positions which takes 56.5ms for meshes in 256^3 and 223.3ms for 512^3, respectively. Meshes computed in 256^3 and 512^3 grids contain about 100k and 400k vertices, respectively. Note that we can create meshes in arbitrary resolutions and our reconstructions can be rendered through sphere tracing without the need to generate an explicit mesh.

2. Additional Results

In the sequel, we show additional results and comparisons. First, we demonstrate how we can automatically rig our reconstructions using a statistical body model. Then we conduct further comparisons on the PeopleSnapshot Dataset [1]. Finally, we show additional qualitative results.

2.1. Animating Reconstructions

In fig. 1, we show examples of rigged and animated meshes created using our method. For rigging, we fit the statistical body model GHUM [9] to the meshes. To this end, we first triangulate joint detections produced by an off-the-shelf 2D human keypoint detector on renderings of the meshes. We then fit GHUM to the triangulated joints and the mesh surface using ICP. Finally, we transfer the joints and blend weights from GHUM to our meshes. We can now animate our reconstructions using Mocap data or by sampling GHUM's latent pose space. By fist reconstructing a static shape that we then rig in a secondary step, we avoid reconstruction errors of methods aiming for animation ready reconstruction in a single step [4, 5].

2.2. Comparisons on the PeopleSnapshot Dataset

We use the public PeopleSnapshot dataset [1, 2] for further comparisons. The PeopleSnapshot dataset contains of people rotating in front of the camera while holding an A-pose. The dataset is openly available for research purposes. For this comparison we use only the first frame

Figure 1. Examples of reconstructions rigged and animated in a post processing step. We show the input image (left) and re-posed reconstructions (right). The reconstructions are rendered under a novel illumination.

of each video. We compare once more with PIFuHD [8] and additionally compare with the model-based approach Tex2Shape [3]. Tex2Shape does not estimate the pose of the observed subject but only its shape. The shape is represented as displacements to the surface of the SMPL body model [7]. In fig. 2 we show the results of both methods side-by-side with our method. Also in this comparison our method produces the most realistic results and additionally also reconstructs the surface color.

2.3. Qualitative Results

We show further qualitative results in fig. 3. Our methods performs well on a wide range of subjects, outfits, backgrounds, and illumination conditions. Further, despite never being trained on this type of data, our method performs extremely well on image of people with solid white background. In fig. 4 we show a number of examples. This essentially means, matting the image can be performed as a pre-processing step to boost the performance of our method in cases where the model has problems identifying foreground regions.

References

[1] https://graphics.tu-bs.de/people-snapshot. 1, 3

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8387–8397. IEEE, 2018. 1

[3] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2Shape: Detailed full human body geometry from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2293–2303. IEEE, 2019. 2, 3

[4] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Int. Conf. Comput. Vis.*, pages 11046–11056, 2021. 1

[5] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3093–3102, 2020. 1

[6] Thomas Lewiner, Hélio Lopes, Antônio Wilson Vieira, and Geovan Tavares. Efficient implementation of marching cubes' cases with topological guarantees. *Journal of Graphics Tools*, 8(2):1–15, 2003. 1

[7] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 2

[8] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 3

[9] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3d human shape and articulated pose models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6184–6193, 2020. 1

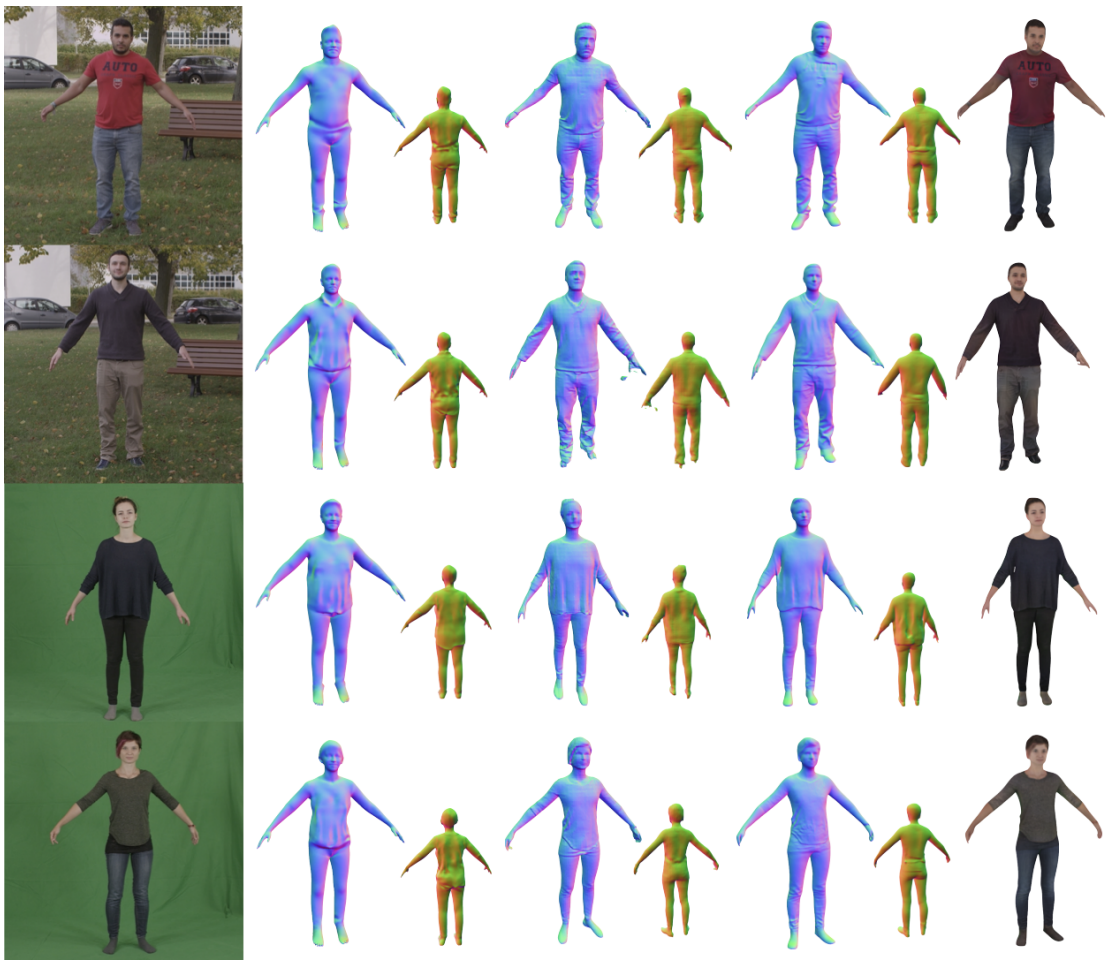

Figure 2. Qualitative comparison on the PeopleSnapshot dataset [1]. From left to right: Input image, geometry produced by Tex2Shape [3], PIFuHD [8], and PHORHUM (ours). We additionally show albedo reconstructions for our method.

Figure 3. Qualitative results on real images featuring various outfits, backgrounds, and illumination conditions. From left to right: Input image, 3D geometry (front and back), albedo reconstruction (front and back), and shaded surface.

Figure 4. Despite never being trained on matted images, our method performs extremely well on images with white background. From left to right: Input image, 3D geometry (front and back), albedo reconstruction (front and back), and shaded surface.