SUPPLEMENTARY MATERIAL
SMPLicit: Topology-aware Generative Model for Clothed People

Enric Corona[1] Albert Pumarola[1] Guillem Alenyà[1] Gerard Pons-Moll[2,3] Francesc Moreno-Noguer[1]
[1]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
[2]University of Tübingen, Germany, [3]Max Planck Institute for Informatics, Germany

In this supplementary material, we describe in detail the implementation of SMPLicit when training and fitting, we provide more qualitative results of fitting on-the-wild images, and discuss the current limitations that should be tackled by future work.

1. Implementation details

The implementation details are summarized in Section 5 of the main document. Here we provide further details that help to reproduce the training of SMPLicit and fitting processes. We will make our code publicly available.

Train data preparation. As mentioned in the main document, we resort to several publicly available datasets and augmentations. For the 3D clothig models that we downloaded from public links, we adjust them to a T-posed canonical body shape $\beta = \mathbf{0}$ and pre-process 100 random body variations, using SMPL's learnt body deformation displacement. In particular, we assign the deformation parameters of the closest SMPL vertex to each of the cloth vertices. At training, we randomly sample one of these models with probability 0.5, and otherwise we use the original data from BCNet [4].

We noticed that the original 3D models of pants often intersect between hips, specially for large SMPL shapes, which in many cases makes the two legs of the reconstructed garment connect. Notably, those produce artifacts when posing the body with the garment. To avoid this issue, we move from T-pose to a X-pose for training and inference only of lower-body models. Since the original training data already intersects, we repose the original pants by picking the skinning weights of the closest SMPL vertices, but also taking into account that their normal vector is similar to the garment model normal.

Training. For the cloth latent space, we set $|\mathbf{z}| = 18$ for upper-body, pants, skirts, hair and $|\mathbf{z}| = 4$ for shoes; the pose-dependent deformation parameters $|\mathbf{z}_\theta| = 128$, number of positional encoding clusters $K = 500$ and iso-surface threshold $t_d = 0.1$ mm. We clip the unsigned distance field to $d_{max} = 10$mm. The implicit network architecture uses three 2-Layered MLPs that separately encode $\mathbf{z}_{\text{cut}}$, $\mathbf{z}_{\text{style}}$ and $\mathbf{P}_{\boldsymbol{\beta}}$ into an intermediate representation before a last 5-Layered MLP predicts the target unsigned distance field, all of them using ReLU nonlinearities. SMPLicit is trained using Adam [5], with an initial learning rate 10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$ for 1M iterations with linear LR decay after 0.5M iterations, on a Nvidia® GTX 1080. We use $BS = 12$, $\sigma_n = 10^{-2}$ and refine a pre-trained ResNet-18 [2] as image encoder f. As [6], we use weight normalization [11] instead of batch normalization [3].

We observe there is a tradeoff between the number of positional encoding clusters and inference time, and set $K = 500$ clusters for a good compromise, as accuracy only improves marginally for higher values. For shoes, we set $K = 100$. We experimented using a KL-Divergence loss instead of L1 loss, but the model performed worse, particularly for garments where we have a small amount of data.

Fitting scans. The method for fitting 3D scans builds upon pointcloud cloth segmentation and the predicted 3D joint pose of the person. We schedule the fitting procedure, initializing β and θ to represent a T-Posed person from which only pose and translation are optimized for 200 iterations with a learning rate of $1e-2$. Then, the body shape β is also optimized with a decreased learning rate of $3e-3$ for 200 additional iterations. Finally, we reduce the learning rate of SMPL pose, shape and translation to $1e-3$ and also fit the upper-body and lower-body cloth parameters for 2000 iterations. During this last step, we let the LR decay linearly towards zero.

For the whole process, we leave the weights of SMPLicit untouched. In each iteration, the model optimizes the cloth parameters for a combination of 1000 uniformly sampled points and 1000 points from the cloth surface.

Fitting images. For fitting SMPLicit to images, we rely on the SMPL estimations from Frankmocap [8] and 2D

Figure 1: Reconstructions of the proposed method when adding realistic wrinkle effects on the fitted clothes.

Input Reconstruction 3D Layered Clothing Input Reconstruction 3D Layered Clothing

Figure 2: **More examples of fitting in multi-person images from the MPII Dataset [1].**

cloth semantic segmentation and instance segmentation obtained from [12]. We iterate over all SMPL detections, projecting the body model onto the image to identify the instance segmentation of the target person, which is used to

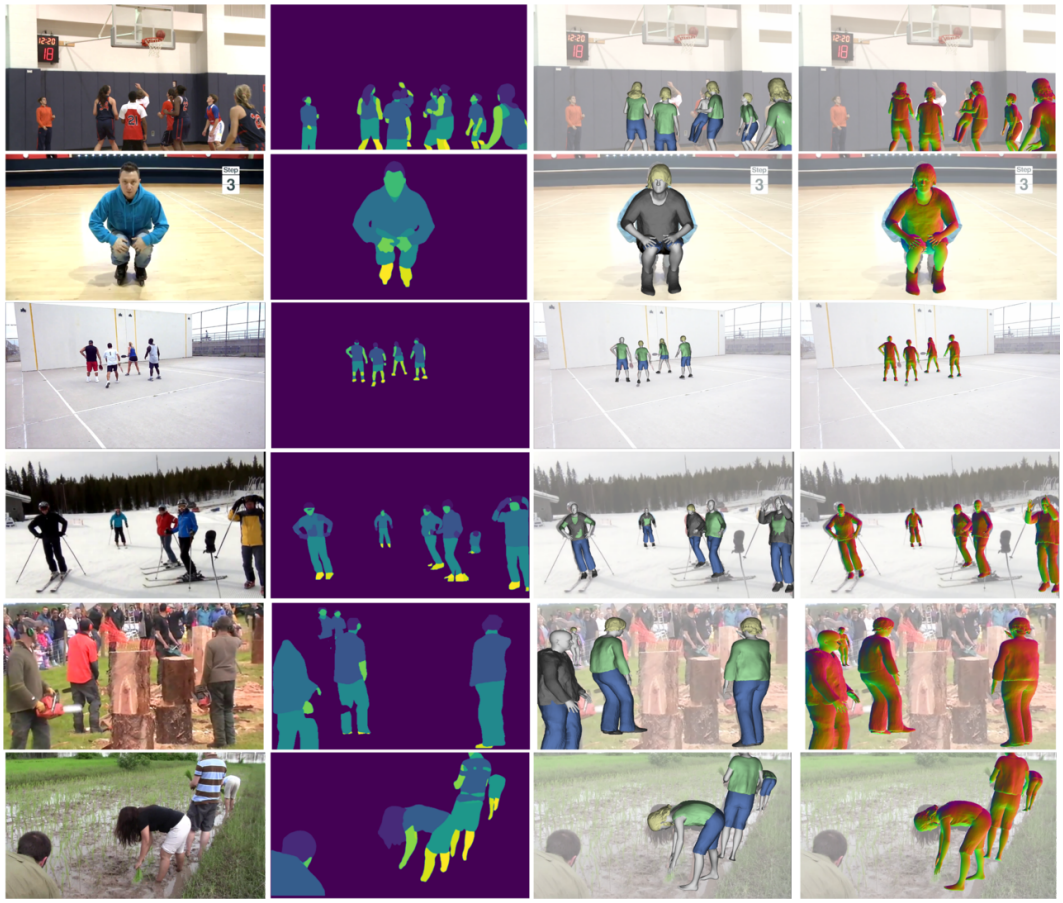

Figure 3: **Failure cases on images from the MPII Dataset [1].**

mask out other people's cloth segmentation.

We select the cloth types that have been segmented in the target person, removing those classes that have been found in less than 50 pixels, typically produced by a noisy segmentation. For each garment type, we uniformly sample points around the T-posed SMPL, and remove those that are too far from the body surface, provided that they are never close to clothing and would not contribute to training. To be more robust to occlusions to other persons or unsegmented objects, we also remove points whose projections do not fall into the instance segmentation mask from the target person.

Finally, for every garment class (upper-clothes, pants, etc), we also remove points whose projection falls into the segmentation class of a garment type that could occlude the target garment, according to a pre-defined sequence of clothes. For instance, we do not optimize T-shirt points that fall into the segmented area of jacket, hair or scarf, or pants points that fall under upper clothes or jacket. This proved particularly important when optimizing layers of clothes as a jacket occludes a large part of upper-clothes.

When fitting shoes, for simplicity we only fit the left-foot shoe and then mirror the reconstruction on T-Pose to generate the right one, before posing the clothes to the final model.

Each optimization takes approximately 90 seconds per cloth when using a uniform sampling of 128^3 points, for the MPII [1] images which we tackle directly at their original resolution (*e.g.* 720x1280 or 1080x1920). Inference time can decrease substantially by sampling less points when working with images of lower resolution.

2. High-Frequency details

Wrinkle details are important for representation of realistic avatars and accurate reconstructions. While not being our primary goal, we show it is also possible to improve SMPLicit models with high-frequency details. For this purpose we follow a similar strategy as in [9, 10, 13] and refine the initially estimated meshes with normal maps predicted from images. These normal maps are predicted using pre-trained PIFuHD's pix2pixHD net-work, thus we here consider humans in upright positions similar to those in their training set, however, the normal prediction network will fail when tackling people with more challenging poses or

noisy segmentations. Fig 1 shows the reconstructions enhanced with wrinkles.

3. Limitations

Regarding the process of generating and animating clothed humans, many previous works have devised two steps in which one first generates smooth garments consistently, and then adds pose-dependent wrinkles. We present SMPLicit-core as a very efficient model for generating different topologies, but it appears harder for implicit functions to provide high frequency deformations. We do leave the topic of cloth high-frequency deformations to further investigation, and actively encourage interested researchers to improve on our method.

We leveraged data very recently published [7, 4] and expect that SMPLicit will be able to represent more cloth variety when trained with larger databases, such as [14] which was not available at time of submission. Furthermore, SMPLicit is easily extendable to new garment types such as glasses or hats.

On the limitations of the image fitting method, Fig. 3 shows failure cases on images of the MPII Dataset [1]. Most wrongly optimized garments are due to inaccuracies and miss-alignments between estimated SMPL body shape and pose and 2D cloth segmentation (rows 1-3) or noisy cloth segmentation areas (Rows 4-5). Some of these problems could be tackled introducing SMPLicit within a deep-learning pipeline given enough training data. Overall, hair is the most difficult human component to represent as a mesh. We show an example of a challenging hair pose-dependent deformation in the last row.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 2, 3, 4

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[3] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 1

[4] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *ECCV*, 2020. 1, 4

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[6] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 1

[7] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *CVPR*, 2020. 4

[8] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020. 1

[9] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019. 3

[10] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 3

[11] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in neural information processing systems*, pages 901–909, 2016. 1

[12] Lu Yang, Qing Song, Zhihui Wang, Mengjie Hu, Chun Liu, Xueshi Xin, Wenhe Jia, and Songcen Xu. Renovating parsing r-cnn for accurate multiple human parsing. In *ECCV*, 2020. 2

[13] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *ICCV*, 2019. 3

[14] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *ECCV*, 2020. 4