Combining Implicit Function Learning and Parametric Models for 3D Human Reconstruction

Supplementary Material

We present IP-Net, **I**mplicit **P**art **Net**work which, given a point cloud (single view, sparse or dense) of a dressed human, reconstructs the outer 3D surface, inner body surface and predicts part-correspondences to the SMPL model. In the following sections we describe our body shape under clothing registrations which builds on top of the method proposed by [6] (code by [6] is not public). We then go on to show additional results for our method in subsequent sections.

1 Data preparation

Body shape under clothing registration. To train IP-Net we require body shape under clothing, for the inner surface prediction. Given a 3D scan, we propose an optimization based approach to register the body mesh, $\mathcal{B}$, under clothing. Our approach builds on [6]. Similar to [6], we model $\mathcal{B} = B(\cdot)$ using a modified SMPL, $M(\cdot)$, which uses pose($\boldsymbol{\theta}$), shape($\boldsymbol{\beta}$) and translation($\boldsymbol{t}$) parameters to model undressed humans in 3D.

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{t}) = W(T(\boldsymbol{\beta}, \boldsymbol{\theta}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) + \boldsymbol{t} \tag{1}$$

$$T(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathbf{T} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}), \tag{2}$$

where $\mathbf{T}$ is a base template mesh with 6890 vertices in a canonical T-pose. $B_p(\cdot)$ represents the pose dependent deformations of a skeleton $J(\boldsymbol{\beta})$. $B_s(\cdot)$ represents the shape dependent deformations. The model is skinned, $W(\cdot)$, with blend weights, $\mathbf{W}$.

We further make the template $\mathbf{T}$ optimizable to model surface variations outside the PCA shape space of the SMPL model. We incorporate translation, $\boldsymbol{t}$ in pose parameters, $\boldsymbol{\theta}$ for brevity in further notation.

$$\mathcal{B} = B(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{T}) = W(T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{T}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) \tag{3}$$

$$T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{T}) = \mathbf{T} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}). \tag{4}$$

We first segment garment and skin parts on the scans using the approach proposed by Bhatnagar *et al.*[3] and initialize the pose and shape parameters using registration proposed in [2,4]. We use a similar objective E_{skin} (Eq. 3 in [6]) to register the visible skin parts on the scans. To register skin parts underneath the garments we make slight modifications to the E_{cloth} term in Eq. 4 [6] by replacing the Geman-McClure cost function by a hinge cost and also add a geodesic

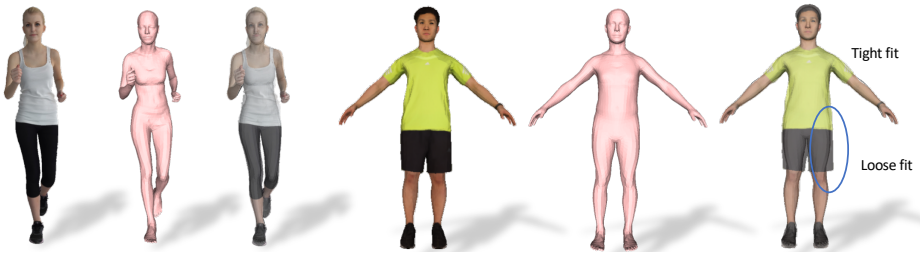

Fig. 1: To train IP-Net we require estimating body shape under clothing from a dressed scan. We show (L to R) input scan, estimated body shape, estimated body overlay-ed with scan.

term to force smoothness near the garment boundaries. The objective can be formally written as follows:

$$E_{cloth} = \sum_{\boldsymbol{v}_i \in \mathcal{S}} g_i * (1 - l_i) * (H(d_1(\boldsymbol{v}_i, \hat{\boldsymbol{p}}_i), c) + d_2(\boldsymbol{v}_i, \hat{\boldsymbol{p}}_i)) \tag{5}$$

$$H(x, c) = \begin{cases} x & \text{if } x < c \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$d_1(\boldsymbol{v}_i, \hat{\boldsymbol{p}}_i) = \begin{cases} d(\boldsymbol{v}_i, \mathcal{B}) & \text{if } \boldsymbol{v}_i \text{ is outside } \mathcal{B} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

$$d_2(\boldsymbol{v}_i, \hat{\boldsymbol{p}}_i) = \begin{cases} w * d(\boldsymbol{v}_i, \mathcal{B}) & \text{if } \boldsymbol{v}_i \text{ is inside } \mathcal{B} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $H(\cdot)$ acts as a hinge for loose clothing, $d_2(\cdot)$ and $d_1(\cdot)$ are the scaled distance functions to ensure that 'body is brought close to the garment surface' and 'body should not intersect the garment surface' respectively. l_i and g_i are the skin identifier label and normalised geodesic cost respectively. We use $w = 20$ and $c = 0.01$.

We additionally enforce facial landmark matching to register better facial details. To get 3D facial landmarks for a scan we render it from multiple viewpoints and run openpose [1] to get 2D facial landmarks on images. We then solve graphcut to lift the multi-view landmarks to 3D (this is similar to what [3] use for lifting 2D segmentation to scans). We use the following objective to match facial landmarks between the body and the scan

$$E_{face} = |L - f(\mathcal{B})|_2, \tag{9}$$

where L, f are facial landmarks on scan and SMPL facial landmark regressor respectively.

In order to ensure that $\mathcal{B}$ is smooth and retains human body like appearance we add the following regularization term. For the skin vertices it is important to ensure that the surface near the garment boundary is tightly coupled to the

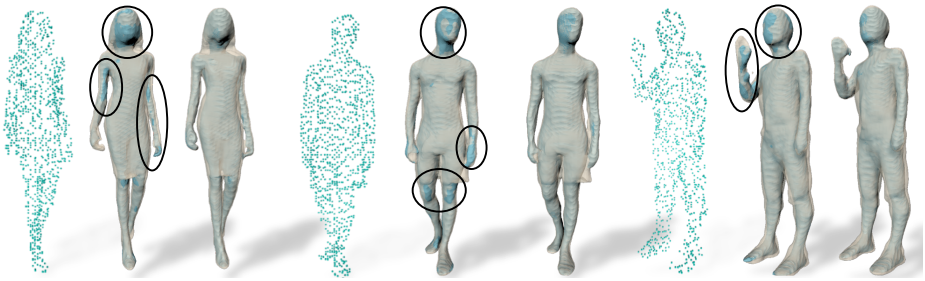

Fig. 2: Advantage of IP-Net being a joint model for inner body surface and outer surface. In each set (L to R) we show input point cloud, inner (blue) and outer (off-white) surface reconstruction by two independent networks, IP-Net reconstructions. reconstructions from IP-Net have visibly fewer inter-penetrations.

underlying body where as vertices away from the boundary can deform to explain hair, hands etc.

$$E_{lap} = \sum_{\boldsymbol{v}_i \in \mathcal{B}} \Big\{ (1 - l_i) * |L_i(\boldsymbol{v}_i^{init}) - L_i(\boldsymbol{v}_i)|_2 + \\ l_i * (1 - g_i) * |L_i(\boldsymbol{v}_i^{init}) - L_i(\boldsymbol{v}_i)|_2 \Big\}. \tag{10}$$

Here L_i is the laplacian operator at vertex $\boldsymbol{v}_i$. l_i and g_i are the skin label and normalised geodesic cost respectively.

Overall objective: We jointly optimise the SMPL parameters ($\boldsymbol{\theta}$, $\boldsymbol{\beta}$) and the template $\mathbf{T}$, to minimise the objectives described above.

$$E(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{T}) = w_{skin} E_{skin} + w_{face} E_{face} + w_{cloth} E_{cloth} + w_{lap} E_{lap}, \tag{11}$$

where w are the weights associated with the corresponding objectives. We found scheduling of weights important for a smooth registration process.

$$w_{\{skin/cloth/face\}} = c_{\{skin/cloth/face\}} * k,$$

$$w_{lap} = c_{lap}/k \tag{12}$$

In our experiments we keep c_{skin}, c_{face}, c_{garm} and c_{lap} as 5, 1, 5 and 100 respectively. k denotes optimization iteration. We show qualitative results in Fig. 1

2 Evaluating body shape prediction by IP-Net

We evaluate our body shape estimation on BUFF [6]. We report the main results in Sec. 4.4 (main paper). We provide more detailed results here in Table 1.

	t-shirt, long pants					soccer outfit			
tilt-twist-left	00005	00096	00032	03223	00114	00005	00032	03223	00114
Yang *et al.*[5]	17.29	18.68	13.76	17.90	15.42	16.77	16.96	20.41	16.40
Zhang *et al.*[6]	2.52	2.83	2.36	2.27	2.31	2.44	2.28	2.17	2.23
Ours	3.98	3.33	4.51	5.34	6.29	2.87	3.65	5.82	4.93
hips	00005	00096	00032	03223	0014	00005	00032	03223	0014
Yang *et al.*[5]	21.02	21.66	15.77	21.84	18.05	22.52	16.81	22.03	17.54
Zhang *et al.*[6]	2.75	2.64	2.63	2.40	2.56	2.58	2.50	2.38	2.51
Ours	3.83	3.86	3.47	5.96	7.31	3.23	3.56	5.72	5.06
shoulders-mill	00005	00096	00032	03223	0014	00005	00032	03223	0014
Yang *et al.*[5]	18.77	NA	18.02	18.15	14.78	18.74	17.88	19.74	16.37
Zhang *et al.*[6]	2.49	NA	2.72	2.26	2.59	2.83	2.28	2.33	2.51
Ours	3.40	NA	3.68	6.12	6.75	3.29	3.85	5.99	6.59

Table 1: Body Shape Evaluation on BUFF [6] We compare vertex-to-surface RMSE (mm). Note that [6] use 4D scan sequence to jointly optimize the shape of a subject whereas we make a prediction using just the first frame of the sequence. Moreover, we do not use BUFF for training. It is interesting to note that we report higher error on subjects 00114 and 03223. These are female subjects in the dataset and we trained IP-Net inner surface classifier with scans registered to male SMPL model. The error for male subjects is significantly low

3 Why not independent networks for inner and outer surfaces?

In Sec. 4.6 (main) paper we quantitatively show that having a joint model significantly reduces the inter-penetrations between the inner and the outer surface predictions. We show qualitative results in Fig. 2

4 IP-Net: Implementation details

The input to IP-Net is a 3D voxel grid obtained by voxelizing the sparse input point cloud into a 128x128x128 grid. IP-Net encoder $f^{\text{enc}}(\cdot|w_{\text{enc}})$, consists of 3x{Conv3D, Conv3D+stride} layers. IP-Net part predictor $f^{\text{part}}(\cdot|w_{\text{part}})$ and IP-Net part-conditioned classifiers $\{f^p(\cdot|w_p)\}_{p=0}^{N-1}$, each consist of 2x{FC} layers. All except the final layer of IP-Net have Relu activation. We use categorical cross-entropy losses with Adam optimizer for training.

5 Limitations

We discuss some important limitations of our approach in Fig. 3

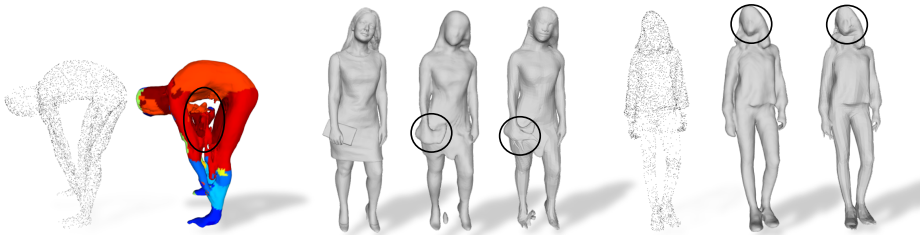

Fig. 3: We present some of the failure cases of our proposed approach. In the first set, we show the input point cloud and the generated surface reconstruction by IP-Net. Unseen poses are difficult for IP-Net. In the second set we show the GT scan with the person holding an object, the IP-Net reconstruction and the resultant registration with artefacts around the hand. Our approach cannot deal with non-clothing objects. In the third set we show the input point cloud, the IP-Net generated surface and the registration. Notice that facial details are missing.

- IP-Net struggles with out of distribution poses. In Fig. 3 first set, we have a person bending forward, and a similar pose was not present in our training set.
- Our registration fails in the presence of non-clothing objects.
- Facial details still need to be improved.

References

1. https://github.com/cmu-perceptual-computing-lab/openpose 2
2. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single RGB camera. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 1
3. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-garment net: Learning to dress 3d people from images. In: IEEE International Conference on Computer Vision (ICCV). IEEE (oct 2019) 1, 2
4. Lazova, V., Insafutdinov, E., Pons-Moll, G.: 360-degree textures of people in clothing from a single image. In: International Conference on 3D Vision (3DV) (sep 2019) 1
5. Yang, J., Franco, J.S., Hétroy-Wheeler, F., Wuhrer, S.: Estimation of human body shape in motion with wide clothing. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) European Conference on Computer Vision. Springer International Publishing (2016) 4
6. Zhang, C., Pujades, S., Black, M., Pons-Moll, G.: Detailed, accurate, human shape estimation from clothed 3D scan sequences. In: IEEE Conf. on Computer Vision and Pattern Recognition (2017) 1, 3, 4