Supplementary Material *for* "3D Clothed Human Reconstruction in the Wild"

Gyeongsik Moon[1*], Hyeongjin Nam[2*], Takaaki Shiratori[1], and Kyoung Mu Lee[2,3]

[1] Meta Reality Labs Research
[2] Dept. of ECE & ASRI, Seoul National University, Korea
[3] IPAI, Seoul National University, Korea
{mks0601,tshiratori}@fb.com, {namhjsnu28,kyoungmu}@snu.ac.kr

In this supplementary material, we present more technical details and additional experimental results that could not be included in the main manuscript due to the lack of space.

A Controlling reconstruction results

Our ClothWild has additional strength that the reconstructed results can be modified easily for other pose, shape, gender, and cloth style. In our framework, the pose, shape, gender, and clothes are disentangled by predicting their latent codes separately. Thus, we can edit the predicted latent codes as we want and forward them into the 3D cloth and human models (*i.e.*, SMPLicit [3] and SMPL [9]) of our framework. Fig. S1 shows an example of controlling a reconstructed 3D clothed human in three different categories. First, we can animate the reconstruction result by editing pose parameters represented by 3D rotations of human body joints (red part of the figure). Second, by modifying the shape and gender, we can change the naked human body of the reconstructed 3D clothed human (blue part of the figure). Third, clothes are represented by cloth existence scores and cloth latent codes of our framework, and we can change clothes by applying different cloth existence scores and cloth latent codes of other clothed humans (green part of the figure).

B Running time of ClothWild's components

Table S1 shows that the ClothNet has an extremely short running time compared to other components in our framework. Furthermore, a significant portion of the running time is occupied by components that generate 3D clothed humans from the latent codes, not our ClothNet. In our experiment in Table S1, the ClothNet and the SMPLicit are running on GPU (*i.e.*, RTX 2080 Ti), and the Marching Cubes and the Pose deformation are running on CPU (*i.e.*, Intel Xeon Gold 6248R) following the convention.

---

* equal contribution
This work was primarily done while Gyeongsik Moon was in SNU.

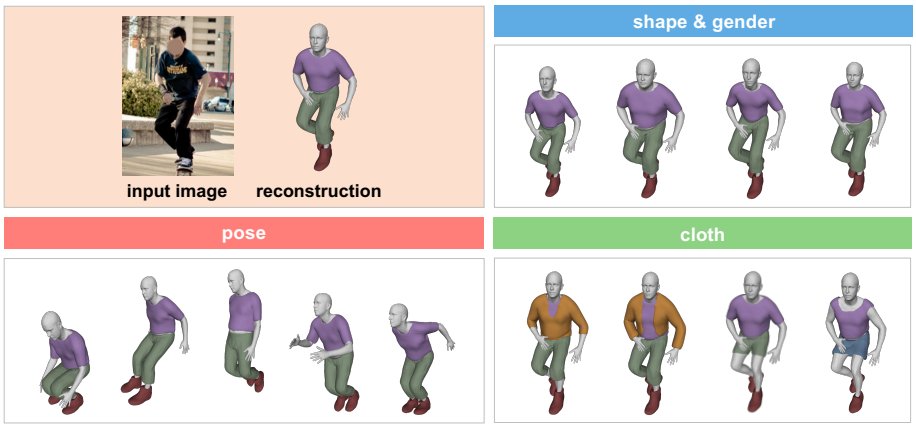


Fig. S1. An example for controlling the reconstructed 3D clothed human. Our reconstruction results are editable for pose, shape, gender, and cloth style.

Table S1. Running time of each component of our ClothWild, where the unit of time is second.

ClothNet	SMPLicit	Marching Cubes	Pose deformation	**Total**
0.01	1.08	3.82	5.10	**10.21**

C Detail of sampling strategy for query point selection

In the query point selection step, we uniformly sample 3D query points at a resolution of $21 \times 21 \times 21$ from a 3D bounding box for each cloth. The 3D bounding boxes are determined by 3D joints coordinates of the T-posed human body, where the 3D joints are defined in SMPL [9]. The 3D bounding box for each cloth and the joints is illustrated in Fig. S2. To the formal description, we define several notations. We denote 3D coordinates of a joint as $\mathbf{J}_{\bullet}$, and its x, y, and z value of the 3D coordinates are denoted as $J^x_{\bullet}$, $J^y_{\bullet}$, and $J^z_{\bullet}$. Additionally, we denote minimum and maximum of z values of T-posed human body vertices as v^z_{min} and v^z_{max}. The 3D bounding box for each cloth is represented as a corner representation $[x_{\text{min}}, y_{\text{min}}, z_{\text{min}}, x_{\text{max}}, y_{\text{max}}, z_{\text{max}}]$, where $\bullet_{\text{min}}$ and $\bullet_{\text{max}}$ are the minimum and maximum values of each coordinate at eight corners of the 3D bounding box. The 3D bounding boxes follow:

$$\begin{aligned}[x_{\text{min}}, y_{\text{min}}, z_{\text{min}}, x_{\text{max}}, y_{\text{max}}, z_{\text{max}}] = \\ [J^x_{\text{R\_hand}}, 2J^y_{\text{pelvis}} - J^y_{\text{chest}}, 1.25v^z_{\text{min}} - 0.25v^z_{\text{max}}, \\ J^x_{\text{L\_hand}}, 3J^y_{\text{chest}} - 2J^y_{\text{pelvis}}, 1.25v^z_{\text{max}} - 0.25v^z_{\text{min}}],\end{aligned} \tag{1}$$

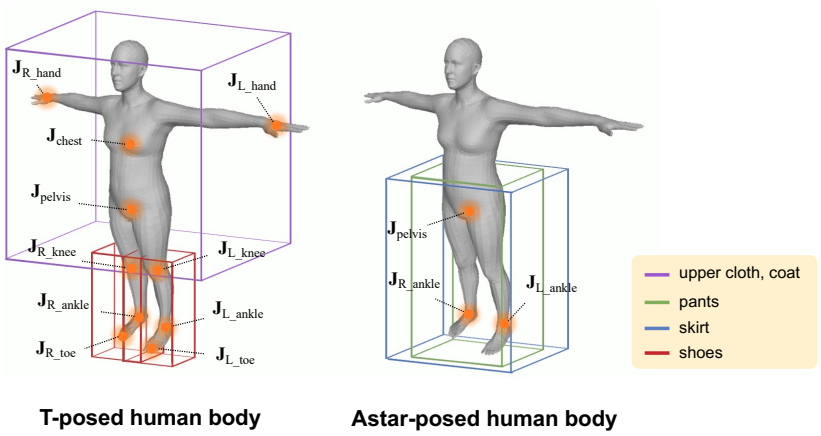


Fig. S2. Illustration of 3D bounding boxes for query point selection.

for upper cloth and coat,

$$\begin{aligned}&[x_{\min}, y_{\min}, z_{\min}, x_{\max}, y_{\max}, z_{\max}] = \\ &[J^x_{\text{L\_ankle}} - 0.15, 1.75J^y_{\text{L\_ankle}} - 0.75J^y_{\text{L\_knee}}, 0.5J^z_{\text{L\_ankle}} + 0.5J^z_{\text{L\_toe}} - 0.25, \\ &J^x_{\text{L\_ankle}} + 0.15, 0.25J^y_{\text{L\_ankle}} + 0.75J^y_{\text{L\_knee}}, 0.5J^z_{\text{L\_ankle}} + 0.5J^z_{\text{L\_toe}} + 0.25],\end{aligned} \tag{2}$$

for left shoe, and

$$\begin{aligned}&[x_{\min}, y_{\min}, z_{\min}, x_{\max}, y_{\max}, z_{\max}] = \\ &[J^x_{\text{R\_ankle}} - 0.15, 1.75J^y_{\text{R\_ankle}} - 0.75J^y_{\text{R\_knee}}, 0.5J^z_{\text{R\_ankle}} + 0.5J^z_{\text{R\_toe}} - 0.25, \\ &J^x_{\text{R\_ankle}} + 0.15, 0.25J^y_{\text{R\_ankle}} + 0.75J^y_{\text{R\_knee}}, 0.5J^z_{\text{R\_ankle}} + 0.5J^z_{\text{R\_toe}} + 0.25],\end{aligned} \tag{3}$$

for right shoe. The 3D bounding boxes for pants and a skirt are formed based on the Astar-posed human body, which legs are slightly wider than the T-posed human body, as follows:

$$\begin{aligned}&[x_{\min}, y_{\min}, z_{\min}, x_{\max}, y_{\max}, z_{\max}] = \\ &[2.3J^x_{\text{R\_ankle}} - 1.3J^x_{\text{pelvis}}, 1.1J^y_{\text{R\_ankle}} - 0.1J^y_{\text{pelvis}}, 1.25v^z_{\min} - 0.25v^z_{\max}, \\ &3.3J^x_{\text{pelvis}} - 2.3J^x_{\text{R\_ankle}}, 1.1J^y_{\text{pelvis}} - 0.1J^y_{\text{R\_ankle}}, 1.25v^z_{\max} - 0.25v^z_{\min}],\end{aligned} \tag{4}$$

for pants, and

$$\begin{aligned}&[x_{\min}, y_{\min}, z_{\min}, x_{\max}, y_{\max}, z_{\max}] = \\ &[3J^x_{\text{R\_ankle}} - 2J^x_{\text{pelvis}}, 1.1J^y_{\text{R\_ankle}} - 0.1J^y_{\text{pelvis}}, 1.25v^z_{\min} - 0.25v^z_{\max}, \\ &4J^x_{\text{pelvis}} - 3J^x_{\text{R\_ankle}}, 1.1J^y_{\text{pelvis}} - 0.1J^y_{\text{R\_ankle}}, 1.25v^z_{\max} - 0.25v^z_{\min}],\end{aligned} \tag{5}$$

for skirt.

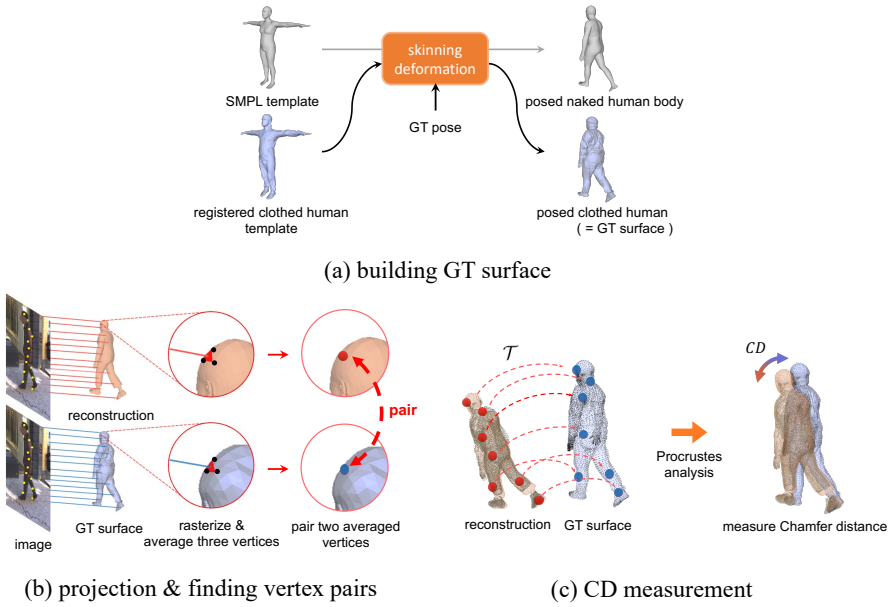


Fig. S3. Detailed illustration of the three steps of the Chamfer distance evaluation.

D Detail of Chamfer distance metric

This section details the Chamfer distance (CD) evaluation metric in three steps: building GT surface, projection & finding vertex pairs, and CD measurement.

Building GT surface. In 3DPW [11], used for evaluation in our work, T-posed clothed human meshes are registered to each subject's 3D scan. The registered 3D clothed human meshes have the same mesh topology as that of SMPL body mesh. We deform the registered 3D clothed humans with GT SMPL pose parameters following the same skinning deformation of SMPL. We use the posed clothed human as GT surface to measure CD with reconstruction result.

Projection & finding vertex pairs. Before measuring CD, we rigidly align global rotation, scale, and translation of the reconstruction to the GT surface. For the alignment, we find semantically matching vertex pairs between the reconstruction and the GT surface based on each 2D projection. First, we project both the reconstruction and GT surface into the input image and rasterize them. By the rasterization, we obtain face index maps of their meshes, where each pixel of a face index map represents a visible face index among mesh faces projected to that pixel location. Second, we average three vertices that make up each face for both the reconstruction and GT surface. Finally, we pair the averaged two vertices (*i.e.*, one from the reconstruction and the other from the GT surface) that correspond to the same pixel location. With these processes, we obtain semantic matching vertex pairs that project to the same pixel of the input image.

CD measurement. Based on the semantic matching vertex pairs, we construct a rigid transformation matrix $\mathcal{T}$ that transforms the first vertices to the second vertices of the pairs. Using the rigid transformation matrix $\mathcal{T}$, we align all of the reconstruction vertices to the GT surface. After alignment, we measure CD between the aligned reconstruction and the GT surface.

E Supervision on cloth existence

GT existence of each cloth is set to `True` if a GT segmentation of the cloth exists in the input image. As cloth segmentations are defined only inside of the image, naively setting the existence to `False` when the segmentation does not exist in the input image can be wrong. For example, if a human is wearing pants and the pants are not included in the input image due to the truncation, the naive solution sets the existence of the pants to `False`, which is wrong.

Instead, we use DensePose to distinguish 1) a human is not wearing a cloth and 2) a human is possibly wearing a cloth but not included in the input image. In the DensePose, human part patches are defined, such as head, torso, arm, and so on. We correspond the upper clothes and the coat into torso and arm parts, the pants and the skirt into leg parts, and the shoes into foot parts. The existence of each cloth is set to `False` if 1) the cloth's human part patches are included in the input image and 2) the cloth's segmentation does not exist in the input image. If the cloth's human part patches are not included in the input image, we do not supervise the predicted cloth existence score for the cloth.

F Evaluation on cloth existence and gender

We evaluate our ClothWild in cloth existence and gender on the MSCOCO [8] validation set. GT cloth existences are obtained cloth segmentation annotations [4] and DensePose in the same way as described in Section E. For gender, since there is no dataset containing GT gender annotations for MSCOCO, we acquire gender annotations by running Homogenus [14] and use its predictions as pseudo-GTs. Measuring accuracy with the obtained GTs, the accuracy of cloth existence and gender are 0.848 and 0.918, respectively.

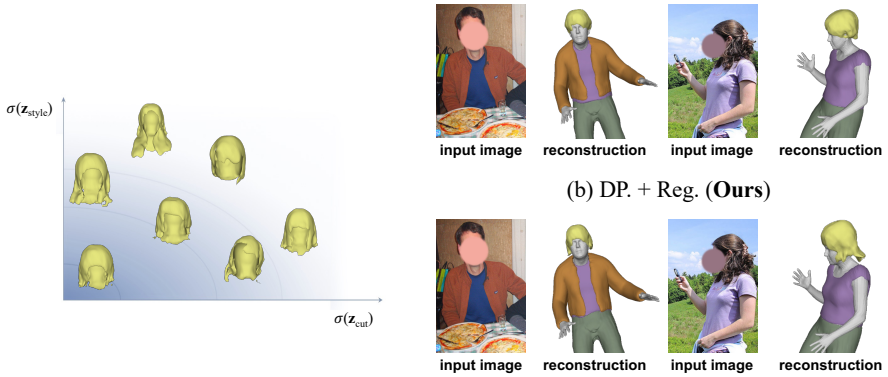
G Limitations

Hair reconstruction. Although the cloth generative model, SMPLicit [3], also supports 3D hair generation, there is a limitation to reconstructing hair in our framework for the following two reasons. First, the SMPLicit cannot cover a wide variety of hair because it is trained mostly from a small set of women's hair. Fig. S4(a) shows that SMPLicit's hair outputs are not diverse and mostly long hair, biased to women's long hair. The hairs of the figure are produced from randomly sampled two Gaussian codes (*i.e.*, $\mathbf{z}_{\text{style}}$ and $\mathbf{z}_{\text{cut}}$) with various standard deviations following the model design of the SMPLicit. Second, a large

part of the hair is quite distant from the human body surface. As our DensePose-based loss function supervises 3D query points close to the human body surface, a large part of the hair is not covered by the DensePose-based loss. Although we use silhouette loss to supervise 3D query points far from the surface, learning 3D hair is not done properly due to the depth ambiguity of 2D supervision targets. Figs. S4(b) and S4(c) show that the reconstructed hair with our framework is unnatural, especially for long hair. Due to these difficulties, learning 3D hair is one of the challenges to be solved.

Shoes reconstruction. Fig. S5(a) shows that our framework often inadequately reconstructs several shoes, such as high heels. We guess the reason is that such shoes have complex geometry, while they do not have many pixels in the cloth segmentations. Therefore, the shoe segmentations are not very informative for learning shoes with complex geometry.

Expression power. There is a limitation to represent cloth details (*e.g.*, wrinkles), as shown in Fig. S5(b). The reason is that the expression power of our framework depends on the cloth generative model. Most of current cloth generative models [2,6,10,1,13], including SMPLicit [3] of our framework, have difficulty in embedding delicate cloth geometry (*e.g.*, wrinkles) and in the cloth latent space. Therefore, we think improving the expression power of cloth generative models should be a future research direction.

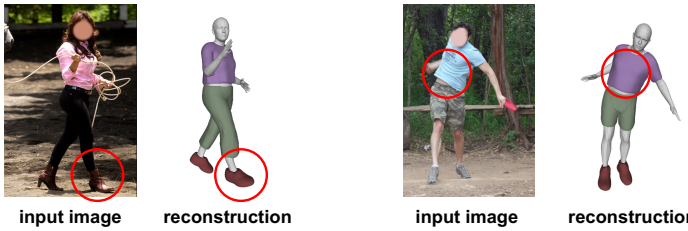


(a) hair distribution on the latent space

(b) DP. + Reg. (**Ours**)

(c) Ours + Sil. (**add for hair only**)

Fig. S4. (a) 3D hair distribution according to the standard deviations of the latent codes. (b) Hair reconstruction examples with our loss configuration. (c) Hair reconstruction examples when adding the silhouette loss for hair only.



(a) failure cases on shoe reconstruction (b) limitation on cloth expression

Fig. S5. Limitations of our framework: shoe reconstruction and expression power.

H More qualitative results

We provide more qualitative result comparisons on the MSCOCO [8] validation set. Fig. S6 shows that our ClothWild performs well on extremely challenging cases, such as overlaps between people (upper left and lower left), occlusions (upper right), and from a low-resolution image (lower right). Figs. S7 and S8 show our ClothWild produces far better results from in-the-wild images compared to previous state-of-the-art 3D clothed human reconstruction methods.

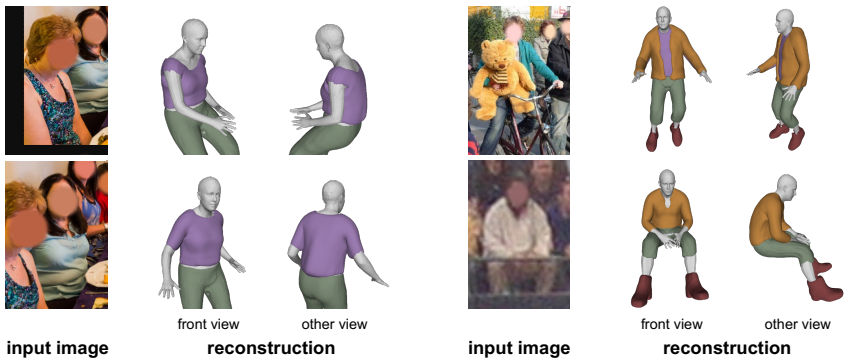


Fig. S6. Qualitative results on challenging cases of MSCOCO [8] validation set. Colors of reconstructed 3D clothes are manually assigned to represent cloth types.

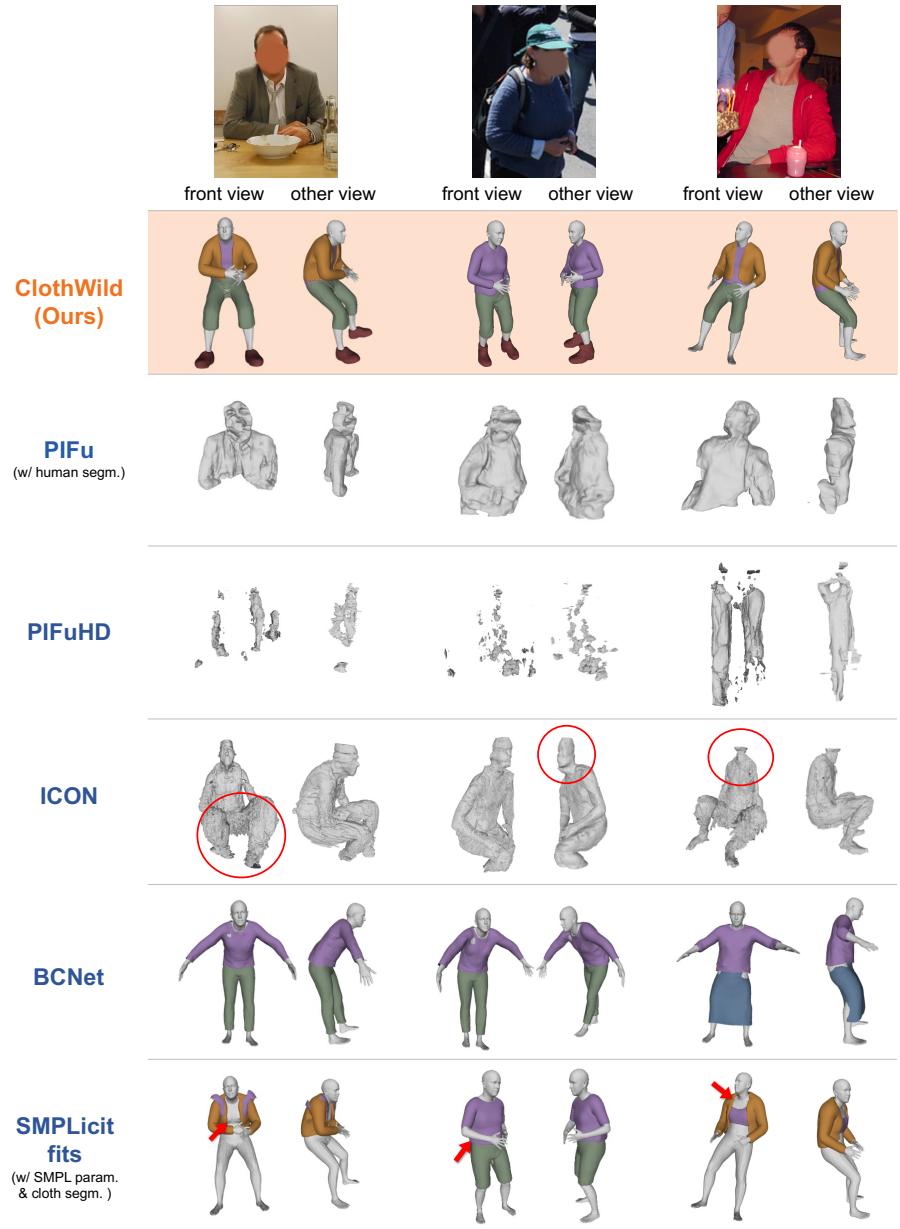


Fig. S7. Qualitative result comparisons on MSCOCO [8] validation set. PIFu additionally uses human segmentation obtained from Mask R-CNN [5] for reconstruction. SMPLicit fits use SMPL parameter and cloth segmentations obtained from Pose2Pose [12] and SCHP [7], respectively. Colors of reconstructed 3D clothes are manually assigned to represent cloth types.

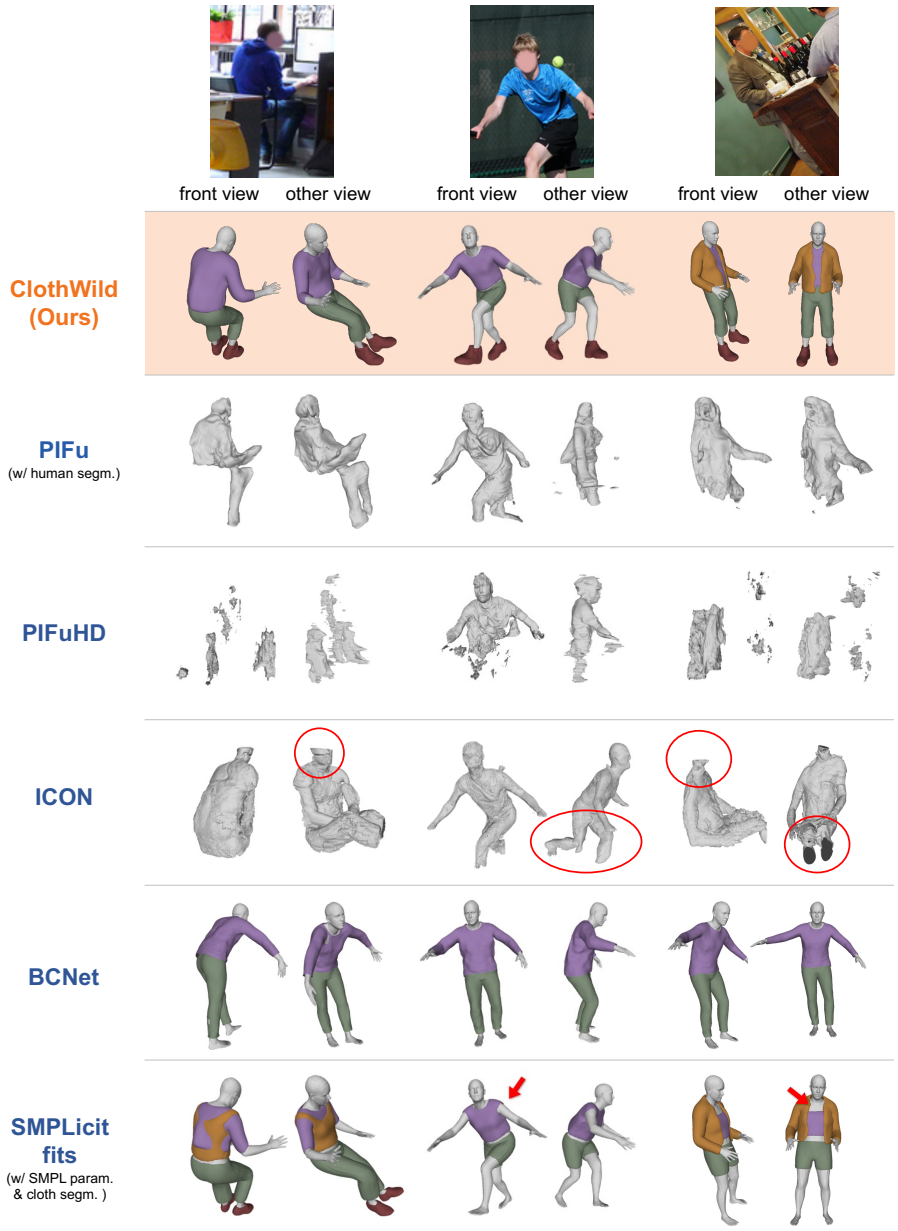


Fig. S8. Qualitative result comparisons on MSCOCO [8] validation set. PIFu additionally uses human segmentation obtained from Mask R-CNN [5] for reconstruction. SMPLicit fits use SMPL parameter and cloth segmentations obtained from Pose2Pose [12] and SCHP [7], respectively. Colors of reconstructed 3D clothes are manually assigned to represent cloth types.

References

1. Bertiche, H., Madadi, M., Escalera, S.: CLOTH3D: Clothed 3D humans. In: ECCV (2020)
2. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-Garment Net: Learning to dress 3D people from images. In: ICCV (2019)
3. Corona, E., Pumarola, A., Alenya, G., Pons-Moll, G., Moreno-Noguer, F.: SMPLicit: Topology-aware generative model for clothed people. In: CVPR (2021)
4. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: CVPR (2017)
5. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
6. Jiang, B., Zhang, J., Hong, Y., Luo, J., Liu, L., Bao, H.: Bcnet: Learning body and cloth shape from a single image. In: ECCV (2020)
7. Li, P., Xu, Y., Wei, Y., Yang, Y.: Self-correction for human parsing. IEEE TPAMI (2020)
8. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
9. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM TOG (2015)
10. Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M.J.: Learning to dress 3D people in generative clothing. In: CVPR (2020)
11. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3D human pose in the wild using imus and a moving camera. In: ECCV (2018)
12. Moon, G., Choi, H., Lee, K.M.: Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In: Computer Vision and Pattern Recognition Workshop (CVPRW) (2022)
13. Patel, C., Liao, Z., Pons-Moll, G.: TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style. In: CVPR (2020)
14. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: CVPR (2019)