

# Supplementary Material for CLOTH4D: A Dataset for Clothed Human Reconstruction



Figure A. Visual Comparisons of Clothed Human Reconstruction Datasets. (a) RenderPeople [1], (b) CAPE [16], and (c) THuman2.0 [27] are scanned datasets, and the rest are synthetic ones. (d) TailorNet [19], (e) ReSynth [17], and (f) cloth3d [4] are not photorealistic, owing to their mannequin-like human avatars. (g) 3DPeople [23], (h) Cloth3d++ [18], and (i) **CLOTH4D** are more photorealistic. Our **CLOTH4D** presents the highest photorealism.

## 1. Visual Comparisons of Datasets

This section mainly serves as the supplementary of Section 2 in the main paper, which visually shows the difference between **CLOTH4D** and previous clothed human reconstruction datasets. As shown in Figure A, the samples in the 1st row belong to the scanned data. The rest belong to the synthetic data. The samples in the 2nd row are noted as ‘not photorealistic’ in the main paper because their avatars are not human-like. On the contrary, the 3rd row exhibits

photorealistic synthetic data. We can see that **CLOTH4D** is more realistic and has more diverse and elaborate clothing than 3DPeople [23] and cloth3d++ [18]. 3DPeople scans are not publicly available, and clothes in cloth3d++ are created using pre-defined garment templates, which are inherently not complicated enough to cover the rich diversity of real-life clothing.

On the contrary, the clothes in **CLOTH4D** are manually made by fashion designers in professional virtual fashion software. A large part of apparel is referred from the newest

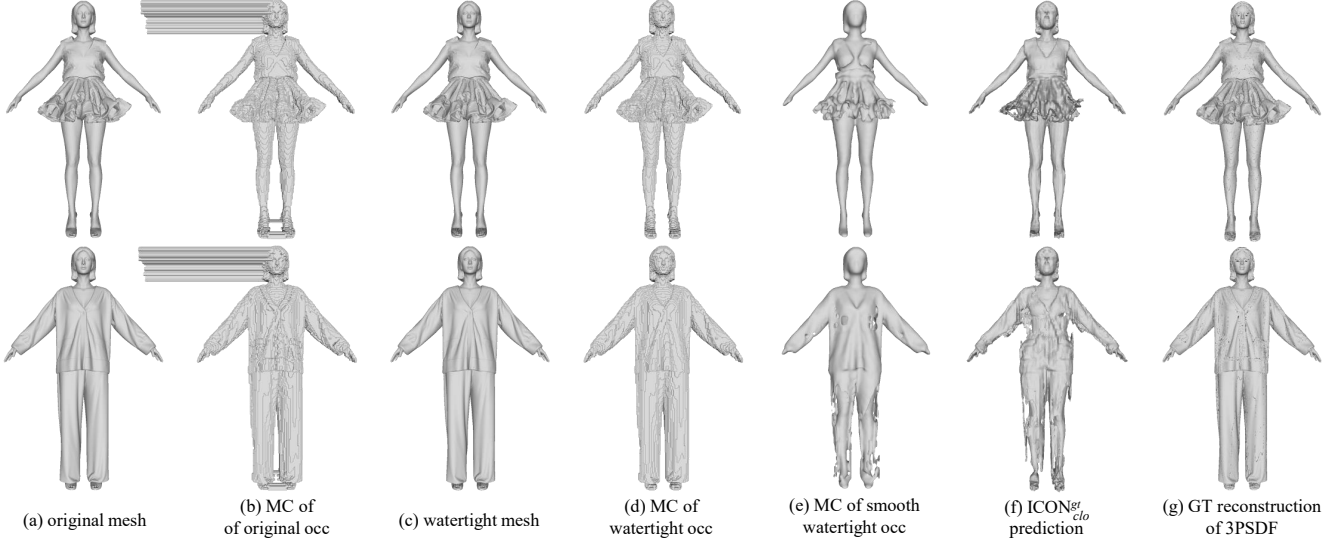


Figure B. To demonstrate the difficulties of reconstructing multi-layer thin structures with occupancy field (the queried occupancy field has a resolution of  $512 \times 512 \times 512$  if not otherwise specified), we show (a) the original non-watertight multi-layer mesh; (b) the marching cubes (MC) [15] results of the original mesh; (c) the watertight mesh converted using [10]; (d) the marching cubes results of the watertight mesh (c); (e) the marching cubes results of the smoothed watertight mesh in (c); (f) the reconstruction results of  $\text{ICON}_{clo}^{gt}$ ; and (g) reconstruction by the state-of-the-art implicit representation 3PSDF [6] given the ground truth mesh.

collections of various fashion houses, including Prada, Moschino, Alexander McQueen, Givenchy, JACQUEMUS, Mugler, Salvatore Ferragamo, Dion Lee, etc. We find that the instances in CLOTH4D are of the highest quality and photorealism compared with other synthetic datasets.

## 2. Mesh Representation.

This section is mainly to serve as the supplementary of Section 4.4 in the main paper, which aims to demonstrate the difficulties of mesh reconstruction by predicting an occupancy field for a multi-layer thin structure. As shown in Figure B, we encounter the following challenges:

1) **Watertight assumption.** The original mesh generated by CLO is not watertight (Figure B(a)), thus preventing us from defining an occupancy field in the 3D space. One may cast a ray from a 3D querying point and calculate the number of intersections between the ray and the mesh, then uses the parity of the number of intersections to define the inside/outside of the point w.r.t. the mesh. Finally, the marching cubes algorithm can be applied to reconstruct the mesh surface. However, Figure B(b) shows that such approximation generates meshes with noticeable artifacts near open surfaces like heads and shoes.

2) **Predicting occupancy field for multi-layer thin structures.** To leverage the occupancy field for surface reconstruction, we convert the original meshes to watertight (Figure B(c)) using ManifoldPlus [10]. And the corresponding reconstructed meshes with ground truth  $512 \times 512 \times 512$  queried occupancy field is shown in Figure B(d).

We note that even with the ground truth occupancy, the reconstructed meshes lose high-frequency details and thin structures due to the limited resolution of the occupancy field, indicating it still remains challenging for the field-to-mesh conversion given uniformly sampled discrete querying points in the 3D space.

Moreover, during inference, the occupancy field is predicted by a network instead of using the ground truth. It is known that a deep model tends to give smooth predictions (e.g., contiguous prediction near the mesh surface). We mimic this behavior by smoothing the ground truth occupancy field with the default setting suggested in [3] (smoothing is also a common pre-processing to avoid the marching cubes results being jagged [3]) and illustrate the reconstruction in Figure B(e). These results present holes and missing clothing pieces similar to the prediction by  $\text{ICON}_{clo}^{gt}$  (Figure B(f)). This implies that the current state-of-the-art occupancy-based implicit functions [24, 25, 30] suffer from preserving complicated clothing structures.

3) **SOTA surface reconstruction method.** To further investigate the challenges brought by our dataset, we reconstruct the mesh surfaces using the SOTA implicit representation 3PSDF [6] given the ground truth 3PSDF values, and the results are shown in Figure B(g). Although targeting to tackle non-watertight arbitrary topologies and complex geometries, 3PSDF is still not robust for the multi-layer thin structures in CLOTH4D, introducing holes, face intersections, and uneven surfaces. Additionally, the training data generation process of 3PSDF [6] is time-consuming for the meshes containing fine details in our dataset. It takes around



Figure C. Cases that will result in simulation artifacts. The animations of these three sequences are Arms Akimbo, Turn Left, and Walk.

50min to generate the training label (*i.e.*, 3PSDF representation) for a mesh on a single CPU core, making it impractical to scale on a large dataset like CLOTH4D.

### 3. Failure Cases of Cloth Simulation

This section is mainly to serve as the supplementary to Section 4.5 in the main paper, which aims to elaborate on the limitations of CLOTH4D due to failures of clothing simulation.

As shown in Figure C, it can be seen that three types of conditions will result in failure clothing simulation: 1) Soft fabric is more prone to unstable simulation. As shown in the 1st row, the light tulle dress looks slightly unnatural at the bottom. 2) Physical contact may bring abnormal adhesion between skin and clothes. It usually occurs when the limbs are close to the clothes or touch them and then move away, as the 2nd row shows in Figure C. 3) Multi-layers cause difficulties in clothing simulation. This has minor effects compared with the previous two since we could delete the inner layers for simulation in most situations. Inherently, the reason for these failure cases is related to the algorithm of cloth simulation. In other words, the quality of synthetic data will highly rely on the performance of the simulation algorithm.

In addition, the inaccurate rigging of Maximo will result in unnatural simulations to a certain extent.

### 4. Detailed Temporal Metrics

The  $\text{Chamfer}_{ddt}$  and  $\text{Chamfer}_{dtd}$  are defined as follows:

$$\text{Chamfer}_{ddt} = \frac{1}{T} \sum_t |d_{CD}(\mathcal{M}_t^{pr}, \mathcal{M}_t^{gt}) - d_{CD}(\mathcal{M}_{t+1}^{pr} - \mathcal{M}_{t+1}^{gt})|, \quad (1)$$

$$\text{Chamfer}_{dtd} = \frac{1}{T} \sum_t |d_{CD}(\mathcal{M}_t^{pr}, \mathcal{M}_{t+1}^{pr}) - d_{CD}(\mathcal{M}_t^{gt} - \mathcal{M}_{t+1}^{gt})|, \quad (2)$$

where  $M_t^{pr}$  and  $M_t^{gt}$  denote the predicted and ground truth mesh at time step  $t$ , respectively.  $d_{CD}(M_1, M_2)$  is the Chamfer distance between two meshes that is formulated as:

$$d_{CD}(\mathcal{M}_1, \mathcal{M}_2) = \frac{1}{\mathcal{M}_1} \sum_{x \in \mathcal{M}_1} \min_{y \in \mathcal{M}_2} \|x - y\|_2^2 + \frac{1}{\mathcal{M}_2} \sum_{x \in \mathcal{M}_2} \min_{y \in \mathcal{M}_1} \|x - y\|_2^2. \quad (3)$$

The  $P2S_{ddt}$  and  $P2S_{dtd}$  are of a similar form by replacing Chamfer distance with point-to-surface distance. For a more comprehensive analysis and discussions on how  $ddt$  and  $dtd$  correlate with human perception, please refer to [7].

## 5. Fitting SMPL Model to Body Mesh

A merit of our dataset is that it provides under-clothing body mesh as shown in Figure D, making it simpler and more accurate to register a SMPL [14] body model to the body mesh. We follow a similar registration pipeline as [20]. More specifically, given a naked body scan  $\mathcal{S}$ , we fit a parametric SMPL body model  $\mathcal{M}(\beta, \theta)$ , where  $\beta$  and  $\theta$  represent the body and body parameters, respectively, such that  $\mathcal{S}$  and  $\mathcal{M}$  are as close as possible.

We first render the textured mesh  $\mathcal{M}$  into multi-view images and use OpenPose [5] to extract 2D keypoints. Then, we initialize the SMPL parameters by running a multi-view SMPLify approach [2, 21], which can roughly align the joints of the human body mesh and the SMPL mesh. To obtain a refined registration, we minimize the following loss function:

$$\begin{aligned} E(\beta_1, \beta_2, \dots, \beta_T, \theta_1, \theta_2, \dots, \theta_T) = & \sum_{t=1}^T (\lambda_{joint} E_{joint,t} + \lambda_{shape} E_{shape,t} \\ & + \lambda_{chamfer} E_{chamfer,t} + E_{reg,t}), \end{aligned} \quad (4)$$

$$E_{shape,t} = |\beta_t - \beta_{t+1}|,$$

$$E_{chamfer,t} = d_{CD}(\mathcal{M}(\beta_t, \theta_t), \mathcal{S}_t),$$

$$E_{reg,t} = \lambda_\beta E_\beta(\beta_t) + \lambda_\theta E_\theta(\theta_t),$$

where we jointly optimize the shape and pose parameters  $\beta_1, \beta_2, \dots, \beta_T, \theta_1, \theta_2, \dots, \theta_T$  for the whole motion sequence of length  $T$ .  $E_{joint,t}$  is the 2D joint reprojection loss minimizing the distance of OpenPose 2D keypoints and SMPL projected keypoints. As a motion sequence is of the

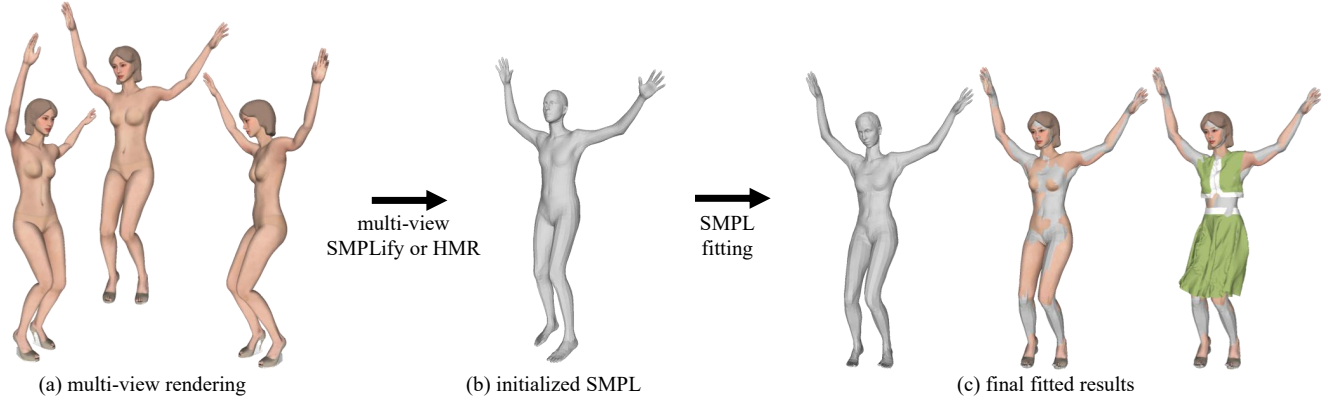


Figure D. Pipeline of registering an SMPL to the avatar body mesh.

Table 1. Quantitative evaluation on CLOTH4D<sub>tight</sub>.

Method	PIFu	PIFu <sub>clo</sub>	PaMIR	PaMIR <sub>clo</sub>	ICON	ICON <sub>clo</sub>
Normals ↓	0.137	0.104	0.173	0.186	0.149	0.147
P2S ↓	3.037	2.167	3.499	3.800	3.334	2.734
Chamfer ↓	3.015	1.795	3.524	3.416	3.116	2.637

same person, we use  $E_{shape,t}$  to constrain the body shape consistency across time.  $E_{chamfer,t}$  is the Chamfer distance (Eq. (3)) between the fitted SMPL mesh  $\mathcal{M}(\beta_t, \theta_t)$  and human mesh  $S_t$ , which enforces the fitted mesh ensemble the original human mesh.  $E_{reg,t}$  regularizes the body parameters  $\beta_t$  and  $\theta_t$  as in [21].  $\lambda_*$  balance different loss terms and are empirically set to make each energy term have the value of a similar scale.

As SMPL mesh can only capture a minimal-clothing body, while the existing scan datasets usually contain clothed people. They have to leverage more off-the-shelf body information (e.g., human parsing), complicated fitting terms (e.g., clothing terms, skin terms), and hyper-parameter tuning to reach satisfactory results [20, 28]. In contrast, CLOTH4D obtains more accurate SMPL fits in a simple and effective way by minimizing the Chamfer distance. We also notice that the scan fitting is pretty robust to the initialized SMPL parameters, thus in practice, the initialization could also be done by running multi-view SMPLify on clothed human images or with a monocular human mesh recovery method like [12, 29].

## 6. More Results

**Results on Tight Clothes** As most SOTAs are designed for datasets with tight clothes, to show how the results vary for different clothing tightness, we made quantitative comparisons on a subset of CLOTH4D with tighter (similar to the scanned dataset) yet still high-quality clothes, i.e., removing the challenging clothes. The results evaluated on CLOTH4D<sub>tight</sub> (~15% of the original test set) are reported in Table 1. The conclusions in Section 4.3 and Section 4.4 of the main paper still hold.

**Results of Multi-layer Algorithms** We further compare

SOTA methods trained on CLOTH4D with other multi-layer methods like BCNet [11]. The results are shown in Figure E(a). These template-based methods avoid generating broken results but fail to model unseen clothing categories (e.g., shoes, dresses) and details, leading to large reconstruction errors.

**Results on Multi-layer Test Sets** In addition to evaluating on scan dataset CAPE in the main paper, Figure E(b) shows some qualitative results on a multi-layer synthetic dataset. It reflects the same conclusions although Cloth3D++ only contains simple clothes and unrealistic images. Training on CLOTH4D achieves better results on Cloth3D++ (Chamfer of PIFu vs. PIFu<sub>clo</sub> is 4.238 vs. 3.310).

## 7. Real-world Usefulness

As illustrated in the experiments, the re-trained SOTA methods have weak generalization ability to real-world images, enlarging the possibility of generating broken clothing. After the reality check, we attribute the worse results of the in-the-wild images with loose clothing to 1) SOTA methods using implicit functions tend to generate broken results (even on the CLOTH4D test set as in Figure 5 and Figure B) when trained on data with thin layers; 2) When only trained on CLOTH4D, these methods overfit and amplify these artifacts even if the domain gap is small. Thus, we expect future research to explore the full potential of CLOTH4D following the suggestions in the paper.

Our paper, as a reality check to SOTA, concludes that more effective solutions are required to capture multi-layer structures for real-world application potential. As a dataset, CLOTH4D itself 1) provides layering concepts for learning multi-layer structures; 2) implies various fashion attributes, making the reconstruction aware of design details; and 3) contains better dynamic and geometric information for generating more accurate results, especially for the side view. Consistent conclusions were made on all tested sets, i.e., CAPE, Cloth3D++ (Figure E (b)), and in-the-wild images (Figure E (c)).



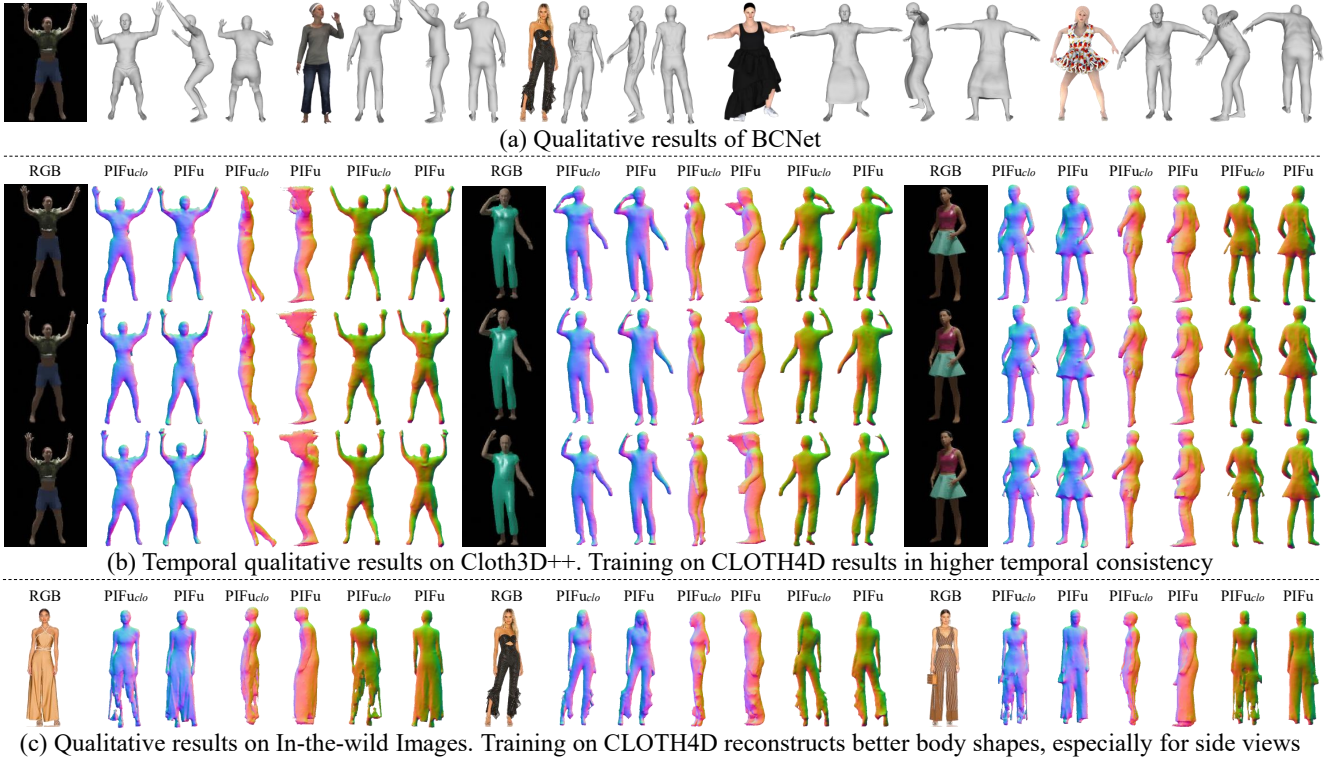


Figure E. More qualitative comparisons. We show input images and the reconstructed front/side/back views.

## 8. Enriching Other Tasks

As mentioned in Section 3 in the main paper, CLOTH4D contains high-quality paired data across multiple modalities. We believe that in addition to clothed human reconstruction, CLOTH4D could facilitate a wider range of computer vision and graphics tasks, *e.g.*, clothing capture [22, 26], human pose transfer [8, 13], video 2D/3D pose estimation [12], fashion attribute classification [31], virtual try-on [9], and other fashion related tasks [32, 33].

## 9. Video Reference

As mentioned in the main paper, we provide two types of videos for a more intuitive demonstration.

**Video A** (Section 3 in the main paper) shows the creation process of a 3D clothing item, ‘Polyfaille Exploded Corset Dress in Bobby Pink’ (the reference is borrowed from Alexander Macqueen<sup>1</sup>). To show the advance of CLOTH4D, we further put more examples in Figure F. It can be seen that the clothes contained in the CLOTH4D are rich in textures, fabrics, silhouettes, prints, and types.

**Video B** (Section 4.3 in the main paper) consists of seven separate sequences exhibiting the qualitative results of all reported models in the main paper. The observations and

conclusions are consistent with that described in the main paper.

## References

- [1] 3dpeople. <https://www.3dpeople.com>. 1
- [2] A multi-view smpl fitting based on smplify-x. <https://github.com/boycehbz/MvSMPLfitting>. 3
- [3] Pymcubes: Marching cubes (and related tools) for python. <https://github.com/pmneila/PyMCubes>. 2
- [4] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: clothed 3d humans. In *European Conference on Computer Vision*, pages 344–359. Springer, 2020. 1
- [5] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3
- [6] Weikai Chen, Cheng Lin, Weiyang Li, and Bo Yang. 3psdf: Three-pole signed distance function for learning surfaces with arbitrary topologies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18522–18531, 2022. 2
- [7] Mikhail Erofeev, Yury Gitman, Dmitriy S Vatin, Alexey Fedorov, and Jue Wang. Perceptually motivated benchmark for video matting. In *BMVC*, pages 99–1, 2015. 3
- [8] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF international*

<sup>1</sup><https://www.alexandermacqueen.com/en-us/ready-to-wear/polyfaille-exploded-corset-dress-699946QEACM5084.html>



Figure F. More samples in CLOTH4D.

conference on computer vision, pages 10471–10480, 2019. 5

- [9] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 5
- [10] Jingwei Huang, Yichao Zhou, and Leonidas Guibas. Manifoldplus: A robust and scalable watertight manifold surface generation method for triangle soups. *arXiv preprint arXiv:2005.11621*, 2020. 2
- [11] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision*, pages 18–35. Springer, 2020. 4
- [12] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 4, 5
- [13] Hongyu Liu, Xintong Han, ChengBin Jin, Huawei Wei, Zhe Lin, Faqiang Wang, Haoye Dong, Yibing Song, Jia Xu, and Qifeng Chen. Human motionformer: Transferring human motions with vision transformers. *arXiv preprint arXiv:2302.11306*, 2023. 5
- [14] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-

person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 3

- [15] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 2
- [16] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020. 1
- [17] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J Black. The power of points for modeling humans in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10974–10984, 2021. 1
- [18] Meysam Madadi, Hugo Bertiche, Wafa Bouzouita, Isabelle Guyon, and Sergio Escalera. Learning cloth dynamics: 3d + texture garment reconstruction benchmark. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track, PMLR*, volume 133, pages 57–76, 2021. 1
- [19] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 1
- [20] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021. 3, 4
- [21] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 3, 4
- [22] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (ToG)*, 36(4):1–15, 2017. 5
- [23] Albert Pumarola, Jordi Sanchez, G. Choi, A. Sanfeliu, and F. Moreno-Noguer. 3dpeople: Modeling the geometry of dressed humans. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2019. 1
- [24] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2314, 2019. 2
- [25] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 2
- [26] Shan Yang, Tanya Ambert, Zherong Pan, Ke Wang, Licheng Yu, Tamara Berg, and Ming C Lin. Detailed garment recovery from a single-view image. *arXiv preprint arXiv:1608.01250*, 2016. 5

- [27] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, June 2021. 1
- [28] Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4191–4200, 2017. 4
- [29] Hongwen Zhang, Yating Tian, Xinchu Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 4
- [30] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3170–3184, 2021. 2
- [31] Xingxing Zou, Xiangheng Kong, Waikeng Wong, Congde Wang, Yuguang Liu, and Yang Cao. Fashionai: A hierarchical dataset for fashion understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 5
- [32] Xingxing Zou and Waikeng Wong. fashion after fashion: A report of ai in fashion. *arXiv preprint arXiv:2105.03050*, 2021. 5
- [33] Xingxing Zou, Wai Keung Wong, and Dongmei Mo. Fashion meets ai technology. In *Artificial Intelligence on Fashion and Textiles: Proceedings of the Artificial Intelligence on Fashion and Textiles (AIFT) Conference 2018, Hong Kong, July 3–6, 2018*, pages 255–267. Springer, 2019. 5