CloSET: Modeling Clothed Humans on Continuous Surface with Explicit Template Decomposition **Supplementary Material**

Hongwen Zhang[1] Siyou Lin[1] Ruizhi Shao[1] Yuxiang Zhang[1] Zerong Zheng[1]
Han Huang[2] Yandong Guo[2] Yebin Liu[1]
[1]Tsinghua University [2]OPPO Research Institute

This Supplementary Material provides additional details about our approach and more experimental results that were not included in the main manuscript due to limited space. In Section A, we present more descriptions of our newly introduced THuman-CloSET dataset. In Section B, we provide more details about the implementation of our approach. Finally, we report more experimental results in Section C. More results are also presented in the Supplementary Video and the project page at https://www.liuyebin.com/closet.

Table A1. Comparison of the scan data used in our experiments.

Datasets	# Outfits	Outfit type	Average # poses per outfit
CAPE [4]	14	real-world, common	1806
ReSynth [5]	12	synthetic, loose	984
THuman-CloSET	15	real-world, loose	140

A. THuman-CloSET Dataset

We introduce THuman-CloSET for the reason that existing pose-dependent clothing datasets [4, 5] are with either relatively tight clothing or synthetic clothing via physics simulation. THuman-CloSET contains more than 2,000 high-quality human scans captured by a dense camera rig. There are 15 different outfits with a large variation in clothing style, including T-shirts, pants, skirts, dresses, jackets, and coats, to name a few. All subjects are guided to perform different poses by imitating the poses in CAPE [4]. For each outfit, there is also a scan of the same subject in minimal clothing so that we can obtain a more accurate body shape. In our dataset, the body model is firstly estimated from the rendered multiview images of the clothed human and further refined with the ICP optimization between the body model and the scan. As shown in Fig. A1, the loose clothing makes the fitting of the underlying body models quite challenging. For more accurate fitting of the body models, we first fit a SMPL-X [7] model on the scan of the subject in minimal clothing and then adopt its shape parameters for the fitting of the outfit scans in different poses. In this way, we ensure that the fitted SMPL-X models of our dataset are overall of good quality. Fig. A4 shows several outfit scans and example scans in various poses of THuman-CloSET. The comparison with CAPE [4], ReSynth [5], and our THuman-CloSET datasets are summarized in Tab. A1. We make THuman-CloSET publicly available for research purposes and hope it can open a promising direction for clothed human modeling and animation from real-world scans.

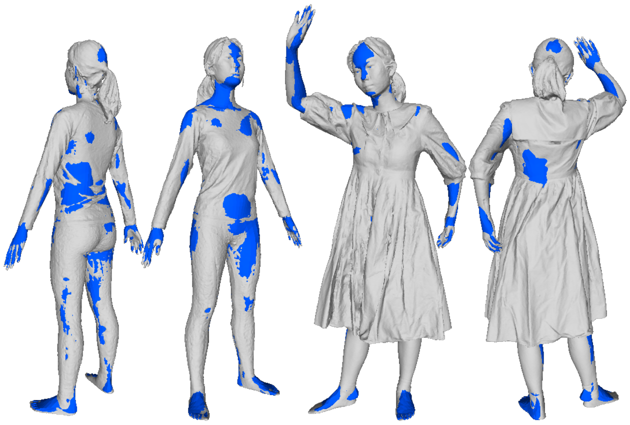

Figure A1. The fitted SMPL-X models (colored with blue) of the same subject in minimal and loose clothing.

B. More Implementation Details

Training. Following POP [5], we train our network for 400 epochs on ReSynth [5] and CAPE [4] datasets, using the Adam [2] optimizer with a batch size of 4 and a learning rate of 3.0×10^{-4}. The loss weights are set to $\lambda_p = 2 \times 10^4$, $\lambda_n = 0.1$, $\lambda_{rgl} = 2 \times 10^3$, $\lambda_{pd} = 1.0$, and $\lambda_{gc} = 5 \times 10^{-4}$ to balance different loss terms. Note that the normal loss is turned on from the 250th epoch for more stable training, as suggested in [5].

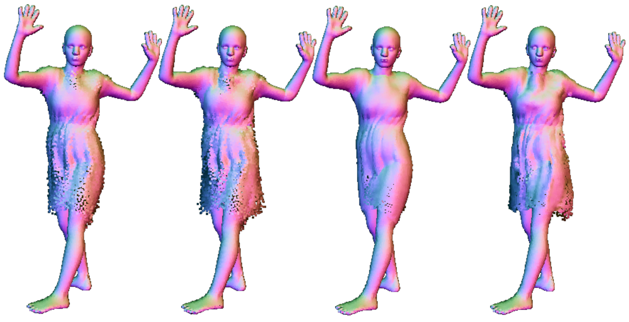
(a) (b) (c) (d)

Figure A2. Ablation results on the usage of garment features in the pose decoder. (a)(b) The temple and clothing deformation results without using garment features. (c)(d) The temple and clothing deformation results with the usage of garment features.

Architecture. In the implementation of our network, the PointNet++ [9] abstracts the point features for $L = 6$ levels, and the numbers of the abstracted points are $2048, 1024, 512, 256, 128,$ and 64, respectively at each level. The pose-dependent and garment-related features have the same length of 64, *i.e.*, $C_p = C_g = 64$. The decoders $\mathcal{D}_g$ and $\mathcal{D}_p$ adopt the same architecture as POP [5]. Tab. A2 reports the network parameters and runtime of POP [5] and our method. Note that the pose and garment encoders in our method can also be replaced with recent state-of-the-art point-based encoders such as PointMLP [6] and PointNeXt [10].

Table A2. Comparison of the network parameters and runtime.

Method	Encoder	# Params	Runtime
POP [5]	UNet [11]	11.33 M	44 ms
Ours	PointNet++ [9]	11.76 M	47 ms

Garment Code. Following POP [5], for a specific outfit (*e.g.*, an individual garment), the garment code is randomly initialized with the shape of $N \times 64$ (N is the vertex number of SMPL(-X)) and shared across all poses. During training, the code is optimized with the back-propagated gradients. When trained with multiple outfits, the pose-dependent deformation should be aware of the outfit type. Hence, the pose decoder takes as input both the garment features $\phi_g(\boldsymbol{p}_i^t)$ and the pose features $\phi_p(\boldsymbol{p}_i^t)$. As shown in Fig. A2, the qualitative results become worse when the garment features are not fed into the pose decoder under the multi-outfit setting.

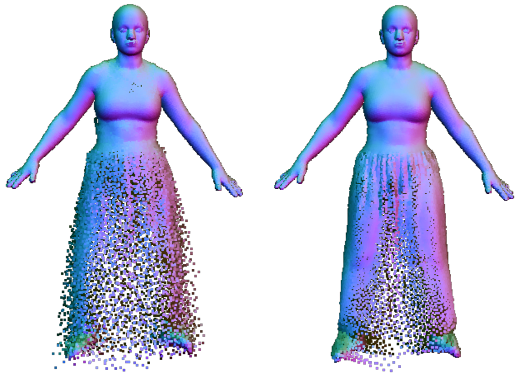
(a) using data term (b) using regularization term

Figure A3. The templates learned with (a) the data term and (b) the regularization term.

Table A3. Ablation study of the efficacy of the explicit template decomposition on different backbones. † denotes the default PointNet [8] and PointNet++ [9].

	Backbone	Size(M)	w/o ETD	w. ETD
UV	Unet	11.33	7.34	7.05
Surface	PointNet †	7.68	7.14	6.94
	PointNet++ †	4.35	7.08	6.71
	PointNet++	11.76	6.53	6.01

C. More Experimental Results.

Template learning. As described in Section 3 in the main paper, the explicit templates are learned under the regularization term. An alternative strategy for template learning is applying the data term directly to the generated point clouds of templates, as done in previous work [3]. However, we found such a strategy leads to worse template learning. As visualized in Fig. A3, the template directly learned with the data term is much nosier than the one learned with the regularization term.

Effect of Explicit Template Decomposition. In Table A3, we further augment the ablation experiments with the backbones of the default PointNet [8] and PointNet++ [9]. We can see that i) learning continuous surface features consistently brings improvements over the UV features, though the default PointNet and PointNet++ have smaller model sizes; ii) PointNet++ is more suitable for surface feature learning than PointNet; iii) ETD consistently improves the results for all backbones. In Fig. A5, we include more rendered results of the clothing deformations learned with and without Explicit Template Decomposition (ETD). In general cases, ETD helps to capture more natural pose-dependent wrinkles. For more qualitative comparisons of SCANimate [12], SNARF [1], POP [5], and our approach, please refer to the supplementary video.

References

[1] Xu Chen, Yufeng Zheng, Michael J. Black, Otmar Hilliges, and Andreas Geiger. SNARF: Differentiable forward skin-

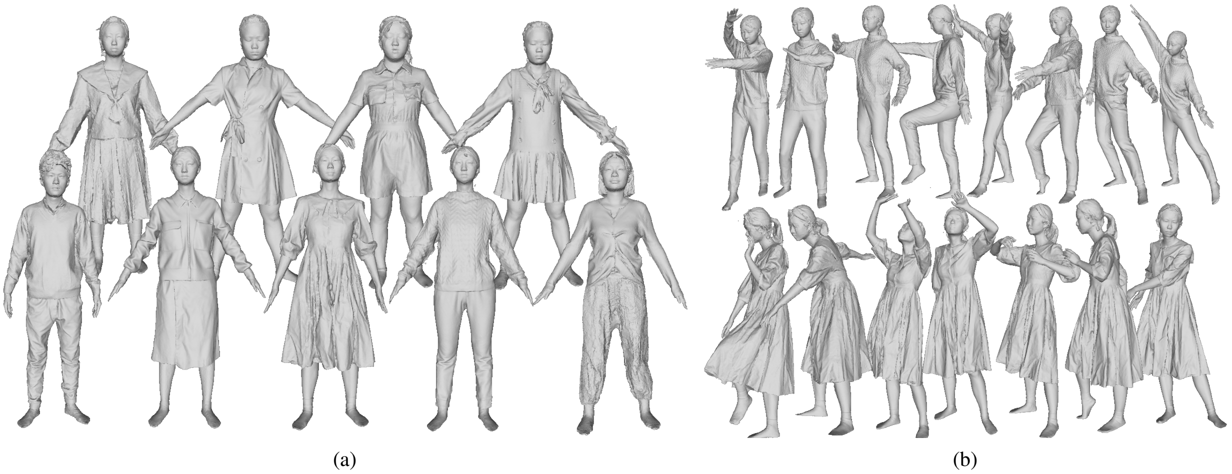

Figure A4. Example scans of our newly introduced THuman-CloSET dataset. (a) Example outfits. (b) Example scans in various poses.

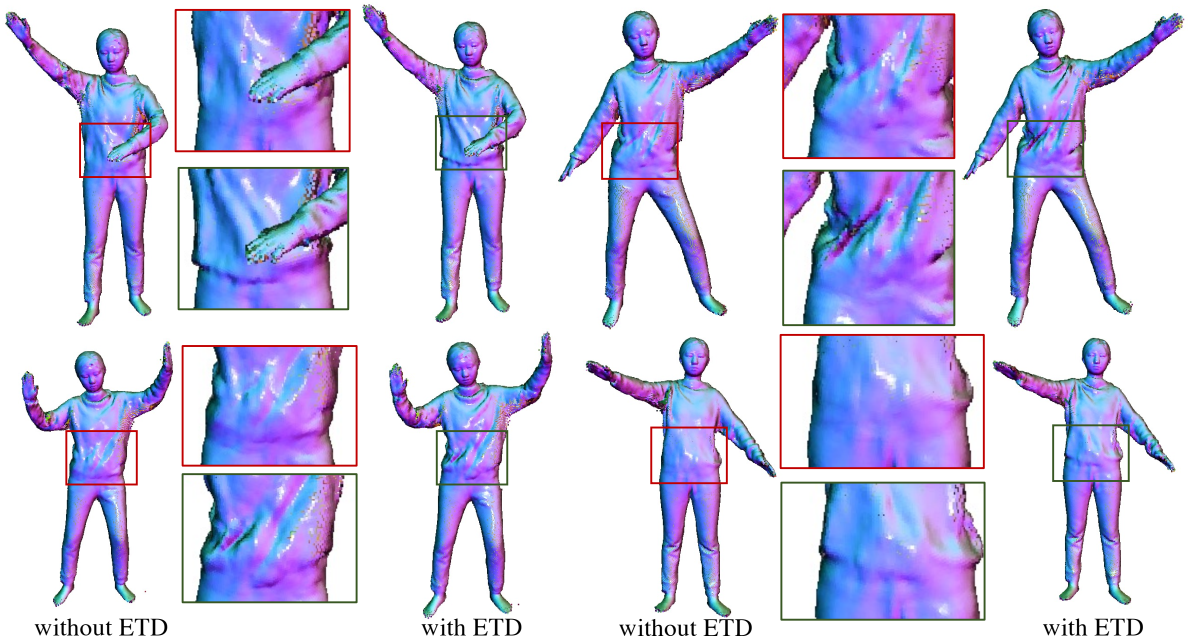

Figure A5. Comparison of the pose-dependent deformations learned with and without Explicit Template Decomposition (ETD).

ning for animating non-rigid neural implicit shapes. In *ICCV*, pages 11594–11604, October 2021. 2

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2014. 1

[3] Qianli Ma, Jinlong Yang, Michael J Black, and Siyu Tang. Neural point-based shape modeling of humans in challenging clothing. *3DV*, 2022. 2

[4] 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 *CVPR*, pages 6469–6478, 2020. 1

[5] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J Black. The power of points for modeling humans in clothing. In *ICCV*, pages 10974–10984, 2021. 1, 2

[6] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. In *ICLR*, 2022. 2

[7] 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 *CVPR*, pages 10975–10985, 2019. 1

[8] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, pages 652–660, 2017. 2

[9] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J

Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 30, 2017. 2

[10] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *NeurIPS*, 35:23192–23204, 2022. 2

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 2

[12] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *CVPR*, pages 2886–2897, 2021. 2