Supplementary Document for “Learning Semantic-Aware Disentangled Representation for Flexible 3D Human Body Editing”

In this document, we provide the following supplementary content:

- Implementation Details.
- Definition of Our Body Parts and Joints.
- More Experiment Details and Results.
- User Study.
- Limitations and Failure Cases.

1. Implementation Details

Hyper-Parameters. For the spiral convolution encoder and decoder, we follow the hyper-parameters of the spiral convolution (*e.g.*, filter size and dilation ratio) in [2]. For DFAUST [1] and SPRING [10] datasets, the sampling factor lists are $[2, 2, 2, 2]$ and $[4, 2, 2, 2]$, respectively. The hyper-parameters $\lambda_{edge}, \lambda_{dis\_shape}, \lambda_{edit\_shape}$ and λ_{norm} are set as 1×10^{-2}, $(\alpha_{min}, \alpha_{max})$ is set as $(0.8, 1.2)$, and σ is set as 72 degrees. Specifically, we do not use volume loss when training on the SPRING dataset, because there are almost no pose changes in the meshes.

Relative Error. We calculate $\lambda_{dis\_shape}, \lambda_{edit\_shape}$ and λ_{norm} in a relative sense for better reconstruction of local details following [4]. In particular, for the ground-truth T and predicted values P, we compute the relative error $\|(T - P)/T\|_1$ instead of $\|T - P\|_1$, which can improve the quality of the editing results.

2. Definition of Our Body Parts and Joints

We define body parts and their joints based on SMPL [8], and Fig. 1 shows the differences between SMPL and ours. Tab. 1 gives specific correspondences. Specifically, we merge the labels of some parts and remove some redundant joints to simplify the structure of the human body. Besides, we define additional joints for feet, hands, and faces to better represent their pose.

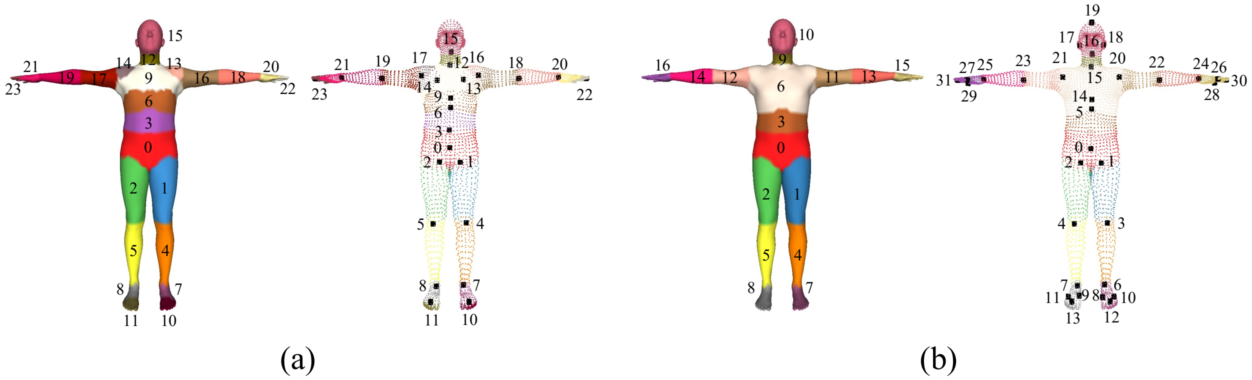
Part	i_{part}^{ours}	i_{part}^{smpl}	i_{joint}^{ours}	i_{joint}^{smpl}
head	10	15	16,17,18,19	15,-,-,-
neck	9	12	15,16	12,15
chest	6	6,9,13,14	14,15	9,12
waist	3	3	0,5	0,6
hip	0	0	0,1,2	0,1,2
left thigh	1	1	1,3	1,4
left shank	4	4	3,6	4,7
left feet	7	7,10	6,8,10,12	7,-,-,10
right thigh	2	2	2,4	2,5
right shank	5	5	4,7	5,8
right feet	8	8,11	7,9,11,13	8,-,-,11
left upperarm	11	16	20,22	16,18
left forearm	13	18	22,24	18,20
left hand	15	20,22	24,26,28,30	20,-,-,22
right upperarm	12	17	21,23	17,19
right forearm	14	19	23,25	19,21
right hand	16	21,23	25,27,29,31	21,-,-,23

Table 1. Correspondences between parts and joints of SMPL [8] and our parts and joints. i_{part} and i_{joint} denote the indices of parts and joints, respectively. -: no corresponding joints.

3. More Experiment Details and Results

Reconstruction Experiments. we use the official implementation of the compared methods [2, 3, 5, 6, 9, 12] with the same sampling factor list, latent space dimension, training strategy and reconstruction loss for a fair comparison. Since DHNN [7] only releases the decoder code, we compare the reconstructed human bodies on their dataset [7] by optimizing our hidden variables like them. Please refer to DHNN [7] for more optimization details. Since the author lost the test list, we randomly split the DHNN dataset into a test set of 320 meshes and a training set of 5274 meshes following its setting. It is worth noting that this is extremely unfair to our approach because most of the test meshes exist in the training set of DHNN. Fig. 2 shows some reconstruction results and error maps on the DHNN dataset [7].

Editing Experiments. For editing bone length and part shape size, we uniformly sample a scalar α in $(0.8, 1.2)$ for



(a)

(b)

Figure 1. The definition of body parts and joints: **(a)** SMPL [8] and **(b)** ours.

Method	E_{joint}	E_{circ}
Ours	**3.75**	15.89
DHNN	41.40	**10.17**
Ours	**11.33**	**20.35**
LIMP	33.90	48.96

Table 2. Quantitative comparison with DHNN [7] and LIMP [4] (in mm).

each component x^k, and then set $\alpha \cdot l(x^k)$ or $\alpha \cdot circ(x^k)$ as the editing target, where $l(\cdot)$ and $circ(\cdot)$ are the functions measuring length and circumference of parts, respectively. In experiments of editing bone orientation, since Unsup [12] cannot work without target meshes, to allow comparison, we randomly select target meshes from the test set, and calculate joint positions after pose changes to obtain meshes. More visual editing results are shown in Figs. 3 to 5.

Advantages over Statistical Models. First, our method has a better representation ability. Our reconstruction error on DFAUST [1] is significantly less than SMPL [8] by an order of magnitude (4.70 mm vs. 25.36 mm). An example is given in Fig. 6. Second, our method has more fine-grained semantics, which enables flexible human body editing unsupported by statistical models.

Comparison with Other Disentangled Works. Because DHNN [7] and LIMP [4] do not support real (direct) body editing, we only compare the performance of pose/shape transfer with them. As shown in Fig. 7 and Tab. 2, our method achieves excellent results in pose/shape transfer without the data constraints required by DHNN [7] and LIMP [4].

4. User Study

To better evaluate the editing capacity of the proposed representation, we perform a perceptual evaluation with a user study that consists of 3 group tests. The first group shows the results of Unsup [12] and our method on 4 cases of editing bone orientation. The last two groups show the results of HBR [11] and our method on 3 cases of editing bone lengths and part shape sizes, respectively. The users need to evaluate the editing capability of the methods from two aspects: whether the edited attributes are changed to the target value in a natural, reasonable and accurate way and whether the other unedited attributes are left intact. We have collected answers from 102 participants, including 28 females and 74 males of different ages (2 users below 18, 96 users between 18 and 40, and 4 users between 40 and 60). We evaluate the percentage of each method considered to have better performance in changing attributes P_{cha} and preserving attributes P_{pre}. The statistical results of 3 group tests are given in Tab. 3, which demonstrates that our approach has a more flexible and accurate editing capability.

5. Limitations and Failure Cases

Editing bone orientation with our method may fail when the target orientation is uncommon in the training data, as illustrated in Fig. 8. In further work, we will dig deeper into the prior knowledge about the human body to improve the generalization capability of our representation.

References

[1] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1, 2, 5

[2] Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, and Stefanos Zafeiriou. Neural 3D morphable models: Spiral convolutional networks for 3D shape representation learning and generation. In *Int. Conf. Comput. Vis.*, 2019. 1

[3] Zhixiang Chen and Tae-Kyun Kim. Learning feature aggregation for deep 3D morphable models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1

[4] Luca Cosmo, Antonio Norelli, Oshri Halimi, Ron Kimmel, and Emanuele Rodola. LIMP: Learning latent shape representations with metric preservation priors. In *Eur. Conf. Comput. Vis.*, 2020. 1, 2

[5] Zhongpai Gao, Junchi Yan, Guangtao Zhai, Juyong Zhang, Yiyan Yang, and Xiaokang Yang. Learning local neighboring

Method	Bone Orientation		Bone Length		Shape Size	
	P_{cha}	P_{pre}	P_{cha}	P_{pre}	P_{cha}	P_{pre}
Unsup [12]	27.95%	36.27%	-	-	-	-
HBR [11]	-	-	27.43%	34.63%	40.20%	40.16%
Ours	**72.05%**	**63.73%**	**72.57%**	**65.37%**	**59.80%**	**59.84%**

Table 3. The percentage of each method considered to have better editing performance in three editing tasks. - : not supported for this task.

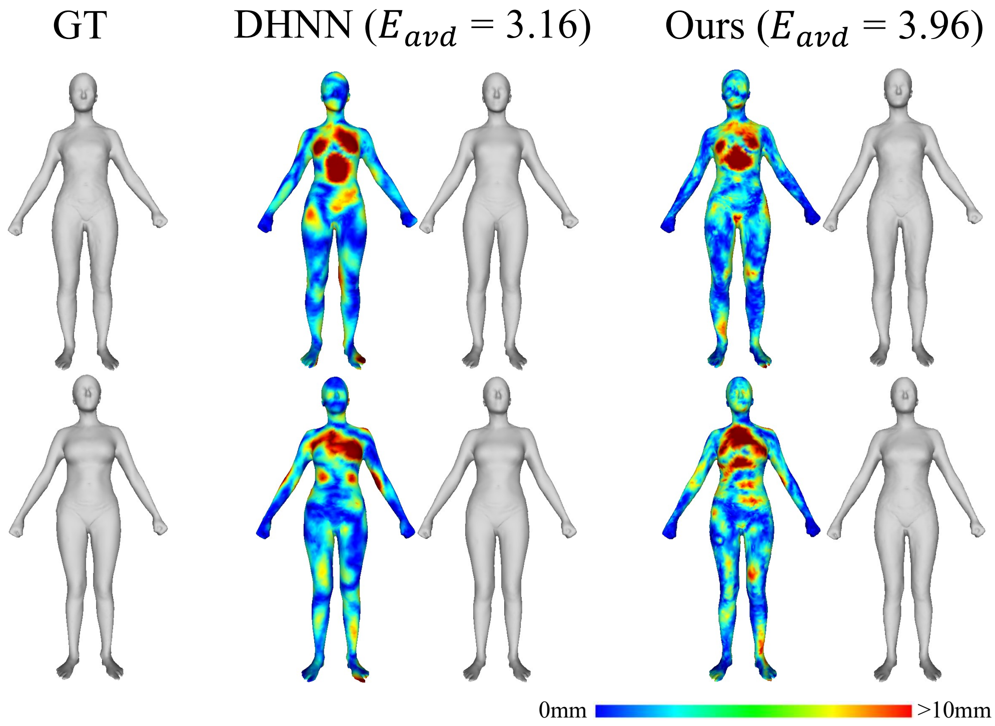


Figure 2. Qualitative reconstruction results on the DHNN dataset [7]. E_{avd} denotes the average point-wise Euclidean distance (in millimeters) between corresponding vertices in the input and its reconstruction.

structure for robust 3D shape representation. In *AAAI*, 2021. 1

[6] Shunwang Gong, Lei Chen, Michael Bronstein, and Stefanos Zafeiriou. SpiralNet++: A fast and highly efficient mesh convolution operator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019. 1

[7] Boyi Jiang, Juyong Zhang, Jianfei Cai, and Jianmin Zheng. Disentangled human body embedding based on deep hierarchical neural network. *IEEE Trans. Vis. Comput. Graph.*, 2020. 1, 2, 3

[8] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.*, 2015. 1, 2

[9] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3D faces using convolutional mesh autoencoders. In *Eur. Conf. Comput. Vis.*, 2018. 1

[10] Yipin Yang, Yao Yu, Yu Zhou, Sidan Du, James Davis, and Ruigang Yang. Semantic parametric reshaping of human body models. In *International Conference on 3D Vision*, 2014. 1

[11] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. 3D human body reshaping with anthropometric modeling. In *International Conference on Internet Multimedia Computing and Service*, 2017. 2, 3

[12] Keyang Zhou, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Unsupervised shape and pose disentanglement for 3D meshes. In *Eur. Conf. Comput. Vis.*, 2020. 1, 2, 3

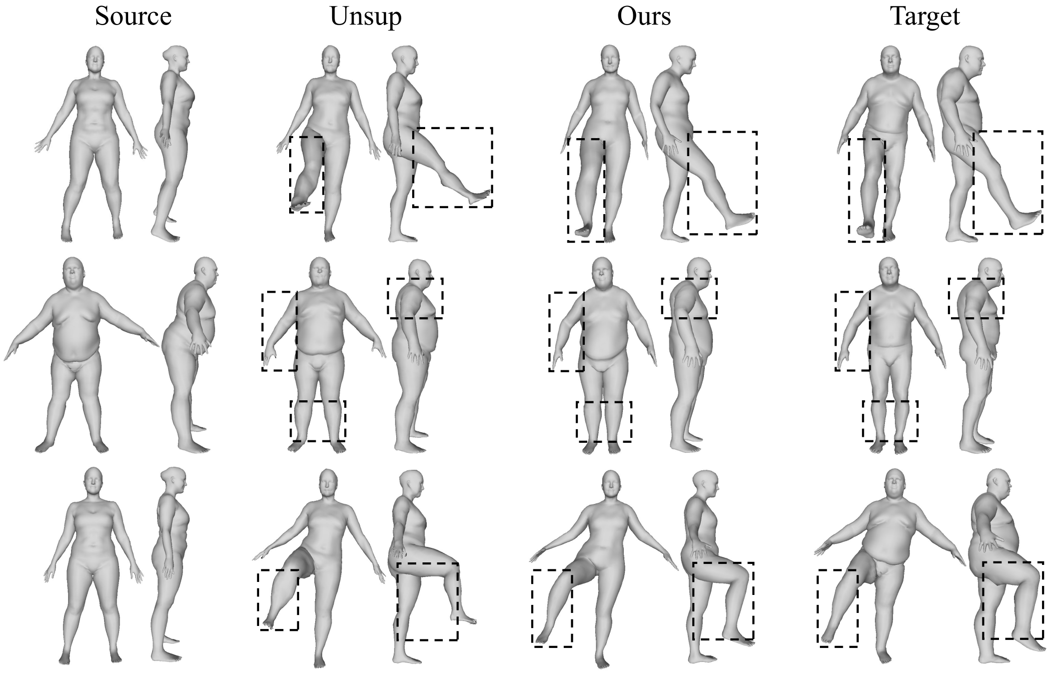


Figure 3. Qualitative results of editing bone orientation.

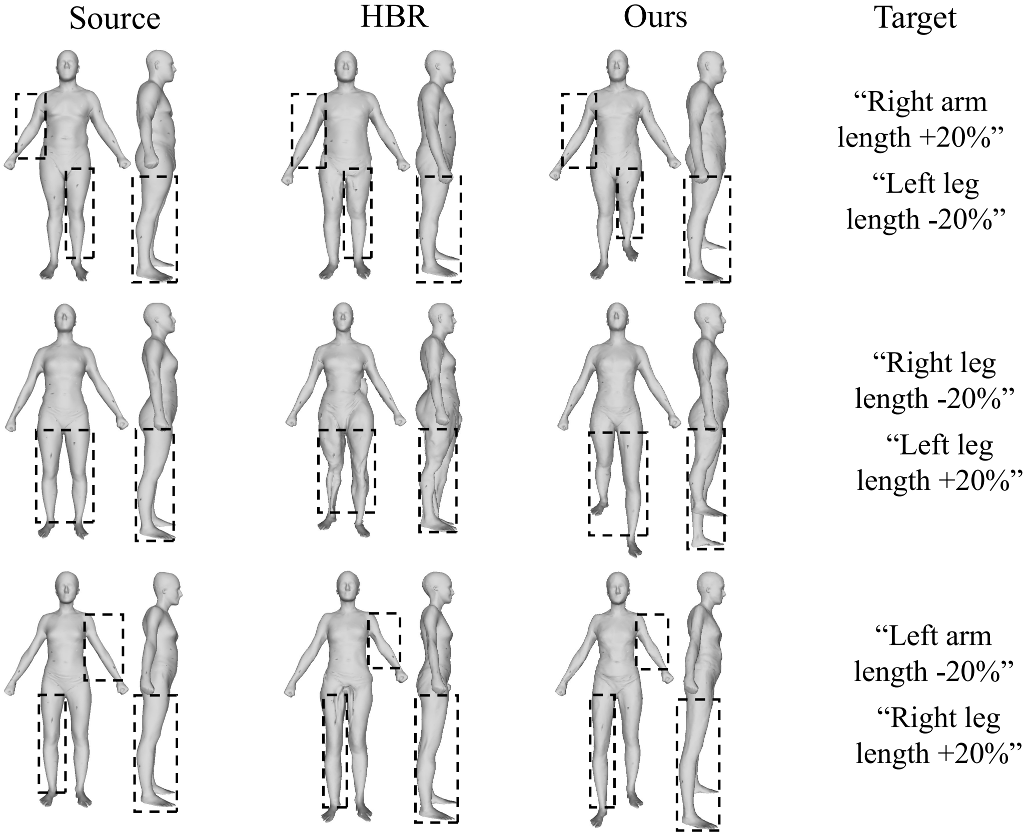


Figure 4. Qualitative results of editing bone lengths.

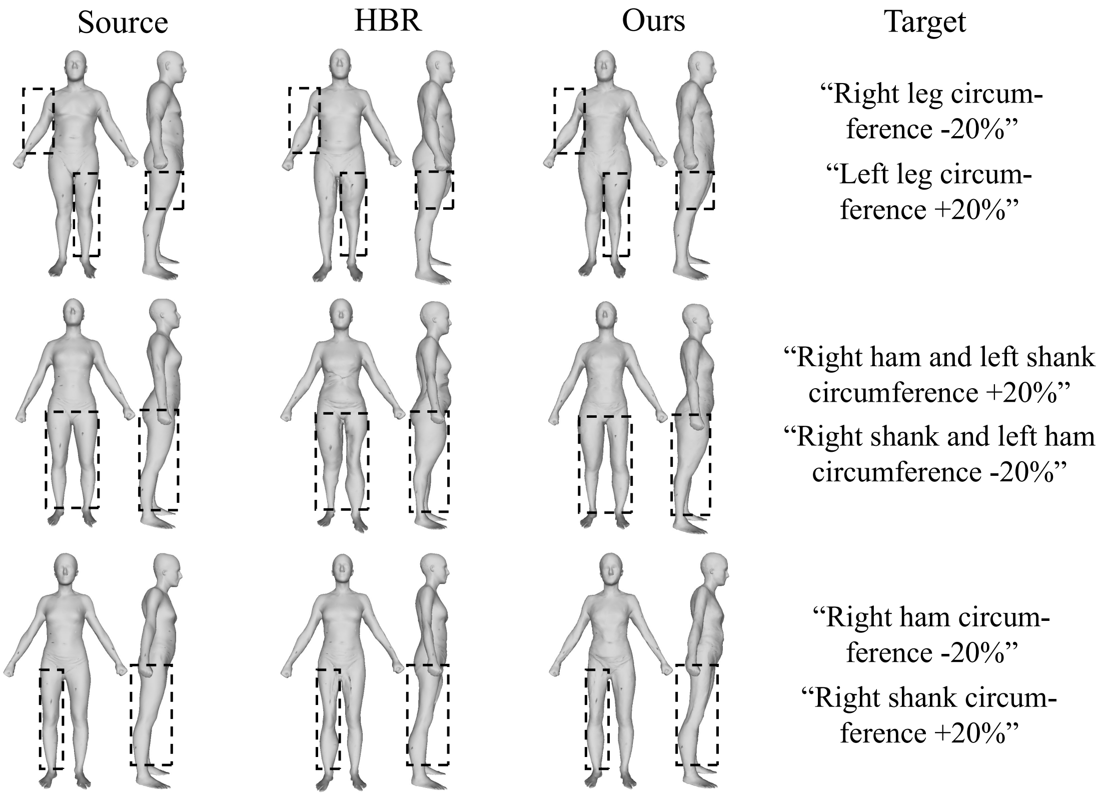

Source
HBR
Ours
Target
“Right leg circumference -20%”
“Left leg circumference +20%”
“Right ham and left shank circumference +20%”
“Right shank and left ham circumference -20%”
“Right ham circumference -20%”
“Right shank circumference +20%”

Figure 5. Qualitative results of editing part shape sizes.

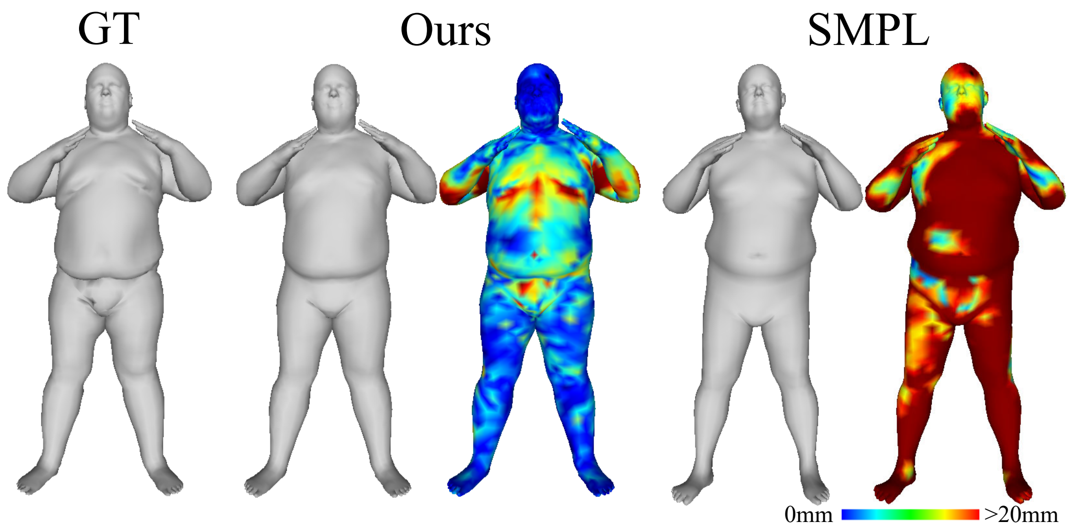


Figure 6. Qualitative reconstruction results on DFAUST [1].

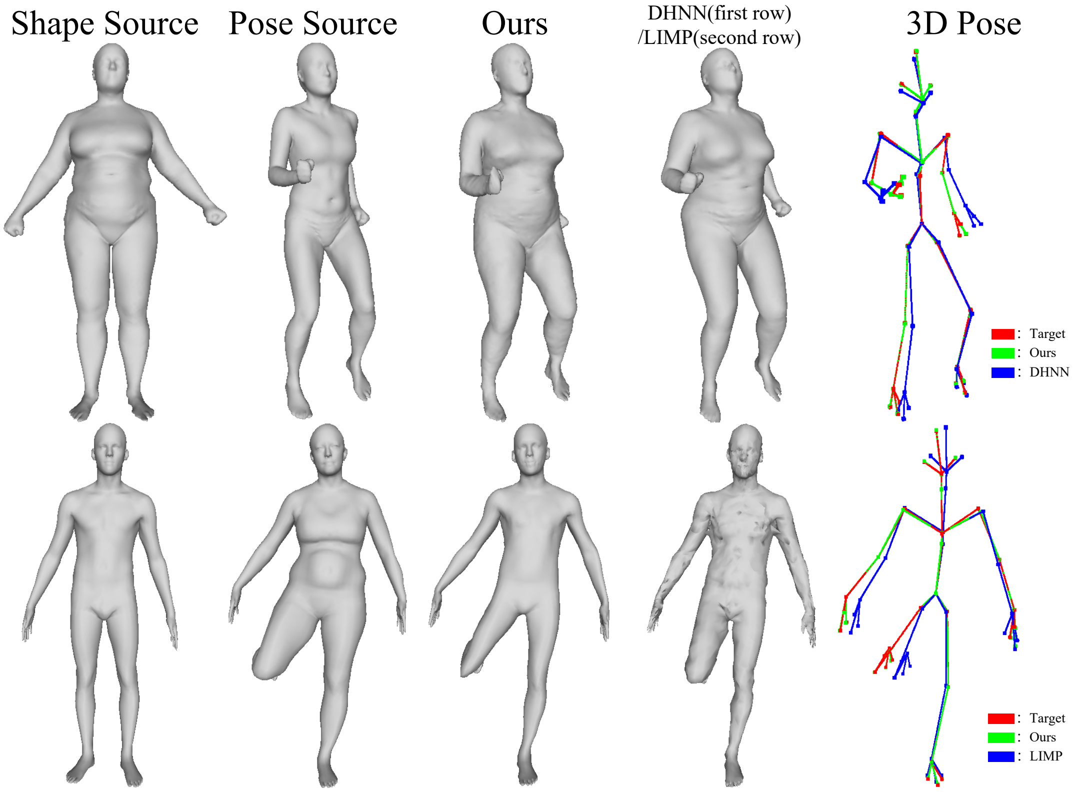


Figure 7. Qualitative pose/shape transfer results.

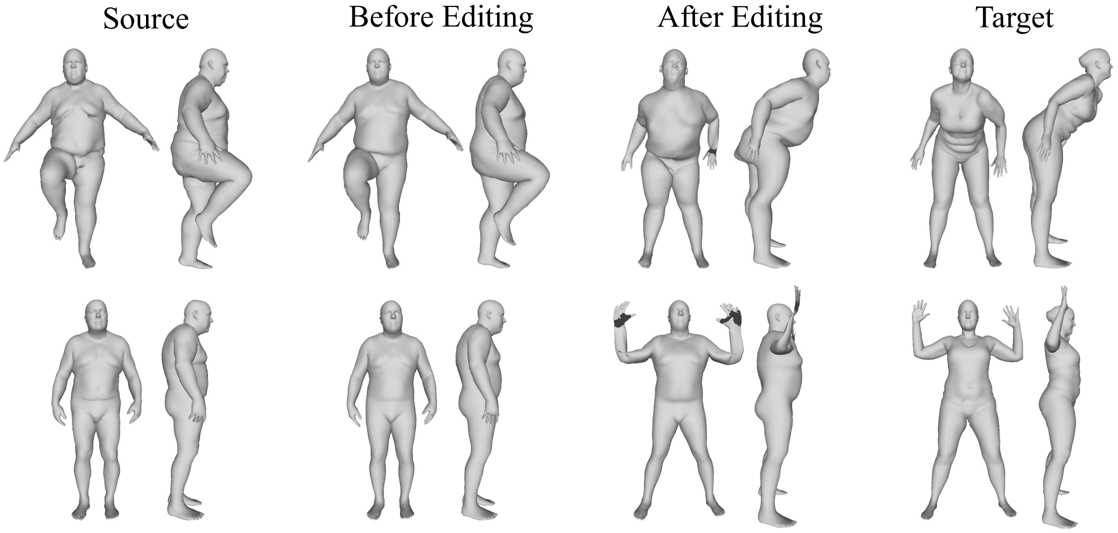


Figure 8. Some examples of failure cases. We show the reconstructed and edited bodies in the second and third columns.