D^3-Human: Dynamic Disentangled Digital Human from Monocular Video

Supplementary Material

1. Experimental details

1.1. Data preprocessing

The matching camera parameters and SMPL [7] parameters can be obtained using VideoAvatar [1]. We construct a tetrahedron [12] that encloses the initial SMPL mesh and slightly expand its range to accommodate scenarios involving loosely fitted clothing. We use SAM2 [9] to obtain 2D human parsing masks and retain the necessary categories such as body and one-piece outfits. Compared to REC-MV [8], our mask does not require additional processing of clothing feature lines. We employ the Sapiens [6] to obtain normals for the clothed human, which are used to optimize the details of the deformation network. To avoid starting training from an initial state that is too far from the target, we pre-warm the SDF network parameters to approximate the shape of the SMPL model. The hmSDF is defined directly at each tetrahedron vertex, without the use of a neural network. It is initialized with random values with a mean of 0 and a variance of 1.

1.2. Body Completion and Fusion.

To separate the invisible body parts, we construct a tetrahedral mesh and SDF representation for the body, where the SDF is initialized to the shape of SMPL. At the same time, a body hmSDF is maintained and optimized in the same way. However, directly extracted body regions occluded by clothing may produce artifacts in areas such as the underarms. Therefore, we further use segmentation edges to cut the densified SMPL mesh. Since the gap between the reconstructed visible body and the invisible SMPL mesh regions is sufficiently small, Poisson reconstruction [5] can easily be used to fuse them together.

2. Experimental Results

2.1. Ablation of Region Aggregation

We present the results before and after using the region aggregation algorithm in Figure 1. Before region aggregation, incorrect segmentation can result in excess fragments and geometric holes. After region aggregation, it is possible to eliminate these fragments and holes.

The clothing and body should each have a fixed number of connected components; for example, the clothing is typically a single complete connection, while the body is usually divided into five connected components (the head and four limbs) due to the presence of clothing. We determine the correctness of each connected component by checking the number of vertices it contains and aggregating incorrectly categorized components into the correct connected components.

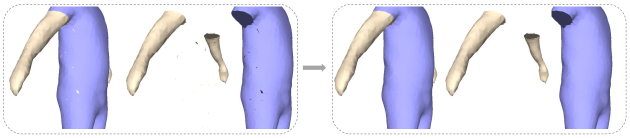

Figure 1. Region aggregation result.

2.2. Abltaion of Collision Distance Value.

When the collision distance value is small, the clothing and body meshes are too close together. Even if the body is indeed inside the clothing, rendering inaccuracies may occur, leading to incorrect judgments of the visible areas. We compare the rendering differences with different collision distance values in Figure 2.

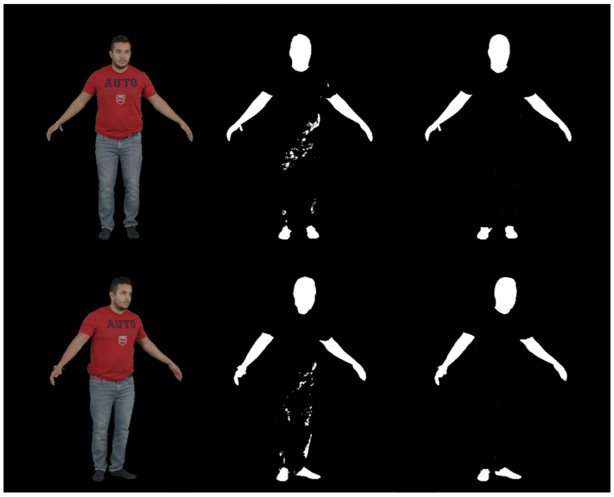

Figure 2. Ablation study on rendering mask inaccuracies caused by collision distance values. When $\epsilon_2 = 0.001$, the clothing and body are too close together, causing rendering inaccuracies that classify some parts of the body occluded by the clothing as visible, resulting in incorrect rendering results. When $\epsilon_2 = 0.005$, the inaccuracies are eliminated, and the rendering results are correct.

2.3. Qualitative Comparisons

We present more qualitative comparisons on the Peoplesnapshot[1] subjects, shown in Figure 3.

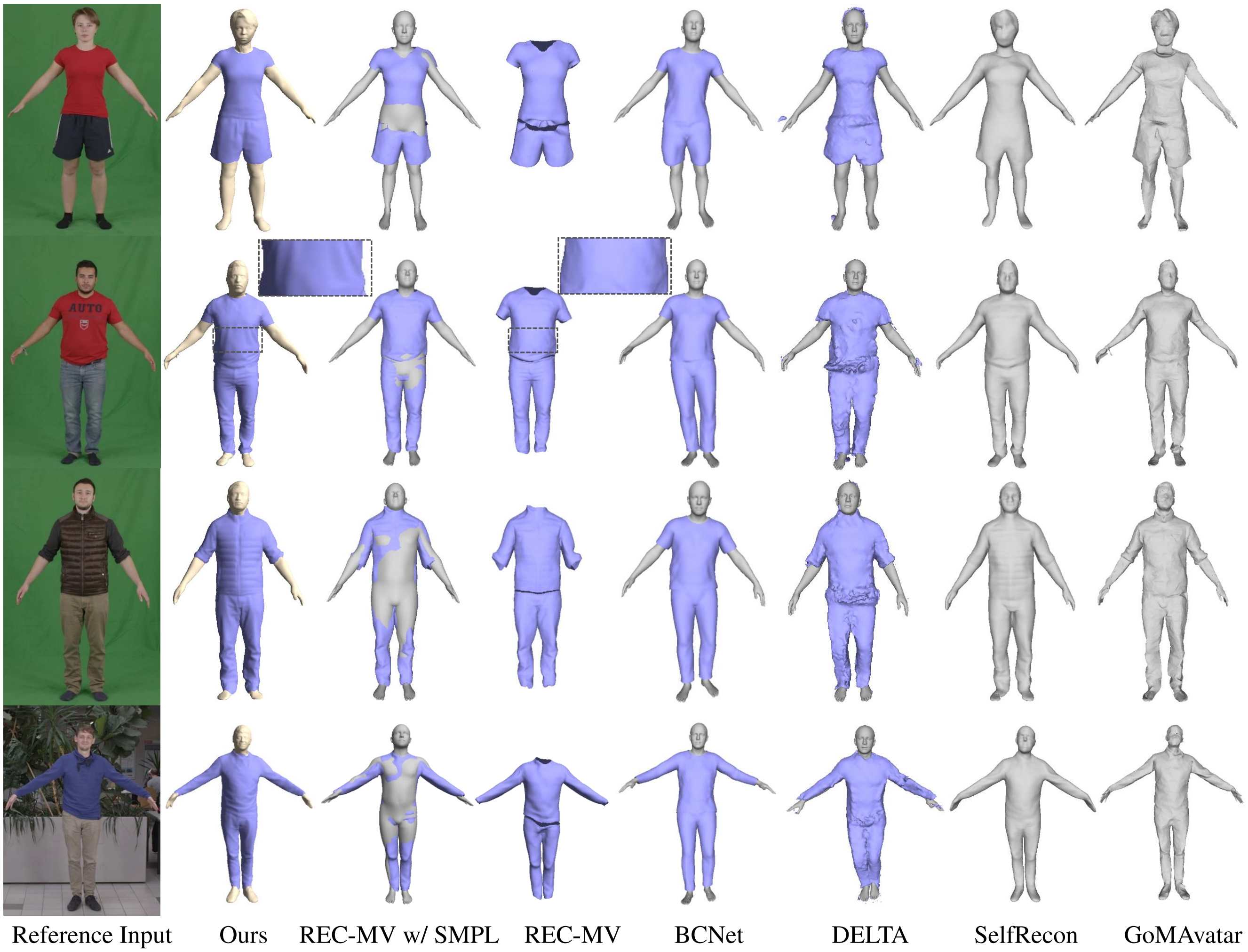

Figure 3. More qualitative comparisons on the Peoplesnapshot dataset. In comparison, our method demonstrates superior qualitative performance in reconstructing detailed clothing and body, as well as in achieving decoupling.

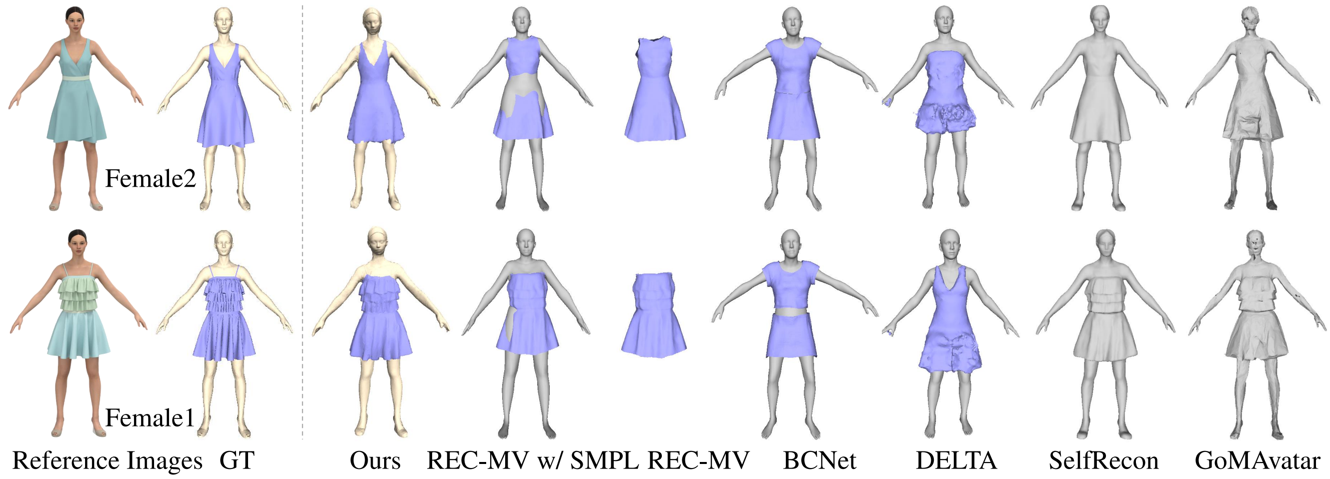

Figure 4. More qualitative comparisons on the SelfRecon dataset.

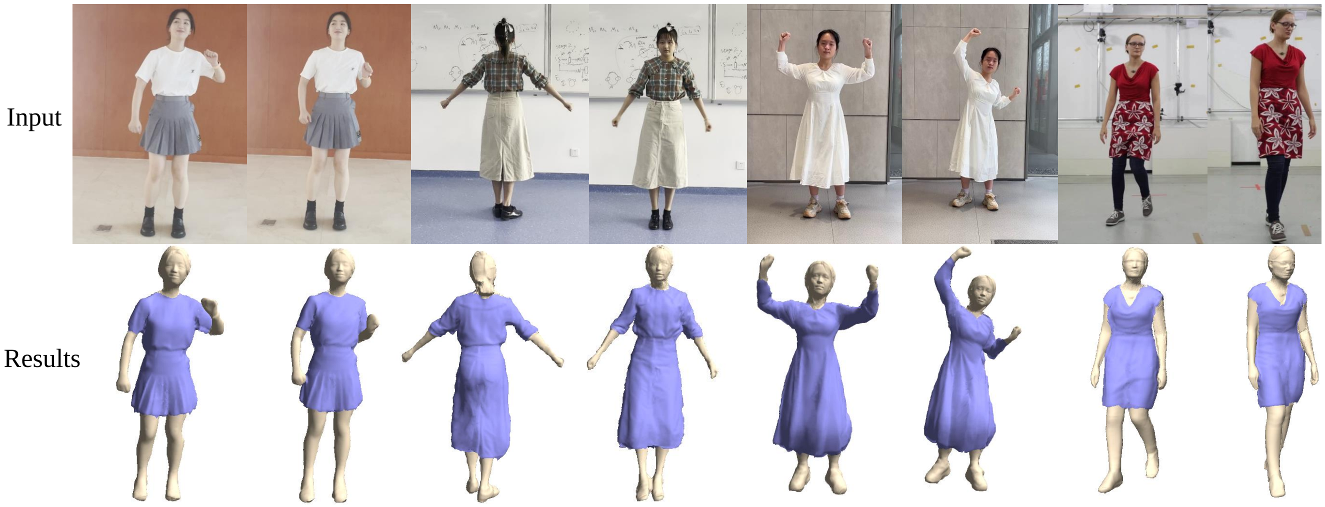

Figure 5. Reconstruction of humans wearing dresses or skirts in in-the-wild poses.

2.4. Quantitative Comparison

We show the visual results of Female 1 and Female 2 from SelfRecon[4] in the quantitative experiments, in Figure 4. Since the reconstruction results of each method may not be in the same space, we use FRICP [13] to align the clothing, body, and clothed human with their respective ground truth counterparts before calculating the Chamfer Distance.

From the results, our method achieves the best numerical accuracy while preserving details during decoupled reconstruction. DELTA [2] attains the second-best accuracy for the body, as it also optimizes the body; however, the visible body appears quite similar to the SMPL model, lacking facial details that closely resemble the input images. REC-MV [8] achieves the second-best accuracy and visual quality for clothing, but its level of detail is less refined compared to our method, and the SMPL model cannot be directly used as the body. SelfRecon [4] produces visually appealing clothed human reconstructions but has lower numerical accuracy due to imprecise body poses, resulting in significant numerical errors. The clothed body reconstructions of GoMAvatar [10] are quite coarse. While BCNet [3] supports accurate decoupled reconstruction, its accuracy is relatively low as it is designed for single-frame scenarios.

2.5. User study of real-world dataset

. We add user studies to the comparison with qualitative comparisons data, in Table 1. For each group of reconstruction results, participants addressed the following queries:

- Reconstruction Quality: Which video demonstrates higher fidelity in body and garment reconstruction (excluding clothed areas, including head, hands, and feet), considering both overall shape similarity and detailed feature preservation compared to the input video?
- Disentangling Ability: Which video achieves proper separation between garments and underlying body geometry while maintaining complete structural integrity?
- Temporal Consistency: Which video best maintains temporal coherence in video sequences, ensuring both reconstruction stability and motion continuity across frames?
- Overall: Considering the combined performance of anatomical reconstruction accuracy, garment-body disentanglement capability, temporal stability, and overall visual plausibility, which video delivers optimal results?

Our method achieves the best performance on all metrics. Our method can represent different types of clothing, while existing methods struggle with open-topology clothing, making their reconstructions less suitable for downstream tasks.

Table 1. Average of 51 user studies, scored from 0 to 1. We assign 1 point to the best for each metric, 0.5 points to the second best, and 0 points to the rest.

Metric	REC-MV	DELTA	BCNet	SelfRecon	GoMavatar	Ours
Reconstruction Quality ↑	0.047	0.060	0.059	0.183	0.004	**0.647**
Disentangling Ability ↑	0.043	0.112	0.135	-	-	**0.710**
Temporal Consistency ↑	0.042	0.031	0.012	0.257	0.002	**0.655**
Overall ↑	0.047	0.049	0.027	0.159	0.00	**0.717**

2.6. Reconstruction in in-the-wild poses

. We show more examples of subjects wearing dresses or skirts, including three public datasets and our collection of long dresses. Figure 5 shows examples including ”female-anran-dance” and ”female-leyang-jump” from REC-MV [8], self-collected data and ”franzi” from MonoPerfCap [11]. The figure selects two frames with different poses from each sequence. Our method can represent common clothing styles (with openings from the manifold surface) in “in the wild” poses (not limited to self-rotation).

2.7. Limitations.

In challenging scenarios such as fast motion or severe occlusions, our method, like non-disentangled approaches, struggles to achieve satisfactory results. The proposed hmSDF can reconstruct common clothing with manifold surface openings but struggles with highly irregular garments such as fringes or stacked long dresses, which is also a limitation of non-decoupled methods.

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[2] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J. Black. Learning disentangled avatars with hybrid 3d representations. *arXiv*, 2023. 3

[3] 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 (ECCV)*, pages 18–35. Springer, 2020. 3

[4] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[5] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006. 1

[6] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024. 1

[7] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multiperson linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 1

[8] Lingteng Qiu, Guanying Chen, Jiapeng Zhou, Mutian Xu, Junle Wang, and Xiaoguang Han. Recmv: Reconstructing 3d dynamic cloth from monocular videos. In *CVPR*, 2023. 1, 3

[9] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[10] Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alex Schwing, and Shenlong Wang. GoMAvatar: Efficient Animatable Human Modeling from Monocular Video Using Gaussians-on-Mesh. In *CVPR*, 2024. 3

[11] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Trans. Graph.*, 37(2):27:1–27:15, 2018. 3

[12] Yuxin Yao, Siyu Ren, Junhui Hou, Zhi Deng, Juyong Zhang, and Wenping Wang. Dynosurf: Neural deformation-based temporally consistent dynamic surface reconstruction. In *European Conference on Computer Vision*, 2024. 1

[13] Juyong Zhang, Yuxin Yao, and Bailin Deng. Fast and robust iterative closest point. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3450–3466, 2021. 3