MultiGO: Towards Multi-level Geometry Learning for Monocular 3D Textured Human Reconstruction

Supplementary Material

1. Implementation Details

1.1. Datasets

The details of the three datasets used in our experiments are listed as follows, including THuman2.0, THuman3.0, and CustomHumans:

- **THuman2.0** [9]: Thuman2.0 is a dataset with 525 high-resolution 3D human scans wearing over 150 different types of clothing. We use this dataset as our training data.
- **THuman3.0** [6]: THuman3.0 is a dataset that contains over 20 combinations of human garments, each containing 15 to 35 high-quality human scans. In this paper, we select 60 scans for all of our experiments and ablation studies. In Table 1, we specifically display the test sample number we have selected.
- **CustomHumans** [1]: CustomHumans is a dataset with 600 high-quality human scans of 80 subjects in over 100 garments and poses. Following the previous work SiTH [2], we selected 60 subjects for all of the experiments and ablation studies.

1.2. Experimental Setting

Training. All experiments are conducted using four NVIDIA A800 GPUs. Our multimodal UNet is initialized with the pre-trained model from the work [7]. We set the input size of the single-view RGB images to 512x512 pixels and the Fourier expansion order (q) to 8. During training, we render human scans online using the official nvdiffrast library. We randomly sample 8 views to generate 8 RGB images, which are used to constrain the proposed Gaussian, and we then select one of these images as the input view, whose camera elevation and azimuth are set to 0,0. Note that we sample the front views at random elevation and azimuth. We only set/assume these degrees of the front view all to zero, to normalize other views. The learning rate for the AdamW [4] optimizer is set to 5×10^{-5}. In training, we use the officially released fitting SMPL-X parameters [9] as input and the default disturbance value (α) is set to 0.25. The pre-trained UNet of the wrinkle-level refinement module is from the work [8]. Specifically, we freeze the VAE and CLIP image encoder and only update the UNet. All of the input images are rendered with nvdiffrast and resized to 512x512 pixels. We randomly selected 8 horizontal camera-rendered images and top and bottom camera-rendered images as our initialization inputs and de-noising conditions. During training, we set the de-noising step k to 1. The learning rate of AdamW optimizer is set to 1×10^{-5}. All models are trained to converge.

Inference. In inference, we estimate the SMPL-X parameters from the input single-view image. SMPL-X parameters are estimated using scripts from SiTH The output of our models is the 3D Gaussian representation, which is transformed into the 3D mesh using the official script file provided by LGM [7]. During the refinement step, the coarse normal map I_n in the input images are rendered online with nvdiffrast library while the RGB images I_c are rendered from the output 3D Gaussian with diff-gaussian-rasterizer [3]. During remeshing, we set the learning rate of the vertices optimizer to 0.3 and the laplacian weight to 0.01. We iteratively update the mesh for 100 steps to obtain the refined mesh.

1.3. CAPE dataset

In this section, we clarify our decision to exclude the CAPE dataset [5] as our test set. As highlighted in the appendix of SiTH [2], the CAPE dataset has several significant shortcomings. It features incomplete input images rendered from unprocessed point clouds, and the ground truth (GT) meshes are of low resolution, failing to accurately correspond to the input images. Additionally, the dataset suffers from limited diversity in human outfitting, as the majority of subjects are depicted in tight clothing such as t-shirts and shorts. We present some samples from the CAPE dataset in Figure 2. Given these limitations, we opted for higher-quality datasets to ensure an unbiased comparison of our method against the latest advancements in the field. Notably, we have identified recently updated datasets, including CustomsHumans [1] and THuman3.0 [6], which offer comprehensive, high-resolution input data and a wider variety of human attire. This choice not only enhances the validity of our comparisons but also reflects our commitment to using the most robust and diverse data available in our research.

2. Limitations

The current implementation of our model faces an efficiency bottleneck during the GS-to-mesh conversion process at the inference stage. This step is resource-intensive, taking approximately 3 minutes to complete and requiring around 50GB of GPU memory. Additionally, while the GS reconstruction step is relatively fast, completing in under a second, and the mesh refinement process is also quicker, taking about 1 minute. Future efforts may include exploring

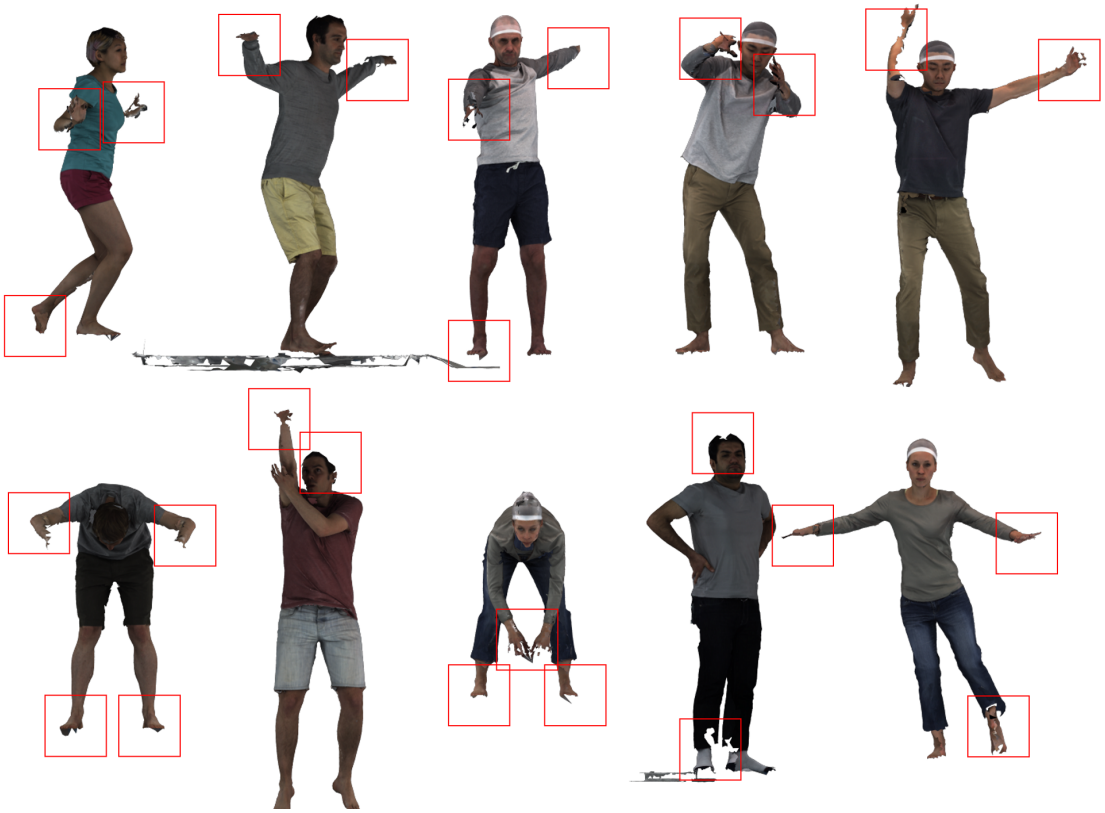

Figure 1. **Samples from the CAPE test set which contain noticeable defects.** We present some input images from the CAPE test set intuitively. From the displayed image, we can see significant issues with the quality of the input image provided by CAPE. Specifically, there are obvious defects in the characters' palms, feet, and head areas. Therefore, to reasonably evaluate the comparison with SOTA methods, we chose the latest and higher quality CustomsHumans and THuman3.0 as our test set.

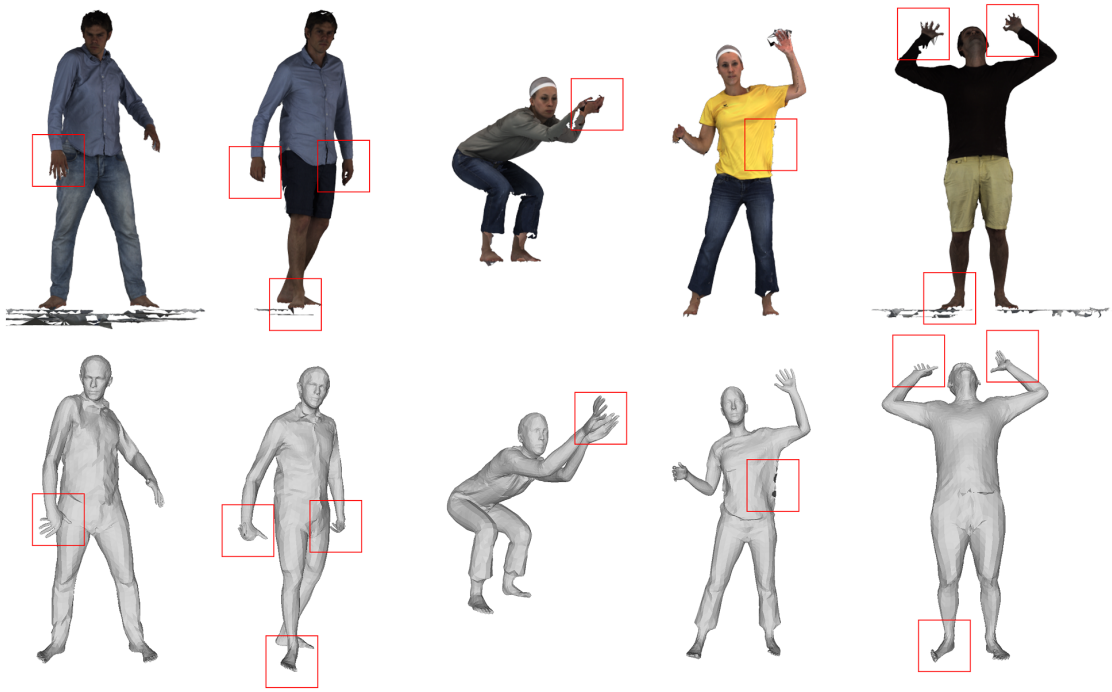

Figure 2. **Samples from the CAPE test set.** Comparing with the input images, it can be found that these GT meshes do not fully correspond to them. For example, in the palms and feet, GT meshes will produce varying degrees of distortion

Figure 3. **More examples to illustrate the effectiveness of the proposed WLR module.** The first row shows the normal map rendered from the mesh before the introduction of the WLR module, and the second row shows the normal map rendered by the mesh after the introduction of the WLR module.

Figure 4. **More examples to illustrate the effectiveness of the proposed WLR module.** The first row shows the normal map rendered from the mesh before the introduction of the WLR module, and the second row shows the normal map rendered by the mesh after the introduction of the WLR module.

alternative algorithms, optimizing existing code, or leveraging more advanced hardware capabilities to alleviate this bottleneck.

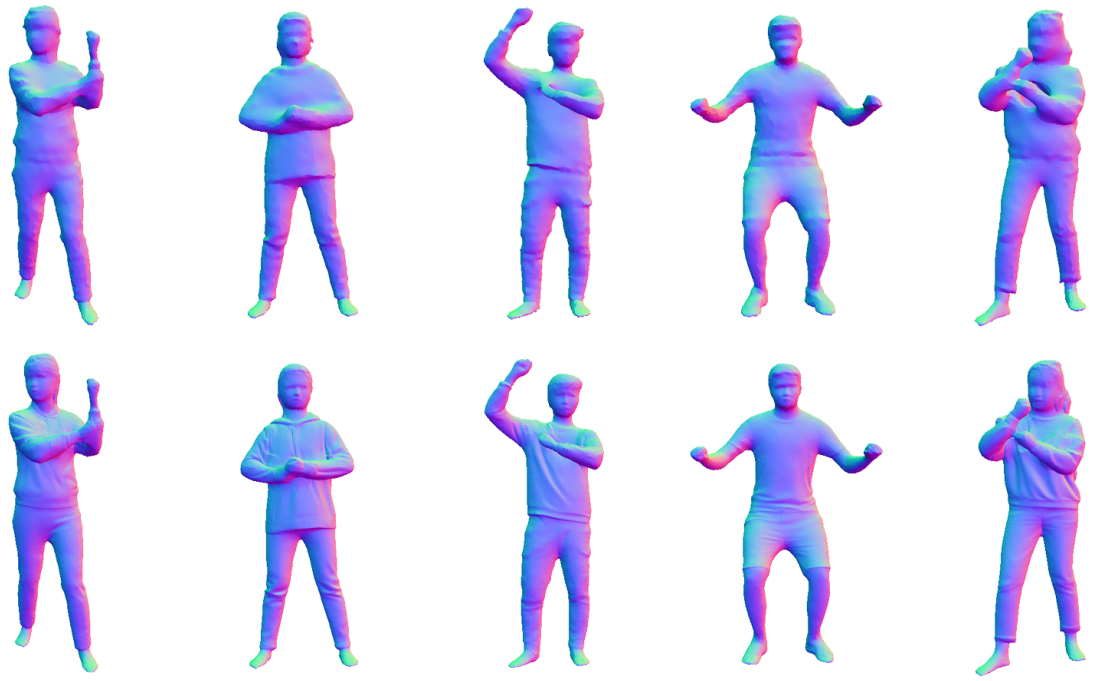

Figure 5. **More examples to illustrate the effectiveness of the proposed WLR module.** The first row shows the normal map rendered from the mesh before the introduction of the WLR module, and the second row shows the normal map rendered by the mesh after the introduction of the WLR module.

Figure 6. **More examples to illustrate the effectiveness of the proposed SLE module.** The first row shows the side rendering of the reconstructed human before the introduction of the SLE module. The second row shows the side rendering of the reconstructed human after the introduction of the SLE module. The third row shows the side the rendering of the ground truth.

Scan ID in our Exp.	Scan ID in THuman3.0	Subject ID	Scan ID in our Exp.	Scan ID in THuman3.0	Subject ID
1	00001_0033	1	31	00008_0042	11
2	00001_0069	1	32	00008_0046	11
3	00001_0070	1	33	00008_0032	11
4	00001_0047	2	34	00008_0049	12
5	00001_0049	2	35	00008_0052	12
6	00001_0052	2	36	00008_0057	12
7	00003_0003	3	37	00023_0012	13
8	00003_0013	3	38	00023_0080	13
9	00003_0018	3	39	00023_0008	13
10	00003_0021	4	40	00024_0014	14
11	00003_0035	4	41	00024_0023	14
12	00003_0036	4	42	00024_0025	14
13	00004_0007	5	43	00024_0039	15
14	00004_0014	5	44	00024_0043	15
15	00004_0022	5	45	00024_0052	15
16	00005_0022	6	46	00025_0004	16
17	00005_0023	6	47	00025_0005	16
18	00005_0005	6	48	00025_0006	16
19	00005_0042	7	49	00026_0026	17
20	00005_0045	7	50	00026_0034	17
21	00005_0048	7	51	00026_0039	17
22	00006_0022	8	52	00027_0005	18
23	00006_0006	8	53	00027_0032	18
24	00006_0007	8	54	00027_0027	18
25	00007_0009	9	55	00028_0034	19
26	00007_0021	9	56	00028_0025	19
27	00007_0030	9	57	00028_0020	19
28	00008_0005	10	58	00060_0018	20
29	00008_0013	10	59	00060_0010	20
30	00008_0044	10	60	00060_0028	20

Table 1. Details about the 60 scans from THuman3.0 used in our experiment. We report the Scan ID in our experiment its corresponding ID in Thuman3.0 and the subject ID.

3. More Experimental Results

3.1. Visualization on Ablation Study

Additional Examples Demonstrating the Effectiveness of the Proposed WLR Module. Figures 3, 4, and 5 provide further insights into the impact of the Wrinkle-Level Refinement (WLR) module by comparing results before and after its implementation. The results clearly illustrate that the WLR module significantly enhances the geometric quality of the reconstructed mesh, particularly in capturing intricate details such as clothing wrinkles and facial features.

Additional Examples Demonstrating the Effectiveness of the Proposed SLE Module. Figure 6 presents additional results that highlight the effects of incorporating the Skeleton-Level Enhancement (SLE) module. The comparison reveals the SLE module effectively aids in reconstructing the target human geometry, resulting in a reconstructed mesh that closely resembles the ground truth mesh.

References

[1] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21024–21035, 2023. 1

[2] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 2023. 1

[4] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[5] 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

[6] Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. Deepcloth: Neural garment representation for shape and style editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1581–1593, 2023. 1

[7] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 1

[8] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv preprint arXiv:2405.20343*, 2024. 1

[9] 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)*, 2021. 1