DeClotH: Decomposable 3D Cloth and Human Body Reconstruction from a Single Image

Supplementary Material

In this supplementary material, we present additional technical details and more experimental results that could not be included in the main manuscript due to the lack of pages. The contents are summarized below:

S1. Controlling reconstruction results

Our proposed DeClotH has the advantage of easily modifying the reconstructed results for virtual try-on and pose deformation. Fig. S1 illustrates the examples of controlling the reconstruction results. First, we can transfer the reconstructed 3D clothes into a new 3D avatar, by fitting 3D clothes based on the SMPL+H human model (blue part of the figure). Second, by forwarding new SMPL+H pose parameters to the linear blend skinning (LBS) of our pipeline, we can animate the reconstruction results (green part of the figure). Like these examples, reconstructing separate 3D geometries is highly useful for applying human reconstruction systems for various downstream applications.

Figure S1. **Examples for controlling 3D reconstruction results.** Our reconstruction results are editable, such as virtual try-on and pose deformation.

S2. Evaluation of pose deformation

We demonstrate that our DeClotH is also superior to existing methods in applying pose deformation to reconstruction results. For the evaluation, we deform the reconstruction results with GT human pose parameters of 4D-DRESS [53] test set. 4D-DRESS contains sequences of 3D cloth and human scans, driven by human pose parameters. Using these pose parameters, we deform the reconstructed meshes to follow the first pose of each sequence. Then, we evaluate the deformed meshes based on the GT 3D scans corresponding to the first pose. The evaluation results are shown in Tab. S1, indicating our DeClotH has the advantage over other methods for animating reconstruction results with novel human poses.

S3. Evaluation with POR Score

We provide more quantitative comparison results through POR Score (pixel-wise object removal score) proposed by Kim *et al.* [24]. The POR Score is devised to evaluate the quality of 3D decomposition in the absence of 3D cloth GT scans. This metric measures the proportion of remaining cloth pixels in rendered human body images, after performing cloth decomposition. Specifically, given a reconstructed 3D human body with the target cloth removed, we render 30 images using uniformly distributed camera viewpoints. Subsequently, we run the off-the-shelf image segmentation method, SAM [26], to obtain the cloth segmentation corresponding to the cloth prompt. Here, the cloth prompts are acquired by running the image captioning method, BLIP [31]. From the obtained segmentations, the POR Score measures the ratio of pixels classified as the tar-

Methods	4D-DRESS (cloth)				4D-DRESS (cloth + human)			
	CD↓	NC↓	PSNR↑	LPIPS↓	CD↓	NC↓	PSNR↑	LPIPS↓
GALA* [24]	5.251	0.044	25.390	0.069	2.844	0.088	19.454	0.117
SiTH [17] + GALA [24]	6.560	0.042	27.578	0.065	3.364	0.078	21.886	0.102
TeCH [20] + GALA [24]	4.425	0.033	29.271	0.044	2.422	0.059	23.276	0.070
DeClotH (Ours)	**2.782**	**0.030**	**31.489**	**0.033**	**2.271**	**0.055**	**23.369**	**0.067**

Table S1. **Quantitative comparisons of pose deformation with 3D cloth decomposition and 3D cloth human reconstruction methods.**

Methods	POR Score↓
GALA* [24]	0.418
SiTH [17] + GALA [24]	0.246
TeCH [20] + GALA [24]	0.225
DeClotH (Ours)	**0.218**

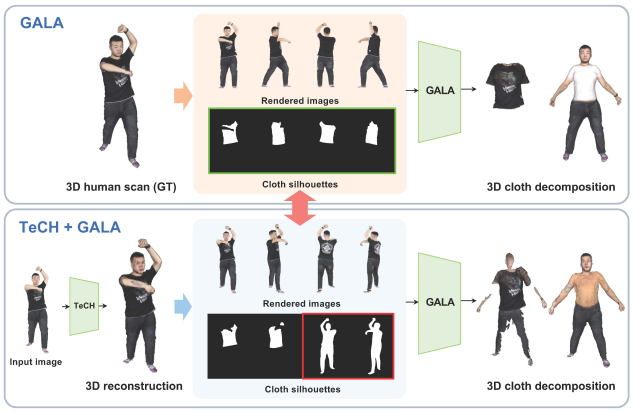
Table S2. **Quantitative comparisons of POR Score [24] with 3D cloth decomposition and 3D cloth human reconstruction methods, on 4D-DRESS [53].**

get cloth in the image. A lower POR Score indicates better performance of the 3D cloth decomposition. As shown in Tab. S2, our framework also outperforms the other methods in POR Score, which demonstrates that DeClotH achieves better results in 3D cloth and human body decomposition.

S4. Implementation details

Network architecture. The DMTets, which are optimized at the geometry stage, are implemented by using two fully-connected layers with 32 hidden dimensions and ReLU activations. The DMTets take the 3D vertex coordinates of the tetrahedral grid $(\mathbf{X}_T, T)$ as input, where the coordinates are normalized between -0.5 and 0.5. Then, the coordinates are encoded by a hash positional encoding [39] with a maximum resolution of 1024 and 16 resolution levels. The MLP networks, which are optimized at the texture stage, are implemented by using a fully-connected layer with 32 hidden dimension and ReLU activations. The MLP networks take the mesh coordinates as input, after applying the hash positional encoding with a maximum resolution of 2048. Additionally, we implement a MLP network, which takes camera parameter $\mathbf{k}$ and produces adaptive background colors of the rendering pipeline, using two fully-connected layers.
Optimization details. PyTorch [43] is used for the implementation. In both the geometry and texture stages, we use Adam optimizer [25] with 4000 optimization steps. The initial learning rate is set to 0.001 and reduced by an exponential scheduler, $\eta = 0.001 \times 0.1^{step/4000}$. During the optimization process, we render 3D cloth and human body based on the spherical coordinate system, (r, θ, ϕ), where r denotes the distance from the spherical origin, θ denotes the elevation angle, and ϕ denotes the azimuth angle. We set $r \in [0.7, 1.3]$, $\theta \in [-30°, 30°]$, and $\phi \in [-180°, 180°]$, with uniform sampling. To capture fine details of human faces, we additionally use zoomed-in camera views for the rendering. Specifically, we set the spherical origin to the 3D position of SMPL+H head keypoint, $r \in [0.3, 0.4]$, $\theta \in [-90°, 90°]$, and $\phi \in [-90°, 90°]$. All the experiments are conducted with an NVIDIA Quadro RTX 8000 GPU.
Training details for ClothDiffusion. To train ClothDiffusion described in Sec. 4, we adopt StableDiffusion [47] in version 1.5. The weights of ClothDiffusion are updated by Adam optimizer [25] with 200k training steps and a mini-batch size of 8. The learning rate is set to 10^{-5}. We train the model with an NVIDIA Quadro RTX 8000 GPU.

(a) Error propagation of 3D human reconstruction



(b) Domain gap in rendered images

Figure S2. **Failure examples of two-stage reconstruction methods**: (a) propagation of 3D reconstruction error and (b) domain gap in rendered images.

S5. Discussion of two-stage reconstruction

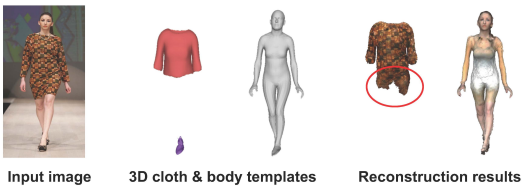
In this section, we provide a deep discussion about the advantages of our DeClotH compared to the two-stage reconstruction methods, SiTH [17] + GALA [24] and TeCH [20] + GALA [24]. We suggest that the two-stage reconstruction methods have two drawbacks: 1) propagation of 3D reconstruction error and 2) domain gap in the rendered images.
Error propagation of 3D human reconstruction. Fig. S2

(a) illustrates that the 3D geometric errors from the 3D human reconstruction significantly affect 3D cloth decomposition errors. In the first row of the figure, GALA [24] accurately decomposes 3D cloth when provided with a 3D human GT scan, which is naturally free of geometric artifacts. On the other hand, in the second row of the figure, TeCH [20] produces the 3D geometric error in reconstructing the ankle part, leading to the annihilation of ankle parts in the 3D cloth decomposition. The primary discrepancy lies in 3D human reconstruction methods overlooking the geometric relationship between the cloth and the human body, leading to overly thick or thin reconstructions. While these thick or thin reconstructions appear visually acceptable, they are critically detrimental to 3D cloth decomposition. Unlike the two-stage approach, our DeClotH considers the volumetric space for 3D cloth decomposition during the reconstruction process. Therefore, DeClotH is free from error propagation issue and provides accurate reconstructions of 3D cloth and human body.

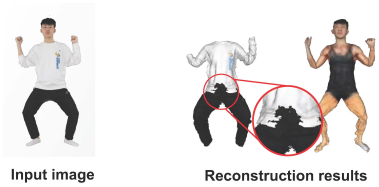
Domain gap in rendered image. Fig. S2 (b) shows that there is a domain gap issue in rendered images between real-world and reconstructed 3D avatars, leading to wrong 3D cloth decomposition. The 3D cloth decomposition method, GALA [24], runs based on the cloth silhouettes from the rendered images of a given 3D avatar. Here, the cloth silhouettes are acquired through the image segmentation method, SAM [26]. GALA (first row of the figure) results in the accurate decomposition result by utilizing correct cloth silhouettes for all rendered images. In contrast, TeCH+GALA (second row of the figure) produces the erroneous result since the cloth segmentation often fails. We conjecture that the failure of the cloth segmentation is the domain gap in rendered images. Based on the 3D human reconstruction results of TeCH [20], its rendered images have artificial appearances compared to real images. Such artificial appearances adversely affect the decomposition of 3D clothes from the 3D human reconstruction results. On the other hand, our proposed DeClotH is a one-stage method that does not require performing segmentation for rendered images. Therefore, our DeClotH does not have the domain gap issue, which is an advantage over the two-stage reconstruction methods.

S6. Limitations and future works

Diversity of cloth shape. There is a limitation in reconstructing diverse cloth types (*e.g.*, dress), as shown in Fig. S3 (a). This is mainly due to the expression power of the cloth template model (*i.e.*, SMPLicit [8]). Most of the existing cloth template models [4, 8, 10, 22, 38] have difficulty in modeling the wide variety of 3D cloth geometries in the real world. Thereby, for several uncommon clothes, predicting 3D cloth templates often fails, and DeClotH's reconstruction based on the cloth templates also produces erroneous results. Improving the expression power of cloth template models should be a future research direction.



(a) Diversity of cloth shape



(b) Inter-penetration

Figure S3. **Failure cases of our proposed framework.**

Inter-penetration. Fig. S3 (b) shows that our framework often suffers from inter-penetration in reconstructed 3D clothes. This inter-penetration issue is extremely challenging, as it requires reasoning not only about the geometric relationship between the cloth and the human body, but also among different clothes. Accordingly, we aim to extend our framework to efficiently reconstruct 3D clothes while overcoming the inter-penetration issue.

S7. More qualitative results

We provide more qualitative comparisons of 3D clothing reconstruction on 4D-DRESS [53] and THuman2.0 [57]. Figs. S4 and S5 show that our DeClotH produces far more accurate reconstructions of 3D cloth and human body compared to the prior arts. Fig. S6 demonstrates that DeClotH also achieves superior reconstruction performance on in-the-wild images.

Fig. S7 shows the qualitative comparison of StableDiffusion [47], HumanDiffusion [60], and our proposed ClothDiffusion. Compared to StableDiffusion and HumanDiffusion, ClothDiffusion specializes in cloth image generation, excluding other contents. Additionally, ClothDiffusion accurately generates cloth images in the desired regions corresponding to the condition images.

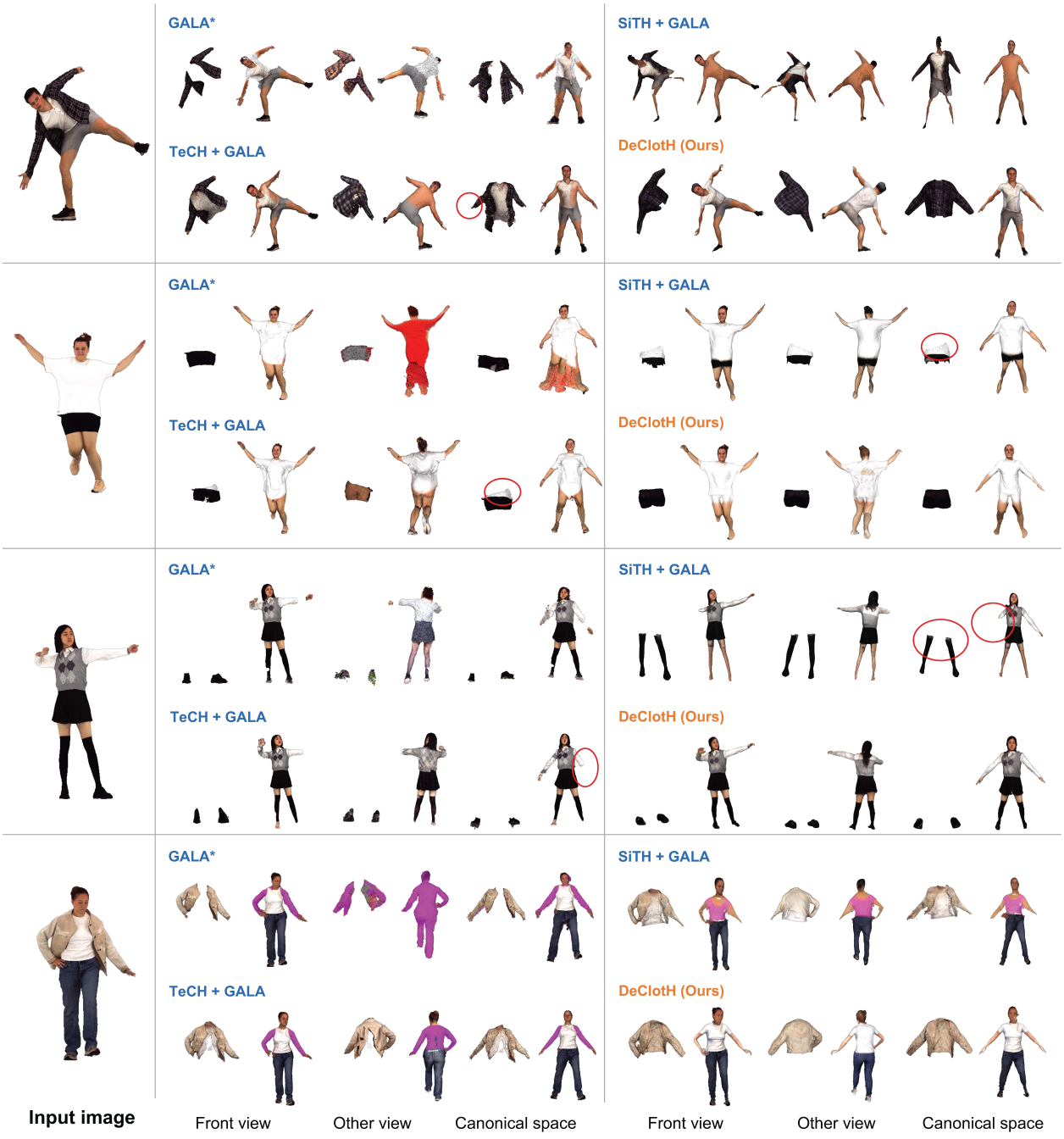


Figure S4. **Qualitative comparison with 3D cloth decomposition and 3D clothed human reconstruction methods: GALA* [24], SiTH [17]+GALA [24], and TeCH [20]+GALA [24], on 4D-DRESS [53].** $*$ denotes the algorithm is modified to take a single image as input instead of a 3D scan. We highlight their representative failure cases with red circles.

Figure S5. **Qualitative comparison with 3D cloth decomposition and 3D clothed human reconstruction methods: GALA* [24], SiTH [17]+GALA [24], and TeCH [20]+GALA [24], on THuman2.0 [57].** $*$ denotes the algorithm is modified to take a single image as input instead of a 3D scan. We highlight their representative failure cases with red circles.

Figure S6. **Qualitative comparison with 3D cloth decomposition and 3D clothed human reconstruction methods: GALA* [24], SiTH [17]+GALA [24], and TeCH [20]+GALA [24], on in-the-wild images.** $*$ denotes the algorithm is modified to take a single image as input instead of a 3D scan. We highlight their representative failure cases with red circles.

Figure S7. **Qualitative comparison of cloth image generation between StableDiffusion [47], HumanDiffusion [60], and our proposed ClothDiffusion.**