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

Hyeongjin Nam[1,3]   Donghwan Kim[1]   Jeongtaek Oh[2]   Kyoung Mu Lee[1,2]

[1]Dept. of ECE&ASRI, [2]IPAI, Seoul National University, [3]KRAFTON

{namhjsnu28, dh971106, ohjtgood, kyoungmu}@snu.ac.kr

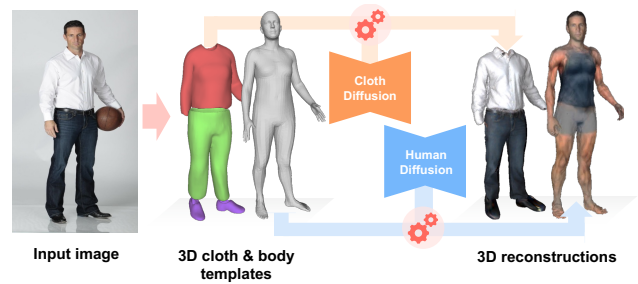https://hygenie1228.github.io/DeClotH/

Abstract

*Most existing methods of 3D clothed human reconstruction from a single image treat the clothed human as a single object without distinguishing between cloth and human body. In this regard, we present **DeClotH**, which separately reconstructs 3D cloth and human body from a single image. This task remains largely unexplored due to the extreme occlusion between cloth and the human body, making it challenging to infer accurate geometries and textures. Moreover, while recent 3D human reconstruction methods have achieved impressive results using text-to-image diffusion models, directly applying such an approach to this problem often leads to incorrect guidance, particularly in reconstructing 3D cloth. To address these challenges, we propose two core designs in our framework. First, to alleviate the occlusion issue, we leverage 3D template models of cloth and human body as regularizations, which provide strong geometric priors to prevent erroneous reconstruction by the occlusion. Second, we introduce a cloth diffusion model specifically designed to provide contextual information about cloth appearance, thereby enhancing the reconstruction of 3D cloth. Qualitative and quantitative experiments demonstrate that our proposed approach is highly effective in reconstructing both 3D cloth and the human body.*

1. Introduction

Reconstructing 3D cloth and human body from a single image is an essential task for various applications including virtual try-on and AR/VR. In recent years, numerous 3D clothed human reconstruction methods [1, 17, 20, 59] have emerged with the advent of diffusion models [18]. Although these methods achieve impressive reconstruction quality, they are inherently designed not to decompose the 3D cloth and human body, limiting their downstream applications. In this regard, we tackle the more challenging task of sepa-

(a) Simplified pipeline of DeClotH

(b) Decomposable 3D cloth and human reconstruction

Figure 1. **Overview of DeClotH.** Given a single image, our framework reconstructs 3D cloth and human body based on the 3D cloth and body templates.

rately reconstructing the 3D cloth and human body directly from a single image, as illustrated in Fig. 1.

Despite the potential applications of decomposable 3D cloth and human body reconstruction, it has not been extensively explored. One major problem is severe occlusion between the cloth and human body, with cloth covering substantial portions of the human body surface. Such an occlusion makes it difficult to infer the overall geometry and texture of the invisible parts between 3D cloth and human body. Additionally, image evidence (*e.g.*, cloth silhouette) of the input image is often imperfect due to occlusion, leading to reconstruction that can overfit to the imperfect evidence. For these reasons, the reconstruction of decompos-

"blue t-shirt" "gray suit jacket"

"yellow sneakers" "red hoodie"

"white short skirt" "orange lossy pants"

Generated image Condition image Generated image Generated image Condition image Generated image

StableDiffusion ClothDiffusion (Ours) StableDiffusion ClothDiffusion (Ours)

Figure 2. **Comparison between an existing diffusion model and ClothDiffusion.** Unlike the representative diffusion model, StableDiffusion [47], our ClothDiffusion generates cloth-specific images and can be controlled by cloth and human body templates.

able 3D cloth and human body is considerably more challenging than the previous tasks, which do not account for the occlusion between cloth and human body.

Recently, score distillation sampling (SDS) [45] loss function has gained popularity in the 3D clothed human reconstruction literature [1, 17, 20, 59] for inferring the geometry and texture of the occluded human parts. The SDS loss enhances reconstruction quality by leveraging the image prior knowledge of a pre-trained text-to-image diffusion model. Although employing the diffusion model is also promising for reconstructing 3D cloth, we observe that naively using such a strategy can provide incorrect guidance about cloth appearance. As shown in Fig. 2, a representative diffusion model, StableDiffusion [47], typically generates images that focus on both cloth and human body, rather than the cloth itself. Since its image prior knowledge contains a mixture of both cloth and human body, the diffusion model is unsuitable for reconstructing 3D cloth from the human body separately. Furthermore, the generated images exhibit excessive diversity in scale, position, and cloth deformation, making it difficult to guide reconstruction of the actual cloth regions.

To address the above challenges, we present **DeClotH** (**De**composable 3D **Clot**h and **H**uman body reconstruction), a template-based optimization framework designed for reconstructing 3D cloth and the human body from a single image. This framework utilizes 3D template models of cloth and human body as strong geometric priors for reconstruction to mitigate erroneous results caused by occlusion. The 3D template models represent the typical shapes of real-world clothes and human bodies by parameterizing them into a low-dimensional latent space. For example, SMPLicit [8] parameterizes 3D clothes with cloth style and

looseness, and SMPL [36] parameterizes 3D human bodies with human pose and shape. Based on these template models, we design a template regularization loss function, which constrains the reconstructed 3D cloth and human body to be close to their 3D template models. Such constraints reduce heavy reliance on the imperfect image evidence caused by occlusions in the input image, leading to more plausible shapes for 3D cloth and human body. Consequently, leveraging 3D template models of cloth and the human body, our framework produces robust reconstructions under severe occlusion between the cloth and human body.

Additionally, we devise a new diffusion model, ClothDiffusion, which overcomes the drawbacks of the existing diffusion model (*i.e.*, StableDiffusion [47]) for 3D cloth reconstruction. Unlike StableDiffusion, which generates mixed content of cloth and human body, ClothDiffusion is specifically trained to generate only cloth images, as shown in Fig. 2. This attribute of ClothDiffusion is highly beneficial for reconstructing 3D cloth, by providing prior image knowledge specialized for cloth geometry and texture. Additionally, ClothDiffusion can be controlled by incorporating 3D template models as regional information for guidance. From the 3D template models, we extract a cloth silhouette and a human skeleton and forward them to ClothDiffusion. By applying this regional information, ClothDiffusion can provide appropriate guidance that aligns with the actual cloth shape and human pose of the input image. Thus, utilizing ClothDiffusion leads to the delicate geometry and texture of 3D cloth along with reconstructing the 3D human body.

Our extensive experiments demonstrate that De-ClotH produces significantly more accurate reconstruction results than baseline methods, for both 3D cloth and human body. Our contributions can be summarized as follows.

- We present DeClotH, which reconstructs a decomposable 3D cloth and human body from a single image, allowing various applications.
- To address occlusion in reconstruction, we propose using 3D template models of the cloth and human body as beneficial constraints during reconstruction.
- To improve the reconstruction of 3D cloth, we introduce a cloth diffusion model that provides contextual information specialized in cloth geometry and texture.

2. Related works

3D clothed human reconstruction. Most of the pioneering works [2, 3, 9, 14, 15, 19, 21, 33, 41, 42, 48, 49, 52, 54, 55, 62–64] of 3D clothed human reconstruction train their networks based on 3D scan data, such as RenderPeople [46] and THuman2.0 [57] datasets. PIFuHD [49] introduced a coarse-to-fine framework to learn the high-resolution geometry of 3D clothed humans. PHORHUM [3] photo-realistically reconstructs the 3D clothed humans while in-

ferring shading. ECON [55] proposed a method that combines human normal maps with a 3D parametric human body for fine geometric details. Recently, several works [1, 17, 20, 30] have leveraged a pre-trained text-to-image diffusion model [47] to reconstruct the geometry and texture of 3D clothed humans, utilizing the strong prior knowledge of the diffusion model. HumanSGD [1] proposed a human mesh inpainting method with a shape-guided diffusion model. SiTH [17] presented a two-stage pipeline that predicts a back-view image and reconstructs a 3D clothed human based on the front- and back-view images. TeCH [20] utilized a Visual Question Answering (VQA) module to obtain descriptive text prompts from the image as input to the diffusion model. The existing methods reconstruct the 3D clothed human as one unified mesh, which cannot be decomposed into the 3D cloth and human body components. On the other hand, our DeClotH enables the separation of 3D cloth and human body from the reconstructions, allowing for a wide range of applications.

3D template models. 3D template models of cloth and human body are essential for recent 3D clothed human reconstruction [4, 8, 10, 22, 38, 65] and 3D human body reconstruction [5–7, 23, 27–29, 32, 37, 40, 58]. These reconstruction methods predict the parameters of their respective 3D template models to reconstruct 3D clothes or human bodies. BCNet [22] presented a 3D clothed human reconstruction system that predicts the PCA coefficients of 3D cloth template models. ClothWild [38] leveraged a weakly supervised learning strategy for 3D clothed humans by using a 3D cloth template model, SMPLicit [8]. HMR [23] proposed a 3D human body reconstruction framework with adversarial loss to learn plausible 3D pose and shape of the 3D human body template model, SMPL [36]. PIXIE [12] proposed a 3D human reconstruction method that estimates 3D hand pose and facial expression using SMPL-X [44]. These template-based reconstruction methods have a critical drawback because they are constrained by pre-defined template topology. Due to this limitation, they generally reconstruct an overly smoothed mesh that does not capture the actual wrinkles of the clothes. On the other hand, our framework covers fine details through the template-based optimization, integrating image evidence (*e.g.*, cloth silhouettes) with 3D template models in the reconstruction.

3D cloth decomposition. Encouraged by recent attention to 3D virtual try-on, several works [24, 51, 53, 56, 66] have proposed methods to decompose 3D clothes from a 3D human scan, multi-view images, or a video. SIZER [51] introduced a method to segment 3D cloth meshes from a 3D human scan by voting on mesh vertices corresponding to cloth labels. 4D-DRESS [53] leveraged graph cut optimization to decompose 3D clothes from a 3D human scan. GALA [24] proposed a method to utilize a text-to-image diffusion model as a valuable prior to decomposition. Com-

pared to these methods, our framework tackles a much more challenging task, reconstructing decomposed 3D cloth and human body from a single image.

3. DeClotH

Fig. 3 illustrates the overall pipeline of our DeClotH. Given an input image $\mathbf{I}$ and a target cloth type C, our framework optimizes both the target 3D cloth mesh $\mathbf{M}_{cloth}$ and the 3D human mesh $\mathbf{M}_{body}$ not wearing the 3D cloth. To simplify the description, we refer to the 3D human without the target cloth as the "human body". In the following sections, we first describe the 3D geometry representation (Sec. 3.1) and image preprocessing (Sec. 3.2) for the optimization of 3D cloth and human body. Subsequently, we provide detailed descriptions of three core loss functions in our framework: template regularization loss (Sec. 3.3), cloth SDS loss (Sec. 3.4), and human SDS loss (Sec. 3.5). Finally, we explain the overall optimization process (Sec. 3.6).

3.1. 3D geometric representation (DMTet)

In our framework, Deep Marching Tetrahedra [50] (DMTet) is utilized as the 3D geometric representation of 3D cloth and human body. DMTet represents 3D geometry with a deformable tetrahedral grid $(\mathbf{X}_T, T)$, where $\mathbf{X}_T$ denotes 3D vertices of the tetrahedral grid and T defines the tetrahedral structure. Specifically, the MLP network of DMTet predicts the signed distance value from the 3D geometry surface for each vertex of the grid. We adopt two DMTets to represent 3D cloth and human body in a canonical pose (A-pose), respectively. To obtain 3D cloth mesh $\mathbf{M}_{cloth}$ and 3D human body mesh $\mathbf{M}_{body}$ from the DMTets, we extract meshes via the Marching Tetrahedra [11] (MT) algorithm and transform the meshes via a linear blend skinning (LBS), which is pre-defined in the SMPL+H [36] human model.

3.2. Image preprocessing

To optimize DMTets of 3D cloth and human body, we gather multiple optimization targets: normal map $\mathbf{N}$, silhouette $\mathbf{S}$, and 3D template meshes ($\mathbf{M}_{cloth}^t$ and $\mathbf{M}_{body}^t$).
Normal & silhouette estimator. The normal maps $\mathbf{N}$ of the front and back views are obtained by the normal estimator of ECON [55] from the input image. The silhouettes $\mathbf{S}$ of cloth and human are acquired by running the off-the-shelf segmentation method, SAM [26], given input image $\mathbf{I}$ and text prompt ($\mathbf{t}_{cloth}, \mathbf{t}_{human}$).
ClothNet. ClothNet predicts a 3D cloth template mesh $\mathbf{M}_{cloth}^t$ of the target cloth from the input image. We use ClothWild [38] as ClothNet, which achieves state-of-the-art performance on in-the-wild images.
BodyNet. BodyNet predicts the 3D body template mesh $\mathbf{M}_{body}^t$, by estimating the 3D human pose and shape of the SMPL + H model [36] from the input image. We modify PIXIE [12] to infer SMPL+H and use it as BodyNet.
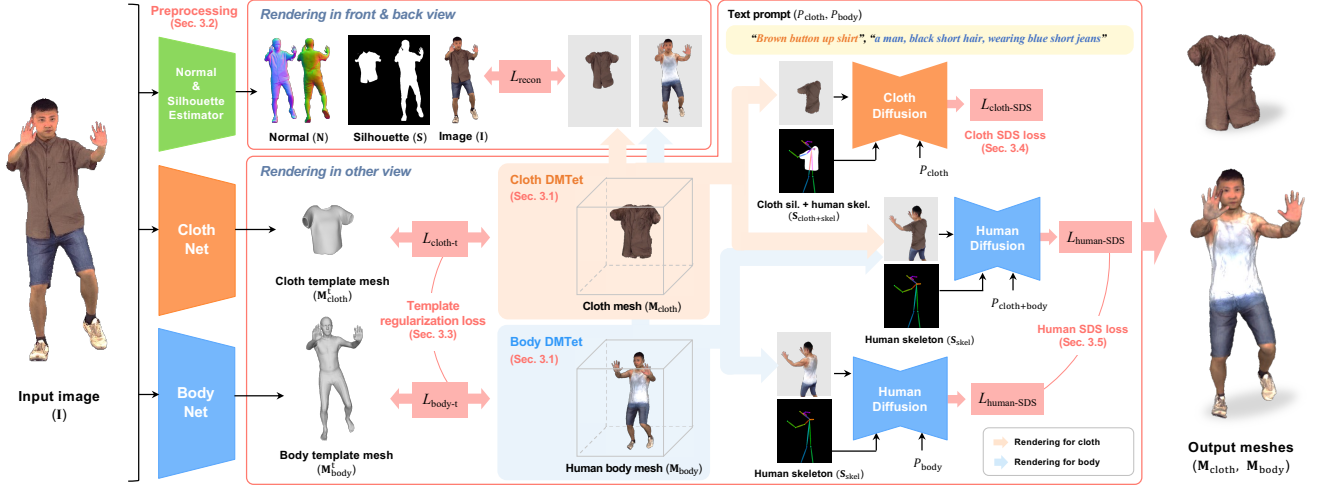
Figure 3. **Overall pipeline of DeClotH.** Given an input image $\mathbf{I}$, DeClotH optimizes 3D cloth and human body, represented by DMTets (Sec. 3.1). For the optimization, we extract normal map $\mathbf{N}$, silhouette $\mathbf{S}$, and 3D template meshes ($\mathbf{M}_{\text{cloth}}^{\text{t}}$ and $\mathbf{M}_{\text{body}}^{\text{t}}$) (Sec. 3.2). Subsequently, the 3D cloth and human body are optimized by three core loss functions: template regularization loss (Sec. 3.3), cloth SDS loss (Sec. 3.4), and human SDS loss (Sec. 3.5).

3.3. Template regularization loss

The template regularization losses, $L_{\text{cloth-t}}$ and $L_{\text{body-t}}$, enforce optimized 3D cloth and human body meshes to be close to their 3D template meshes ($\mathbf{M}_{\text{cloth}}^{\text{t}}$ and $\mathbf{M}_{\text{body}}^{\text{t}}$). The loss functions are defined as

$$L_{\text{cloth-t}} = \|\mathcal{R}_{\text{sil}}(\mathbf{M}_{\text{cloth}}, \mathbf{k}) - \mathcal{R}_{\text{sil}}(\mathbf{M}_{\text{cloth}}^{\text{t}}, \mathbf{k})\|_2, \quad (1)$$

$$L_{\text{body-t}} = \|\mathcal{R}_{\text{sil}}(\mathbf{M}_{\text{body}}, \mathbf{k}) - \mathcal{R}_{\text{sil}}(\mathbf{M}_{\text{body}}^{\text{t}}, \mathbf{k})\|_2, \quad (2)$$

where $\mathcal{R}_{\text{sil}}$ is a silhouette renderer, and $\mathbf{k}$ is a camera parameter for the rendering. These losses constrain the projected 3D meshes to be close to the projection of their 3D template meshes. ClothNet and BodyNet, which produce the 3D template meshes, are trained on in-the-wild datasets containing various images with occlusions. The 3D template meshes exhibit high robustness against occlusions, leveraging learned data-driven knowledge from the in-the-wild datasets. Consequently, the template regularization loss prevents the erroneous reconstruction of 3D cloth and human body from occlusion.

3.4. Cloth SDS loss

Cloth SDS loss $L_{\text{cloth-SDS}}$ supervises the geometry and texture of 3D cloth, particularly for regions that are not visible in the front view. Our cloth SDS loss follows the basic formulation of SDS loss proposed by DreamFusion [45]. Given a 3D mesh parameterized with ϕ, the SDS loss updates ϕ based on the rendered image $\mathbf{x}$ of the 3D mesh, image condition $\mathbf{c}$, and text prompt P for the diffusion model. The gradient of the SDS loss is calculated as

$$\nabla_\phi L_{\text{SDS}}(\mathbf{x}, \mathbf{c}, P) = \mathbb{E}[w_t(\hat{\epsilon}(\mathbf{x}_t; \mathbf{c}, P, t) - \epsilon)\frac{\partial \mathbf{x}_t}{\partial \phi}], \quad (3)$$

where t denotes a noise level, $\mathbf{x}_t$ is a rendered image with noise, and w_t is a weighting variable dependent on the noise level t. This loss function computes the distance between the predicted noise $\hat{\epsilon}(\cdot)$ and the sampled noise ϵ from the diffusion model. Accordingly, the SDS loss guides the 3D mesh rendering to follow a visually coherent appearance that aligns with the image condition and text prompt.

Unlike the basic SDS loss, the cloth SDS loss function exploits a new diffusion model, ClothDiffusion, instead of StableDiffusion [47]. ClothDiffusion is trained to generate cloth-specific images using cloth silhouettes and human skeletons as conditioning inputs. Cloth SDS loss supervises the rendered normal maps and RGB images as follows:

$$L_{\text{cloth-SDS}}^{\text{norm}} = L_{\text{SDS}}(\mathbf{N}_{\text{cloth}}^{\mathbf{k}}, \mathbf{S}_{\text{cloth+skel}}^{\mathbf{k}}, P_{\text{cloth}}), \quad (4)$$

$$L_{\text{cloth-SDS}}^{\text{rgb}} = L_{\text{SDS}}(\mathbf{I}_{\text{cloth}}^{\mathbf{k}}, \mathbf{S}_{\text{cloth+skel}}^{\mathbf{k}}, P_{\text{cloth}}), \quad (5)$$

where $\mathbf{N}_{\text{cloth}}^{\mathbf{k}}$ and $\mathbf{I}_{\text{cloth}}^{\mathbf{k}}$ are the normal map and the RGB image rendered from the 3D cloth mesh $\mathbf{M}_{\text{cloth}}$ with a camera parameter $\mathbf{k}$. The $\mathbf{S}_{\text{cloth+body}}^{\mathbf{k}}$ is a combination of two silhouettes of 3D cloth template mesh $\mathbf{M}_{\text{cloth}}^{\text{t}}$ and 3D human skeleton extracted from the 3D body template mesh $\mathbf{M}_{\text{body}}^{\text{t}}$. The cloth SDS loss provides rich contextual information on cloth appearances using cloth-specific prior knowledge of ClothDiffusion. Additionally, cloth SDS loss can accurately guide the reconstruction of cloth regions by utilizing cloth silhouettes and human skeletons as regional information.

3.5. Human SDS loss

Human loss of SDS $L_{\text{human-SDS}}$ supervises the geometry and texture of the occluded parts of both the 3D cloth and the human body. This loss is based on HumanDiffusion, an-

other diffusion model that generates human images conditioned on human skeletons. We adopt pre-trained weights of ControlNet [60] for the HumanDiffusion. Human SDS loss function supervises the rendered normal map and RGB image as follows:

$$L_{\text{human-SDS}}^{\text{norm}} = L_{\text{SDS}}(\mathbf{N}_{\text{body}}^{\mathbf{k}}, \mathbf{S}_{\text{skel}}^{\mathbf{k}}, P_{\text{body}}) \\ + L_{\text{SDS}}(\mathbf{N}_{\text{cloth+body}}^{\mathbf{k}}, \mathbf{S}_{\text{skel}}^{\mathbf{k}}, P_{\text{cloth+body}}), \quad (6)$$

$$L_{\text{human-SDS}}^{\text{rgb}} = L_{\text{SDS}}(\mathbf{I}_{\text{body}}^{\mathbf{k}}, \mathbf{S}_{\text{skel}}^{\mathbf{k}}, P_{\text{body}}) \\ + L_{\text{SDS}}(\mathbf{N}_{\text{cloth+body}}^{\mathbf{k}}, \mathbf{S}_{\text{skel}}^{\mathbf{k}}, P_{\text{cloth+body}}), \quad (7)$$

where $\mathbf{N}_{\text{body}}^{\mathbf{k}}$ and $\mathbf{I}_{\text{body}}^{\mathbf{k}}$ are the normal map and the RGB image rendered from the 3D human body mesh $\mathbf{M}_{\text{body}}$. Further, $\mathbf{N}_{\text{cloth+body}}^{\mathbf{k}}$ and $\mathbf{I}_{\text{cloth+body}}^{\mathbf{k}}$ are rendered from composition of the 3D cloth and human body meshes ($\mathbf{M}_{\text{cloth}}$ and $\mathbf{M}_{\text{body}}$). $\mathbf{S}_{\text{skel}}^{\mathbf{k}}$ is the rendered image from 3D human skeleton extracted from the 3D body template mesh $\mathbf{M}_{\text{body}}^{\text{t}}$. By incorporating human skeletons, the human SDS loss provides accurate guidance for human regions.

3.6. Optimization procedure

Based on the above loss functions, we optimize the 3D cloth mesh $\mathbf{M}_{\text{cloth}}$ and the 3D human body mesh $\mathbf{M}_{\text{body}}$ through two stages: geometry stage and texture stage.

Geometry stage. In the geometry stage, we optimize the DMTets of 3D cloth and human body by minimizing loss function L_{geo}, defined as follows:

$$L_{\text{geo}} = L_{\text{cloth-t}} + L_{\text{body-t}} + L_{\text{cloth-SDS}}^{\text{norm}} + L_{\text{human-SDS}}^{\text{norm}} + L_{\text{recon}}^{\text{geo}}. \quad (8)$$

$L_{\text{recon}}^{\text{geo}}$ is defined as

$$L_{\text{recon}}^{\text{geo}} = L_{\text{normal}} + L_{\text{sil}} + L_{\text{pen}}. \quad (9)$$

L_{normal} is the L2 distance between rendered normal maps and their optimization targets $\mathbf{N}$ in the front and back view. L_{sil} is the L2 distance between the rendered silhouettes and their optimization targets $\mathbf{S}$ in the front and back view. L_{normal} and L_{sil} are calculated for both the cloth and the human body. L_{pen} penalizes intersection between the 3D cloth and human body meshes.

Texture stage. In the texture stage, we optimize the texture of the meshes ($\mathbf{M}_{\text{cloth}}$ and $\mathbf{M}_{\text{body}}$) obtained from the geometry stage. To this end, we construct MLP networks (Ψ_{cloth} and Ψ_{human}) that predict RGB color given the vertex coordinate of the mesh. These MLP networks are trained by minimizing the loss function L_{tex}, defined as:

$$L_{\text{tex}} = L_{\text{cloth-SDS}}^{\text{rgb}} + L_{\text{human-SDS}}^{\text{rgb}} + L_{\text{recon}}^{\text{tex}}. \quad (10)$$

$L_{\text{recon}}^{\text{tex}}$ is a combination of L2 and LPIPS [61] distances between rendered RGB images and their optimization target (*i.e.*, $\mathbf{I}$) in the front and back view. $L_{\text{recon}}^{\text{tex}}$ is calculated for both cloth and human body.

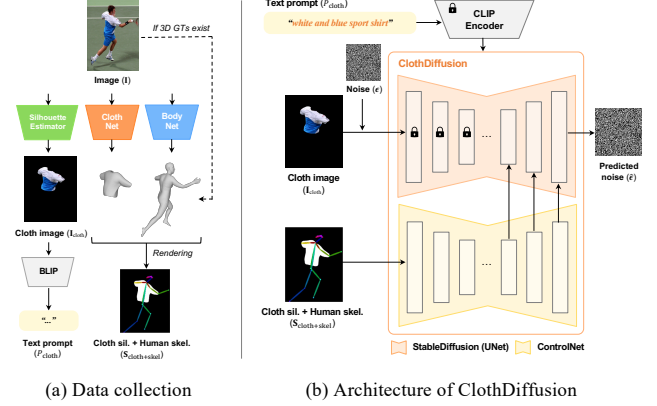(a) Data collection   (b) Architecture of ClothDiffusion

Figure 4. **Training process of ClothDiffusion.** We train the ClothDiffusion based on our collected cloth-specific training data. The ClothDiffusion follows ControlNet architecture with the pretrained StableDiffusion.

4. ClothDiffusion

Fig. 4 shows the training process of ClothDiffusion, which is used in the cloth SDS loss.

4.1. Training data collection

To train ClothDiffusion, we collect three data types: 1) cloth images, 2) conditional images for generation, and 3) text prompts. The cloth images are obtained by running the silhouette estimator [26] from images of the training datasets. The conditional images are acquired by projecting a 3D cloth template mesh and a human skeleton extracted from a 3D body mesh. The 3D cloth and human templates are derived from an image using ClothNet and BodyNet unless 3D ground-truths (GTs) are available in the training datasets. The text prompts are acquired from BLIP [31], a state-of-the-art image captioning method. The acquired text prompt provides detailed descriptions of cloth color, shape, and style in the cloth image.

4.2. Learning cloth generation

ClothDiffusion consists of two networks: StableDiffusion [47] and ControlNet [60]. StableDiffusion estimates noise, given a latent embedding of a cloth image with sampled noise. ControlNet takes a conditional image (*i.e.*, cloth silhouette and human skeleton) and gives additional information for StableDiffusion to generate a plausible image based on the conditional image. Using the pre-trained weights of StableDiffusion, we train ControlNet and partial layers of StableDiffusion while keeping its other layers frozen, following the fine-tuning strategy of ControlNet. This fine-tuning strategy allows ClothDiffusion to generate realistic cloth images by leveraging the strong prior knowledge embedded in the pre-trained StableDiffusion.

5. Experiments

5.1. Datasets

4D-DRESS. 4D-DRESS [53] contains high-quality 3D scans of 64 clothing sequences, with diverse human poses. This dataset is used solely for evaluation purposes. For each clothing sequence in 4D-DRESS, we randomly sample one human pose and render the corresponding 3D scan from pre-defined camera viewpoints to obtain the test images. The evaluation set includes 64 test images, and 124 clothes appear in the test images.

THuman2.0. THuman2.0 [57] contains 3D scans of 525 clothed humans. From all scans, we uniformly sample 52 scans for evaluation. As THuman2.0 does not contain GTs of 3D cloth, we generate 3D cloth pseudo-GTs by running GALA [24] on the 3D human scans. We obtain test images by rendering the 3D scans with pre-defined camera viewpoints. The evaluation set includes 52 test images, and 104 clothes appear in the test images.

Training datasets of ClothDiffusion. DeepFashion [35], SHHQ [13], MSCOCO [34], and THuman2.0 [57] are used to train ClothDiffusion. DeepFashion [35] and SHHQ [13] are large-scale datasets that contain diverse images of clothed humans. We use the official training set from Deep-Fashion and uniformly sample 10,000 images from SHHQ. MSCOCO [34] contains a wide variety of human poses, and we utilize its official training set. THuman2.0 [57] provides multi-view images of clothed people, excluding duplicates from the evaluation set.

5.2. Evaluation metrics

3D geometry reconstruction. We evaluate the geometry of 3D reconstructions by measuring CD (chamfer distance) and NC (normal consistency), following Huang *et al*. [20]. Specifically, we apply Procrustes alignment on the reconstructed 3D meshes based on the SMPL-X meshes of 3D GTs. With the aligned 3D meshes, we measure CD and NC between reconstructed and GT meshes. To measure NC, we calculate the average L2 distance between normal images of reconstructed and GT meshes, rendered at $\{0°, 90°, 180°, 270°\}$ from fixed viewpoints.

3D texture reconstruction. We evaluate the texture of 3D reconstructions, using PSNR (peak signal-to-noise ratio) and LPIPS [61]. (learned perceptual image patch similarity). Before evaluation, we align the reconstructions with 3D human GT poses to eliminate the influence of geometric errors. Then, we calculate PSNR and LPIPS between RGB images of reconstructed and GT meshes rendered at $\{0°, 90°, 180°, 270°\}$ as in the NC measurement.

2D image generation of diffusion model. To verify the feasibility of our proposed ClothDiffusion, we evaluate the generated images from the diffusion model via CLIP-Score [16] and Cloth-IoU. CLIP-Score [16] measures the
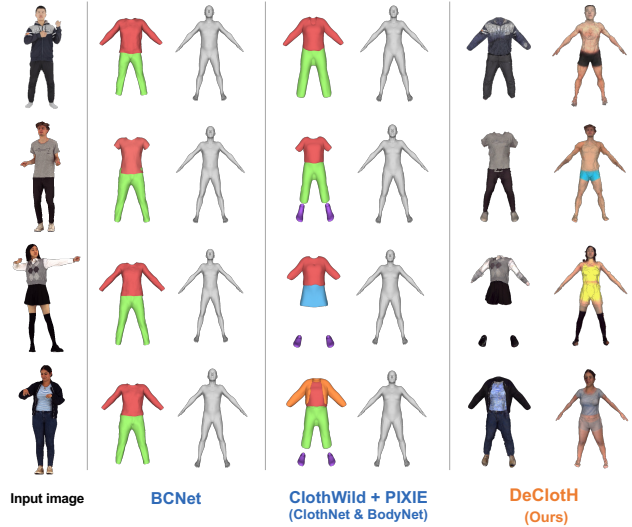

Figure 5. **Effects of the optimization process of DeClotH.**

Methods	4D-DRESS (cloth)		4D-DRESS (cloth+human)	
	CD$\downarrow$	NC$\downarrow$	CD$\downarrow$	NC$\downarrow$
BCNet [22]	4.387	0.046	3.925	0.090
SMPLicit [8]	4.080	0.038	3.605	0.091
ClothWild [38] + PIXIE [12] (ClothNet & BodyNet)	4.100	0.038	3.526	0.087
DeClotH (Ours)	**3.902**	**0.037**	**3.292**	**0.079**

Table 1. **Effectiveness of the DeClotH's optimization process compared to 3D template meshes of ClothNet and BodyNet.**

correlation between cloth text prompts and generated images. Cloth-IoU measures a proportion of the intersection between generated cloth images and GT counterparts. Specifically, we extract cloth silhouettes from the generated images by running SAM [26] with cloth text prompts. Then, we calculate the IoU between the extracted silhouettes and GT counterparts. The cloth text prompts for evaluation are obtained by running BLIP [31] from the test images.

5.3. Ablation study

We carry out the ablation study on 4D-DRESS [53]. As 4D-DRESS does not contain 3D human body scans excluding clothes, we compare methods in two tracks: evaluating 3D cloth (cloth) and evaluating the composition of 3D cloth and the human body (cloth + human).

Effectiveness of template-based optimization. Fig. 5 and Tab. 1 show that our optimization framework produces significantly more realistic 3D reconstruction results, compared to the 3D template meshes from ClothNet and BodyNet. ClothNet and BodyNet are model-based methods that estimate the shape of pre-defined 3D template models of cloth and human body. Thereby, their 3D template meshes cannot deviate from the topology of the template models. Our framework, DeClotH builds upon these
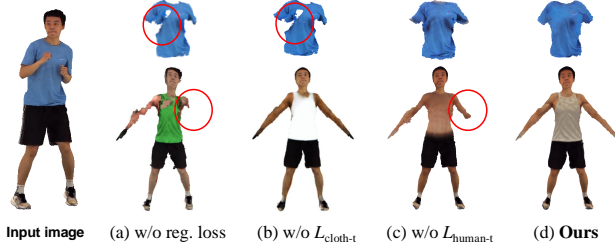
Figure 6. **Effects of template regularization loss.**

Figure 7. **Effects of cloth and human SDS loss.**

		4D-DRESS (cloth)				4D-DRESS (cloth + human)			
$L_{\text{cloth-t}}$	$L_{\text{body-t}}$	CD$\downarrow$	NC$\downarrow$	PSNR$\uparrow$	LPIPS$\downarrow$	CD$\downarrow$	NC$\downarrow$	PSNR$\uparrow$	LPIPS$\downarrow$
✗	✗	8.245	0.047	28.959	0.051	3.880	0.093	22.992	0.071
✗	✓	8.386	0.047	29.028	0.055	3.567	0.084	23.149	0.070
✓	✗	4.078	0.038	30.828	0.038	3.586	0.085	23.422	0.067
✓	✓	**3.902**	**0.037**	**31.582**	**0.033**	**3.292**	**0.079**	**23.921**	**0.065**

Table 2. **Ablation studies for template regularization loss.**

		4D-DRESS (cloth)				4D-DRESS (cloth + human)			
$L_{\text{cloth-SDS}}$	$L_{\text{human-SDS}}$	CD$\downarrow$	NC$\downarrow$	PSNR$\uparrow$	LPIPS$\downarrow$	CD$\downarrow$	NC$\downarrow$	PSNR$\uparrow$	LPIPS$\downarrow$
✗	✗	4.704	0.041	28.151	0.047	3.635	0.086	22.895	0.083
$\rightarrow L_{\text{SD-SDS}}$	✓	4.922	0.040	30.310	0.041	3.711	0.085	23.267	0.068
✓	$\rightarrow L_{\text{SD-SDS}}$	3.955	**0.037**	31.295	0.034	3.601	0.085	23.389	0.067
✓	✓	**3.902**	**0.037**	**31.582**	**0.033**	**3.292**	**0.079**	**23.921**	**0.065**

Table 3. **Ablation studies for cloth and human SDS loss.**

	4D-DRESS (cloth)	
Networks	CLIP-Score$\uparrow$	Cloth-IoU$\uparrow$
StableDiffusion [47]	0.696	0.112
HumanDiffusion [60]	0.683	0.454
ClothDiffusion (Ours)	**0.713**	**0.548**

Table 4. **Ablation studies for ClothDiffusion network.**

3D template meshes by optimizing the 3D cloth and human body with the optimization targets (*e.g.*, normal maps and silhouettes), which have rich geometric information in the input image. Through the optimization process, our framework can reconstruct fine geometric details, such as cloth wrinkles, while keeping the overall shape of 3D templates. Furthermore, our framework reconstructs the texture of cloth and human body along with the geometry, resulting in more lifelike reconstructions.

Effectiveness of template regularization loss. Fig. 6 and Tab. 2 show that the template regularization loss significantly drops the 3D reconstruction error, especially when the cloth and human body are occluded. Without the template regularization loss, the optimization of 3D cloth and human body highly relies on the optimization targets (*e.g.*, cloth silhouette) of the front views. Our proposed template regularization loss reduces the heavy reliance, by using the 3D template models as helpful supervision for the occluded parts. Thus, the template regularization loss alleviates the occlusion issue and enhances the 3D reconstruction in both the cloth and the human body.

Effectiveness of cloth & human SDS loss. Fig. 7 and Tab. 3 show that the cloth and human SDS losses are much more effective in reconstructing geometry and texture than vanilla SDS loss (*i.e.*, $L_{\text{SD-SDS}}$) that uses StableDiffusion for reconstruction guidance. As shown in Fig. 7 (c), the reconstructed 3D jacket has an artifact near the boundary between cloth and human body, when using the vanilla SDS loss. This artifact indicates that StableDiffusion has inadequate prior knowledge to separately reconstruct 3D cloth from the human body. On the other hand, the cloth SDS loss using ClothDiffusion effectively supervises the 3D geometry to follow the desired cloth shape.

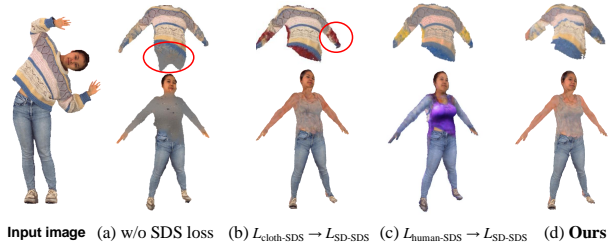**Ablation on ClothDiffusion network.** Tab. 4 shows our

proposed ClothDiffusion is superior to other diffusion models in generating cloth images. StableDiffusion [47] and HumanDiffusion [60] commonly produce images that contain not only cloth but also other contents (*e.g.*, human body and background scene). Thus, StableDiffusion and HumanDiffusion have a low CLIP-Score [16] on text prompts that describe clothes only. On the other hand, ClothDiffusion has a high CLIP-Score since it is specialized to generate cloth images without including other contents. Additionally, ClothDiffusion achieves the highest Cloth-IoU among the diffusion models, since ClothDiffusion takes a cloth silhouette with a human skeleton as a condition, accurately guiding the region where the cloth will be generated. Such superiority of ClothDiffusion in generating cloth images is beneficial in accurate guidance for reconstructing 3D cloth.

5.4. Comparison with state-of-the-art methods

We compare our method with recent 3D cloth decomposition and 3D clothed human reconstruction methods: GALA* [24], SiTH [17] + GALA [24], and TeCH [20] + GALA [24]. GALA* is a modified version of the original GALA to take a single image as input instead of a 3D scan. Specifically, GALA* only applies loss functions corresponding to the front view, ignoring other view directions. SiTH + GALA and TeCH + GALA are two-stage reconstruction methods that first reconstruct a 3D clothed human, as one unified mesh, and decompose the 3D cloth
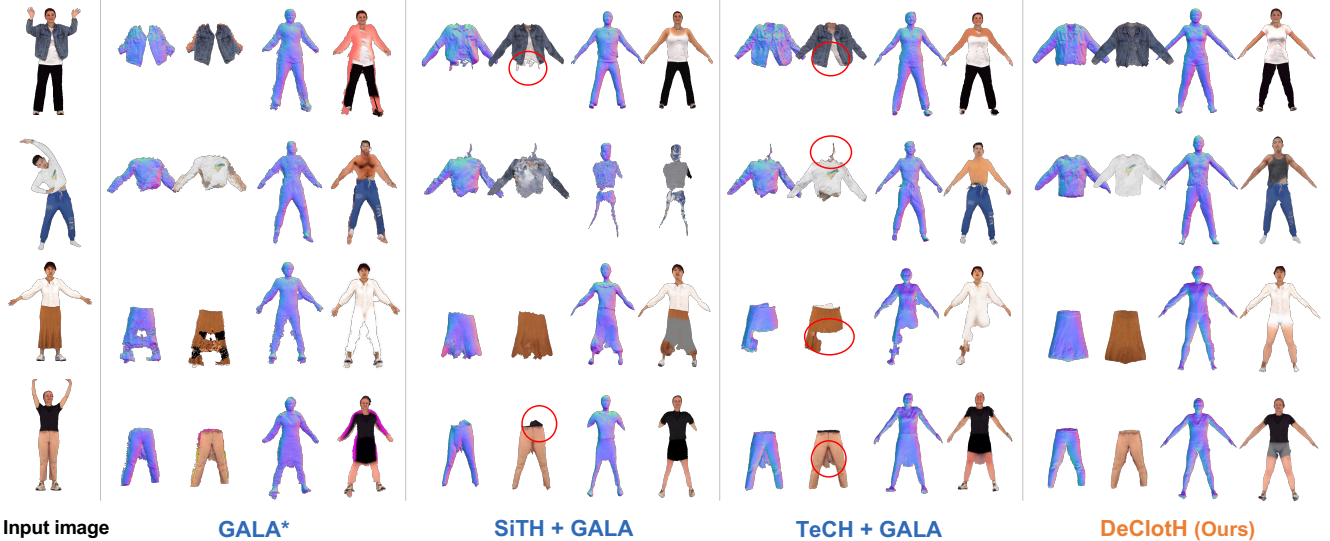
Figure 8. **Qualitative comparison with 3D cloth decomposition and 3D clothed human reconstruction methods: GALA* [24], SiTH [17]+GALA [24], and TeCH [20]+GALA [24], on 4D-DRESS [53] and THuman2.0 [57].** * denotes the algorithm is modified to take a single image as input instead of a 3D scan. We highlight their representative failure cases with red circles.

Methods	4D-DRESS (cloth)				4D-DRESS (cloth + human)				THuman2.0 (cloth)				THuman2.0 (cloth + human)			
	CD↓	NC↓	PSNR↑	LPIPS↓	CD↓	NC↓	PSNR↑	LPIPS↓	CD↓	NC↓	PSNR↑	LPIPS↓	CD↓	NC↓	PSNR↑	LPIPS↓
GALA* [24]	5.790	0.049	25.363	0.068	3.451	0.102	19.490	0.117	2.211	0.059	26.553	0.050	2.132	0.118	22.183	0.079
SiTH [17] + GALA [24]	6.980	0.046	27.725	0.064	3.737	0.087	22.009	0.102	2.486	0.062	28.615	0.056	1.828	0.094	25.009	0.060
TeCH [20] + GALA [24]	5.043	0.039	29.123	0.045	3.334	0.083	23.140	0.076	2.112	0.051	29.584	0.044	1.900	0.091	**25.618**	0.048
DeClotH (Ours)	**3.902**	**0.037**	**31.582**	**0.033**	**3.292**	**0.079**	**23.921**	**0.065**	**1.756**	**0.044**	**30.612**	**0.032**	**1.812**	**0.089**	25.421	**0.046**

Table 5. **Quantitative comparison with existing 3D cloth decomposition and 3D clothed human reconstruction methods.**

mesh from the reconstructed 3D clothed human mesh.

Fig. 8 and Tab. 5 show the superior performance of our DeClotH compared to the prior arts on 4D-DRESS [53] and THuman2.0 [57]. GALA* [24] highly relies on the front view, which results in undesirable appearances, especially when the cloth is occluded. On the other hand, DeClotH produces much more robust results from occlusion by using 3D template models of cloth and human body as strong geometric priors for reconstruction. The two-stage reconstruction methods, SiTH [17] + GALA [24] and TeCH [20] + GALA [24], also suffer from undesirable artifacts of reconstruction results, such as torn clothes. The reason for such artifacts is that the 3D geometric error of the first stage, 3D clothed human reconstruction, fatally propagates to the second stage, 3D cloth decomposition. The 3D clothed human reconstruction contains inevitable 3D geometric errors because of the ill-posedness of reconstruction. The error of the first stage would be a bad source for the 3D cloth decomposition method, resulting in a failure of decomposition. Compared to the two-stage reconstruction methods, DeClotH is a one-stage reconstruction method that is free from the above issue. Additionally,

while the two-stage methods do not consider the geometries of the cloth and human body during the reconstruction, DeClotH effectively guides the geometries by leveraging 3D template models and ClothDiffusion. We provide further discussion in the supplementary material.

6. Conclusion

We introduce DeClotH, a novel and powerful framework that reconstructs decomposable 3D cloth and human body from a single image. Based on 3D template models of cloth and human body, our proposed template regularization loss and cloth diffusion model effectively infer geometry and texture in the invisible regions of 3D cloth and human body. As a result, our framework significantly outperforms baseline methods, qualitatively and quantitatively.

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
- S2. Evaluation of pose deformation
- S3. Evaluation with POR Score
- S4. Implementation details
- S5. Discussion of two-stage reconstruction
- S6. Limitations and future works
- S7. More qualitative results

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.

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

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

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.**

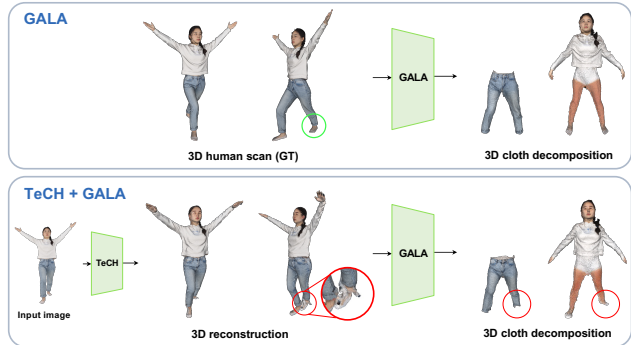| Methods	POR Score↓
GALA* [24]	0.418
SiTH [17] + GALA [24]	0.246
TeCH [20] + GALA [24]	0.225
DeClotH (Ours)	**0.218**

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

(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.

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.

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 re-

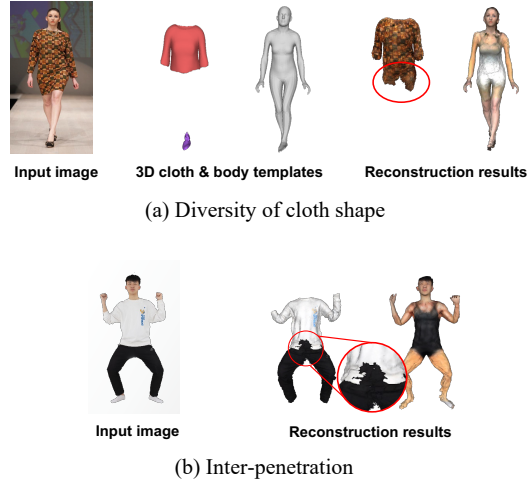(a) Diversity of cloth shape

(b) Inter-penetration

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

construction based on the cloth templates also produces erroneous results. Improving the expression power of cloth template models should be a future research direction.

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.
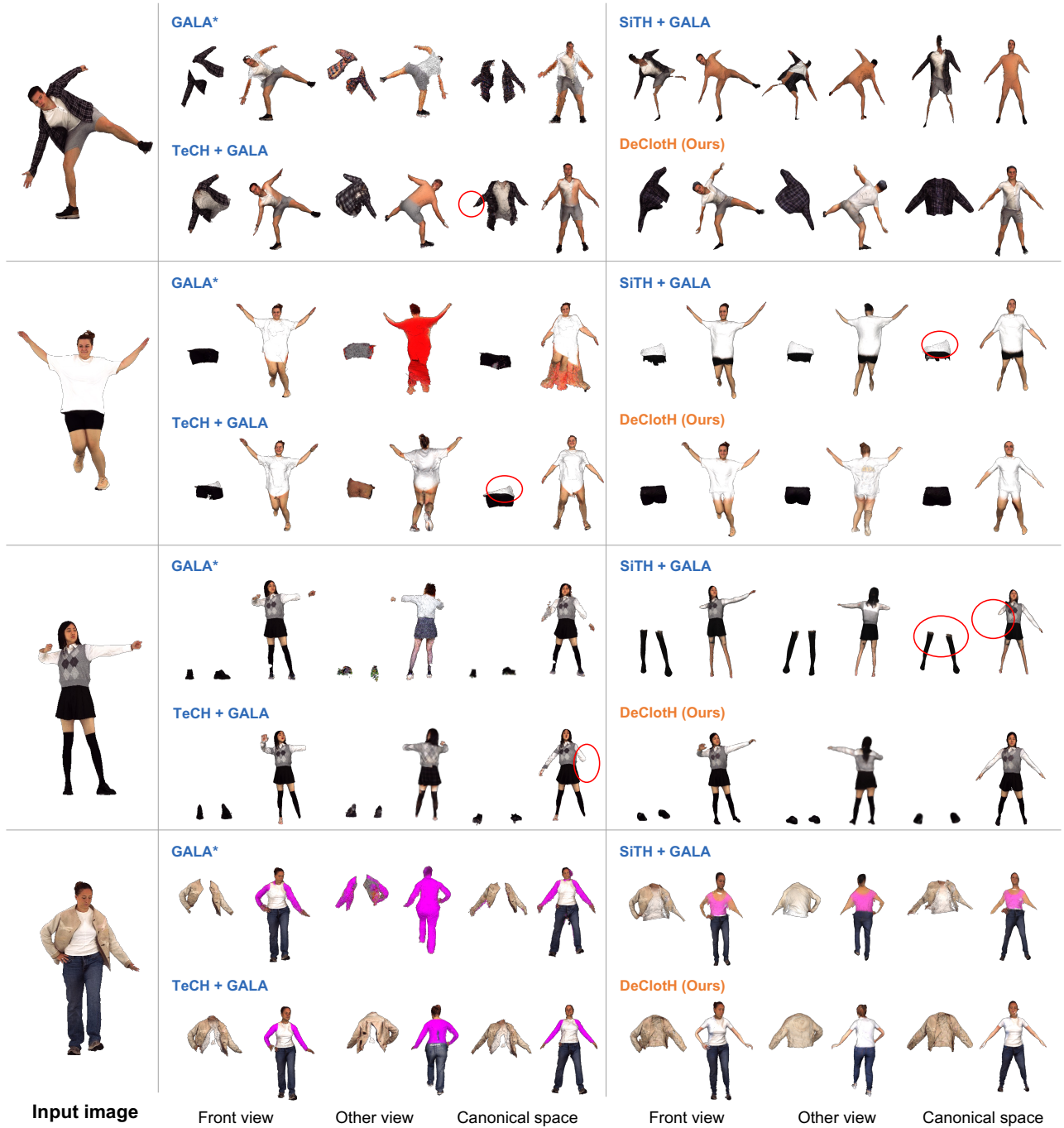
Figure S4. **Qualitative comparison with 3D cloth decomposition and 3D clothed human reconstruction methods: GALA**[*] **[24], SiTH [17]+GALA [24], and TeCH [20]+GALA [24], on 4D-DRESS [53].** [*] denotes the algorithm is modified to take a single image as input instead of a 3D scan. We highlight their 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.
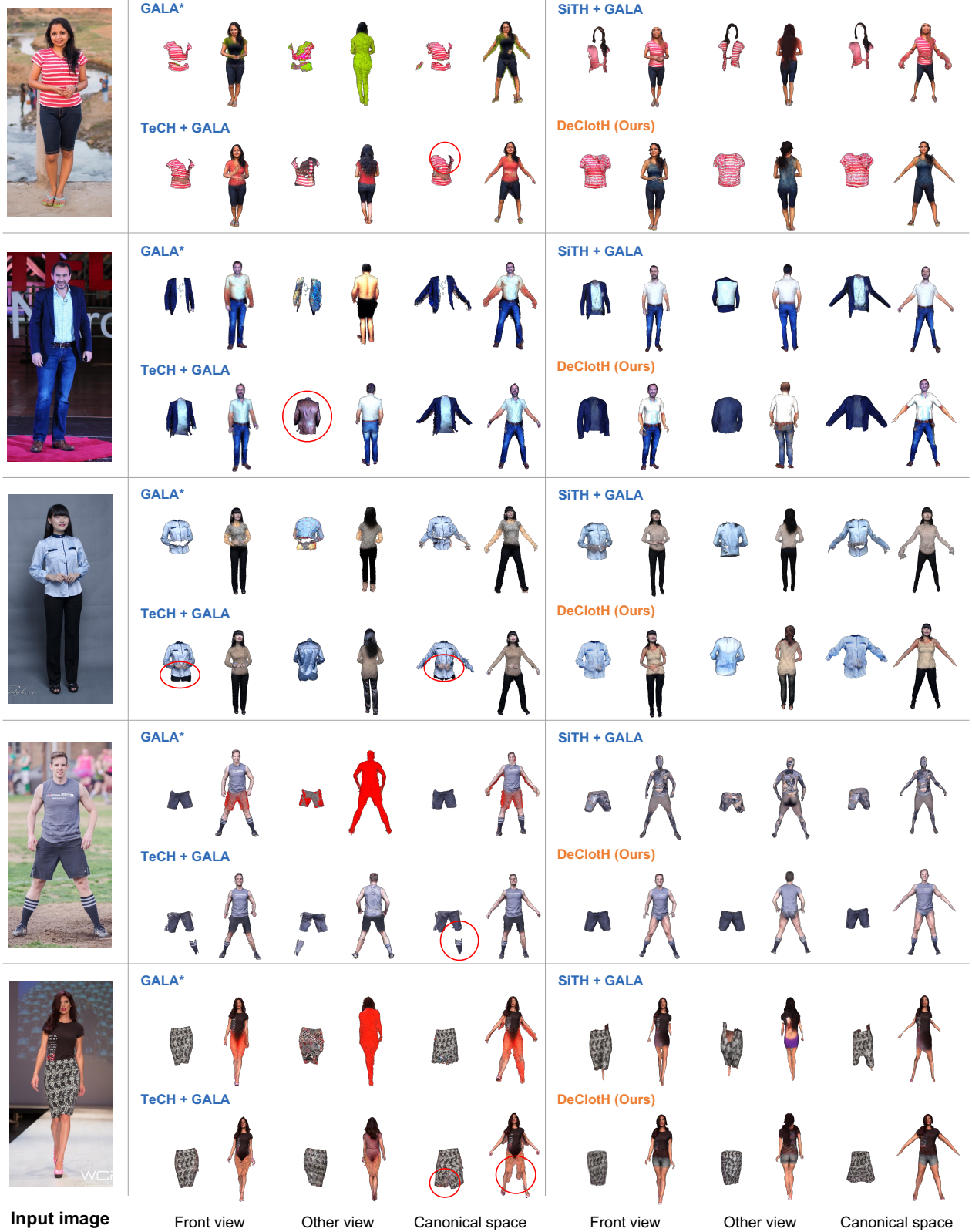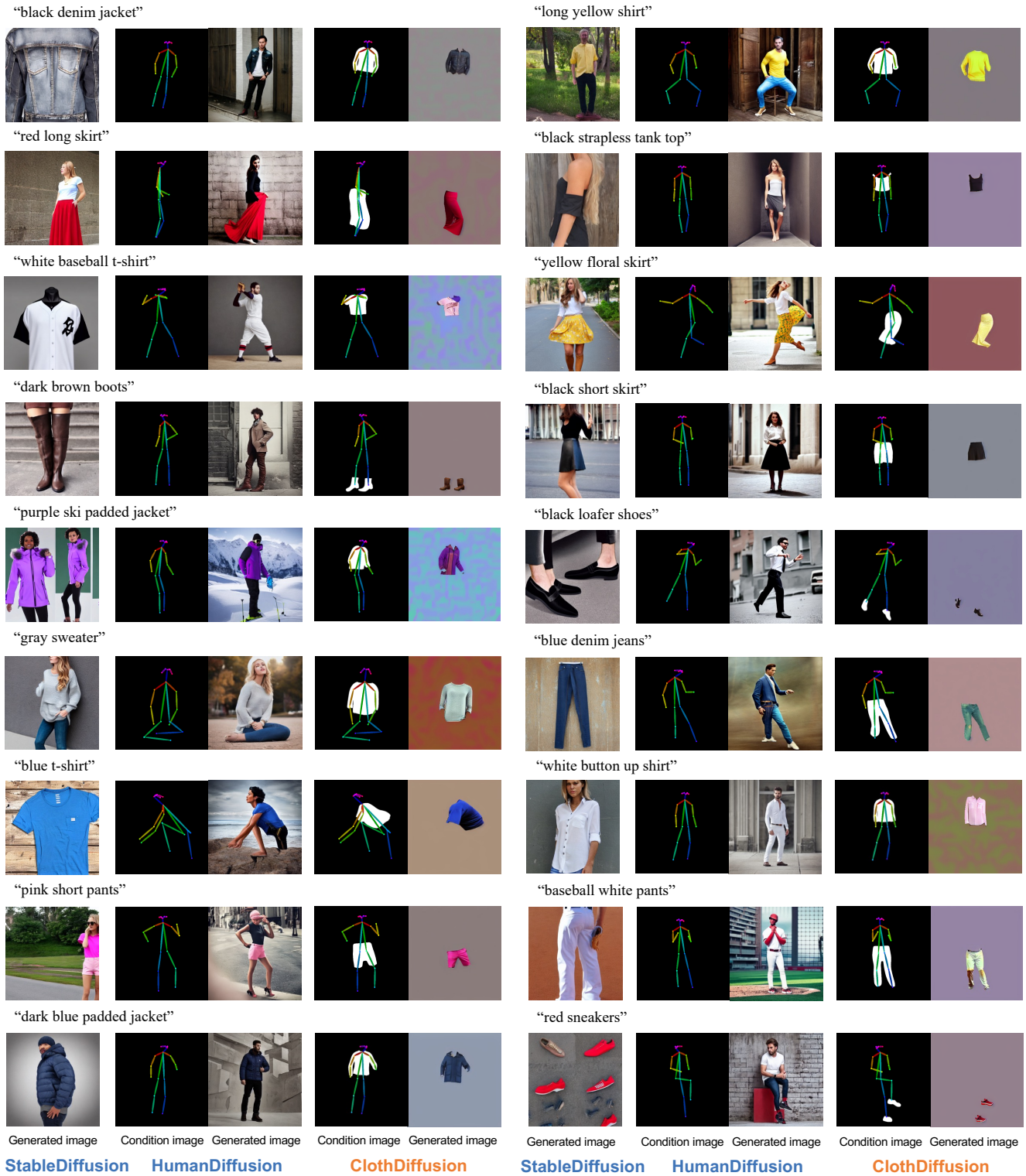
"black denim jacket" "long yellow shirt"

"red long skirt" "black strapless tank top"

"white baseball t-shirt" "yellow floral skirt"

"dark brown boots" "black short skirt"

"purple ski padded jacket" "black loafer shoes"

"gray sweater" "blue denim jeans"

"blue t-shirt" "white button up shirt"

"pink short pants" "baseball white pants"

"dark blue padded jacket" "red sneakers"

Generated image   Condition image   Generated image   Condition image   Generated image
StableDiffusion   **HumanDiffusion**   **ClothDiffusion**

Generated image   Condition image   Generated image   Condition image   Generated image
StableDiffusion   **HumanDiffusion**   **ClothDiffusion**

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

References

[1] Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. Single-image 3D human digitization with shape-guided diffusion. In *SIGGRAPH Asia*, 2023. 1, 2, 3

[2] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3D reconstruction of humans wearing clothing. In *CVPR*, 2022. 2

[3] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3D reconstruction of humans wearing clothing. In *CVPR*, 2022. 2

[4] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-Garment Net: Learning to dress 3D people from images. In *ICCV*, 2019. 3

[5] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3D human pose and shape from a video. In *CVPR*, 2021. 3

[6] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3D human mesh from in-the-wild crowded scenes. In *CVPR*, 2022.

[7] Hongsuk Choi, Hyeongjin Nam, Taeryung Lee, Gyeongsik Moon, and Kyoung Mu Lee. Rethinking self-supervised visual representation learning in pre-training for 3D human pose and shape estimation. In *ICLR*, 2023. 3

[8] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. SMPLicit: Topology-aware generative model for clothed people. In *CVPR*, 2021. 2, 3, 6

[9] Enric Corona, Mihai Zanfir, Thiemo Alldieck, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Structured 3D features for reconstructing controllable avatars. In *CVPR*, 2023. 2

[10] Luca De Luigi, Ren Li, Benoit Guillard, Mathieu Salzmann, and Pascal Fua. DrapeNet: Garment generation and self-supervised draping. In *CVPR*, 2023. 3

[11] Akio Doi and Akio Koide. An efficient method of triangulating equi-valued surfaces by using tetrahedral cells. *IEICE TRANSACTIONS on Information and Systems*, 1991. 3

[12] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *3DV*, 2021. 3, 6

[13] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. StyleGAN-Human: A data-centric odyssey of human generation. In *ECCV*, 2022. 6

[14] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3D human digitization from single 2k resolution images. In *CVPR*, 2023. 2

[15] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. ARCH++: Animation-ready clothed human reconstruction revisited. In *ICCV*, 2021. 2

[16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. *EMNLP*, 2021. 6, 7

[17] I Ho, Jie Song, Otmar Hilliges, et al. SiTH: Single-view textured human reconstruction with image-conditioned diffusion. In *CVPR*, 2024. 1, 2, 3, 7, 8, 4, 5, 6

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1

[19] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. SHERF: Generalizable human NeRF from a single image. *ICCV*, 2023. 2

[20] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided reconstruction of lifelike clothed humans. In *3DV*, 2024. 1, 2, 3, 6, 7, 8, 4, 5

[21] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable reconstruction of clothed humans. In *CVPR*, 2020. 2

[22] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. BCNet: Learning body and cloth shape from a single image. In *ECCV*, 2020. 3, 6

[23] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 3

[24] Taeksoo Kim, Byungjun Kim, Shunsuke Saito, and Hanbyul Joo. GALA: Generating animatable layered assets from a single scan. In *CVPR*, 2024. 3, 6, 7, 8, 1, 2, 4, 5

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

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 3, 5, 6, 1

[27] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 3

[28] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021.

[29] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 3

[30] Nikos Kolotouros, Thiemo Alldieck, Enric Corona, Eduard Gabriel Bazavan, and Cristian Sminchisescu. Instant 3D human avatar generation using image diffusion models. In *ECCV*, 2024. 3

[31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 5, 6, 1

[32] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 3

[33] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. High-fidelity clothed avatar reconstruction from a single image. In *CVPR*, 2023. 2

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 6

[35] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 6

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 2015. 2, 3

[37] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3D hand pose estimation for whole-body 3D human mesh estimation. In *CVPRW*, 2022. 3

[38] Gyeongsik Moon, Hyeongjin Nam, Takaaki Shiratori, and Kyoung Mu Lee. 3D clothed human reconstruction in the wild. In *ECCV*, 2022. 3, 6

[39] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 2022. 2

[40] Hyeongjin Nam, Daniel Sungho Jung, Yeonguk Oh, and Kyoung Mu Lee. Cyclic test-time adaptation on monocular video for 3D human mesh reconstruction. In *ICCV*, 2023. 3

[41] Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. HumanSplat: generalizable single-image human gaussian splatting with structure priors. *NeurIPS*, 2024. 2

[42] Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. HumanSplat: Generalizable single-image human gaussian splatting with structure priors. *NeurIPS*, 2025. 2

[43] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017. 2

[44] 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*, 2019. 3

[45] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *ICLR*, 2023. 2, 4

[46] RenderPeople, 2018. https://renderpeople.com/3d-people. 2

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 4, 5, 7

[48] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019. 2

[49] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *CVPR*, 2020. 2

[50] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3D shape synthesis. In *NeurIPS*, 2021. 3

[51] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. SIZER: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing. In *ECCV*, 2020. 3

[52] Junying Wang, Jae Shin Yoon, Tuanfeng Y Wang, Krishna Kumar Singh, and Ulrich Neumann. Complete 3D human reconstruction from a single incomplete image. In *CVPR*, 2023. 2

[53] Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 4D-DRESS: A 4D dataset of real-world human clothing with semantic annotations. In *CVPR*, 2024. 3, 6, 8, 1, 2, 4

[54] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. ICON: Implicit clothed humans obtained from normals. In *CVPR*, 2022. 2

[55] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. ECON: Explicit clothed humans optimized via normal integration. In *CVPR*, 2023. 2, 3

[56] Yuliang Xiu, Yufei Ye, Zhen Liu, Dimitrios Tzionas, and Michael J Black. PuzzleAvatar: Assembling 3D avatars from personal albums. *arXiv preprint arXiv:2405.14869*, 2024. 3

[57] 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 *CVPR*, 2021. 2, 6, 8, 3, 5

[58] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *ICCV*, 2021. 3

[59] Jingbo Zhang, Xiaoyu Li, Qi Zhang, Yanpei Cao, Ying Shan, and Jing Liao. HumanRef: Single image to 3D human generation via reference-guided diffusion. In *CVPR*, 2024. 1, 2

[60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 5, 7, 3

[61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5, 6

[62] Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. Global-correlated 3D-decoupling transformer for clothed avatar reconstruction. *NeurIPS*, 2024. 2

[63] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. DeepHuman: 3D human reconstruction from a single image. In *ICCV*, 2019.

[64] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE TPAMI*, 2021. 2

[65] Heming Zhu, Lingteng Qiu, Yuda Qiu, and Xiaoguang Han. Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images. In *CVPR*, 2022. 3

[66] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3D gaussian avatars. In *3DV*, 2025. 3