TeCH: Text-guided Reconstruction of Lifelike Clothed Humans

Yangyi Huang[1*], Hongwei Yi[2*], Yuliang Xiu[2*], Tingting Liao[3], Jiaxiang Tang[4], Deng Cai[1], Justus Thies[2]

[1]State Key Lab of CAD & CG, Zhejiang University [2]Max Planck Institute for Intelligent Systems
[3]Mohamed bin Zayed University of Artificial Intelligence [4]Peking University

huangyangyi@zju.edu.cn, {hongwei.yi, yuliang.xiu, justus.thies}@tuebingen.mpg.de
tingting.liao@mbzuai.ac.ae, tjx@pku.edu.cn, dengcai@cad.zju.edu.cn

Figure 1. Given a single image, TeCH reconstructs a lifelike 3D clothed human. **"Lifelike"** refers to 1) a detailed full-body geometry, including facial features and clothing wrinkles, in both frontal and unseen regions, and 2) a high-quality texture with consistent color and intricate patterns. The key insight is to guide the reconstruction using a personalized Text-to-Image (T2I) diffusion model and textual information derived via visual questioning answering (VQA). Multi-view supervision is established through Score Distillation Sampling (SDS).

Abstract

Despite recent research advancements in reconstructing clothed humans from a single image, accurately restoring the "unseen regions" with high-level details remains an unsolved challenge that lacks attention. Existing methods often generate overly smooth back-side surfaces with a blurry texture. But how to effectively capture all visual attributes of an individual from a single image, which are sufficient to reconstruct unseen areas (e.g. the back view)? Motivated by the power of foundation models, TeCH reconstructs the 3D human by leveraging 1) descriptive text prompts (e.g. garments, colors, hairstyles) which are automatically generated via a garment parsing model and Visual Question Answering (VQA), 2) a personalized fine-tuned Text-to-Image diffusion model (T2I) which learns the "indescribable" appearance. To represent high-resolution 3D clothed humans at an affordable cost, we propose a hybrid 3D representation based on DMTet, which consists of an explicit body shape grid and an implicit distance field. Guided by the descriptive prompts + personalized T2I diffusion model, the geometry and texture of the 3D humans are optimized through multi-view Score Distillation Sampling (SDS) and reconstruction losses based on the original observation. TeCH produces high-fidelity 3D clothed humans with consistent & delicate texture, and detailed full-body geometry. Quantitative and qualitative experiments demonstrate that TeCH outperforms the state-of-the-art methods in terms of reconstruction accuracy and rendering quality. The code will be publicly available for research purposes at huangyangyi.github.io/TeCH

---

[*]These authors contributed equally to this work.

1. Introduction

High-fidelity 3D digital humans are crucial for various applications in augmented and virtual reality, such as gaming, social media, education, e-commerce, and immersive telepresence. To facilitate the creation of digital humans from easily accessible in-the-wild photos, numerous approaches focus on reconstructing a 3D clothed human shape from a single image [12, 38, 39, 46, 67, 72, 102–104, 119–121, 136]. However, despite the advancements made by previous approaches, this specific problem can be considered ill-posed due to the lack of observations of non-visible areas. Efforts to predict *invisible* regions (*e.g.* back-side) based on *visible* visual cues (*e.g.* colors [5, 46, 103], normal estimates [104, 120, 121]) have proven unsuccessful, resulting in blurry texture and smoothed-out geometry, see Fig. 8. As a result, inconsistencies arise when observing these reconstructions from different angles. To address this issue, introducing multi-view supervision could be a potential solution. But is it feasible given only a single input image? Here, we propose TeCH to answer this question. Unlike prior research that primarily explores the connection between visible frontal cues and non-visible regions, TeCH integrates textual information derived from the input image with a personalized Text-to-Image diffusion model, *i.e.*, DreamBooth [101], to guide the reconstruction process.

Specifically, we divide the information from the single input image into the semantic information that can be accurately described by texts and subject's distinctive and fine-detailed appearance which is not easily describable by text: **1) Describable** semantic prompts, including the detailed descriptions of colors, styles of garments, hairstyles, and facial features, are *explicitly* parsed from the input image using a garment parsing model (*i.e.* SegFormer [117]) and a pre-trained visual-language VQA model (*i.e.* BLIP [65]). **2) Indescribable** appearance information, which *implicitly* specifies the subject's distinctive appearance and fine-grained details, is embedded into a unique token "$[V]$", by a personalized Text-to-Image (T2I) diffusion model [101].

Based on these information sources, we optimize the 3D human using multi-view Score Distillation Sampling (SDS)[94], reconstruction losses based on the original observations, and regularization obtained from off-the-shelf normal estimators, to enhance the fidelity of the reconstructed 3D human models while preserving their original identity. To represent a high resolution geometry at an affordable cost, we propose a hybrid 3D representation based on DMTet [32, 106]. This hybrid 3D representation combines an explicit tetrahedral grid to approximate the overall body shape and implicit Signed Distance Function (SDF) and RGB fields to capture fine details in geometry and texture. In a two-stage optimization process, we first optimize this tetrahedral grid, extract the geometry represented as a mesh, and then optimize the texture.

TeCH enables the reconstruction of high-fidelity 3D clothed humans with detailed full-body geometry, and intricate textures with consistent color and patterns. As a result, it facilitates various downstream applications such as novel view rendering, character animation, and shape & texture editing. Quantitative evaluations performed on 3D clothed human datasets, covering various poses (CAPE [93]) and outfits (THuman2.0 [126]), have demonstrated TeCH's superiority in reconstructing geometric details. Qualitative comparisons conducted on in-the-wild images, accompanied by a perceptual study, further confirm that TeCH surpasses SOTA methods in terms of rendering quality. The code will be publicly avaiable for research purpose at *huangyangyi.github.io/TeCH*

2. Related Work

TeCH **reconstructs** a high-fidelity clothed human from a single image, and **imagine** the missing parts through the aid of descriptive prompts and a personalized diffusion model. We relate TeCH to both image-based human reconstructors (Sec. 2.1) and 3D human generators (Sec. 2.2). Human reconstructors could be grouped as: 1) Explicit-shape-based, 2) Implicit-function-based, and 3) NeRF-based methods. The human generators are categorized w.r.t. their training data: 1) directly learned from 3D real captures or 2) indirectly learned from large-scale 2D images. In addition, there is a line of image-to-3D works focusing on general objects, which will be discussed in Sec. 2.3.

2.1. Image-based Clothed Human Reconstruction

Explicit-shape-based Methods. Human Mesh Recovery (HMR) from a single RGB image is a long-standing problem that has been thoroughly explored. A lot of methods [26, 53, 56–59, 64, 66, 68, 129, 131] use mesh-based parametric body models [51, 78, 92, 123] to regress the shape and pose of minimally-clothed 3d body meshes. To account for the 3D garments, 3D clothing offsets [1–4, 63, 116, 139] or deformable garment templates [9, 49] are used on top of a body model. Also, non-parametric explicit representations, such as depth maps [29, 108], normal maps [121], and point clouds [127] could be leveraged to reconstruct the clothed human. However, explicit shapes often suffer from restricted topological flexibility, particularly, when dealing with outfit variations in real-world scenarios, *e.g.*, dress, skirt, and open jackets.

Implicit-function-based Methods. Implicit representations (occupancy/distance field) are topology-agnostic, thus, can represent 3D clothed humans, with arbitrary topologies, such as open jackets and loose skirts. A line of works regresses the free-form implicit surface in an end-to-end manner [5, 103, 104], leverages a 3D geometric prior [12, 21, 38, 39, 46, 72, 120, 124, 136], or progressively
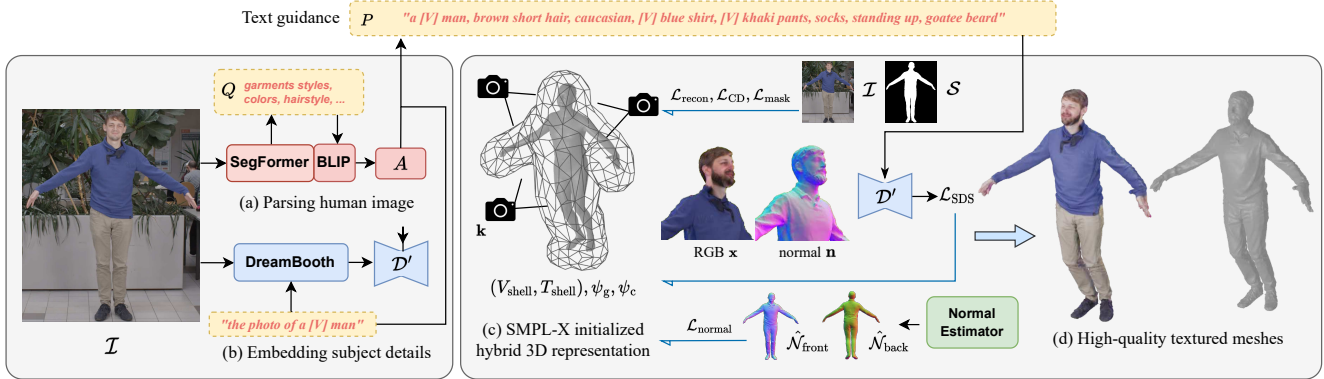
Figure 2. **Method overview.** TeCH takes an image $\mathcal{I}$ of a human as input. Text guidance is constructed through **(a)** using garment parsing model (SegFormer) and VQA model (BLIP) to parse the human attributes A with pre-defined problems Q, and **(b)** embedding with subject-specific appearance into DreamBooth $\mathcal{D}'$ as unique token $[V]$. Next, TeCH represents the 3D clothed human with **(c)** SMPL-X initialized hybrid DMTet, and optimize both geometry and texture using $\mathcal{L}_{\text{SDS}}$ guided by prompt $P = [V] + P_{\text{VQA}}(A)$. During the optimization, $\mathcal{L}_{\text{recon}}$ is introduced to ensure input view consistency, $\mathcal{L}_{\text{CD}}$ is to enforce the color consistency between different views, and $\mathcal{L}_{\text{normal}}$ serves as surface regularizer. Finally, the extracted high-quality textured meshes **(d)** are ready to be used in various downstream applications.

builds up the 3D human using a "sandwich-like" structure and implicit shape completion [121]. Among these works, PIFu [103], ARCH(++) [39, 46], and PaMIR [136] infer the full texture from the input image. PHORHUM [5] and S3F [21] additionally decompose the albedo and global illumination. However, the lack of multi-view supervision often results in depth ambiguities or inconsistent textures.

NeRF-based Methods. There is a separate line of research that focuses on optimizing neural radiance fields (NeRF) from a single image. SHERF [43] and ELICIT [45] optimize a generalized human NeRF, incorporating model-based priors (SMPL-X [92]). While SHERF complements missing information from partial 2D observations, ELICIT leverages appearance prior from CLIP [97].

2.2. Generative Modeling of 3D Clothed Humans

3D Human Generator Trained on 3D Data. Statistical body models [51, 78, 92, 123] can be considered as 3D generative models of the human body. These models are trained on numerous 3D scans of minimally-clothed bodies, and can generate posed bodies with varying shapes, but without clothing. To account for the outfits, CAPE [79] learns a clothing offset layer based on the SMPL-D model, from registered human scans, Chupa [55] "carves" the SMPL mesh by dual normal maps generated by pose-conditioned diffusion model; Alternatively, gDNA [17], NPMs [88], and SPAMs [89], learn the implicit clothed avatars from normalized raw captures (*i.e.*, scans, depth maps). Unfortunately, all the aforementioned methods to learn generative 3D humans with diverse shapes and appearances require 3D data, which is both limited and expensive to acquire. Rodin [113] has recently employed large-scale 3D synthetic head avatars in combination with a diffusion model

to develop a high-fidelity head avatar generator. However, the scarcity of datasets containing real 3D clothed humans [11, 18, 47, 126, 135] limits the model's generalization ability and may lead to overfitting on small datasets.

3D Human Generator from 2D Image Collections. In contrast to 3D data, large-scale 2D human images are widely avaible from DeepFashion [34, 77], SHHQ [28] and LAION-5B [105]. Related human generators represent 3D humans using meshes [36, 40, 50], DMTet [33], Tri-planes [8, 25, 85, 109, 132], implicit functions [118], or neural fields [13, 41, 60, 128]. Some methods adapt GANs [54] by integrating diff-renderer [8, 25, 36, 85, 109, 110, 118, 132], while others leverage diffusion models [13, 40, 44, 60, 130]. Despite the demonstrated quality of these methods in generating textured avatars, a significant gap still exists in achieving "lifelike" avatars with detailed geometry and texture, consistent with the input.

In contrast, TeCH excels at generating "lifelike" 3D characters from a single image, incorporating consistent texture with intricate patterns like checkered or overlapped designs. It relies on a pretrained diffusion model which is trained on a billion-level data, LAION-5B [105], and offers the ability to **imagine the non-visible regions**, guided by descriptive prompts. Furthermore, it leverages the image-based reconstruction approach to faithfully **reconstruct the visible regions** from a single input image.

2.3. Image-to-3D for General Objects

Lifting 2D to 3D for general objects is a longstanding problem with valuable explorations. Here, we mainly focus on diffusion-guided approaches. Initially, CLIP [97] semantic consistency loss [48], Score Jacobian Chaining (SJC) [112] and Score Distillation Sampling (SDS) [94]
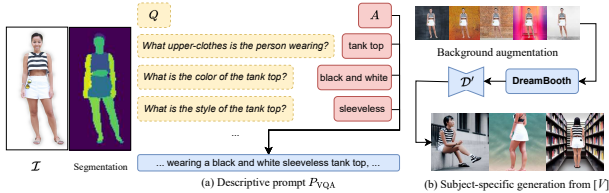
3

Figure 3. **Prompt construction ($P = P_{\mathrm{VQA}} + [V]$).** (a) Inquire VQA model with predefined questions on individual appearance to construct *describable* prompts P_{VQA}. (b) Fine-tuned DreamBooth with background-augmented images to embed *indescribable* subject-specific details into unique identifier $[V]$.

Figure 4. **The effects of text guidance.** We compare the effectiveness of using only VQA descriptions (TeCH$_{\mathrm{vqa}}$), only DreamBooth identity token (TeCH$_{\mathrm{db}}$), and both of them (TeCH).

are proposed to leverage pretrained 2D diffusion models for 3D content generation. Subsequently, there is a line of works [22, 81, 98, 111, 122] that address this problem, by incorporating textural inversion [30], DreamBooth [101], CLIP-guided diffusion prior, depth prior, and reconstruction loss. In addition to aforementioned "reconstruct via multi-view SDS" scheme, recent attention has been drawn to the "reconstruct via direct view-conditional generation" [14, 75, 76, 95, 114, 138]. In contrast, TeCH aims to recover pixel-aligned models with intricate texture, even in non-visible regions, which is a challenging scenario where existing solutions have not shown promising results.

3. Method

Given a single image as input, TeCH aims at reconstructing a high-fidelity 3D clothed human. Here, "high-fidelity" refers to the inclusion of consistent texture with intricate patterns, as well as detailed full-body geometry. To achieve this, TeCH follows a two-step procedure: Firstly, a text prompt that describes the human in the input image is obtained via the human parsing model SegFormer [117] and the VQA model BLIP [65] (Sec. 3.1). This descriptive prompt is used to guide the generation process in DreamBooth [101], a personalized Text-to-Image diffusion model fine-tuned on augmented input images. Secondly, the 3D human, which is represented as hybrid DMTet and initialized with SMPL-X (Sec. 3.2), is optimized with SDS losses [94] computed from the personalized DreamBooth (Sec. 3.3). The Score Distillation Sampling (SDS) loss has been introduced in DreamFusion [94] for the task of Text-to-3D generation of general objects, by optimizing a neural radiance field (NeRF) with gradients from a frozen diffusion model. In our case, we utilize the SDS loss to guide the reconstruction of a 3D human from a single input image, employing a multi-stage optimization strategy (Sec. 3.3) to get a consistent alignment of geometry and texture.

3.1. Extracting Text-guidance from the Observation

Parsing human attributes. As depicted in Fig. 3, given the input image of a human, SegFormer [117], which is fine-tuned on ATR dataset [70, 71], is applied to recognize each part of the garments (*e.g.* hat, skirt, pants, belt, shoes). To obtain detailed descriptions (*i.e.* color and style) of the parsed garments, we utilize the vision-language model BLIP [65] as VQA captioner. This model has been pre-trained on a vast collection of image-text pairs, enabling it to automatically generate descriptive prompts. Rather than using naive image captioning, we employ a series of fine-grained VQA questions $\{Q_i\}$ (see Appx.'s Sec. B) as input to BLIP. These questions cover garment styles, colors, facial features, and hairstyles, with the corresponding answers denoted as $\{A_i\}$. The set of $\{A_i\}$ will be inserted into a pre-defined template to create text prompts P_{VQA}, which will serve as text-guidance to condition the text-to-image diffusion model, recap the full method overview in Fig. 2.

Embedding subject-specific appearance. Does the text prompt P_{VQA} comprehensively capture all the visual characteristics of the subject? No, a picture is worth a thousand words. Thus, we utilize DreamBooth [101] to learn the *indescribable* visual appearance. DreamBooth is a method for "personalizing" a diffusion model through few-shot tuning (3∼5 images). We perform DreamBooth's fine-tuning on a pre-trained Stable Diffusion (v1.5) as the base model. To generate the needed inputs, we augment the single input image with five different backgrounds, as shown in Fig. 3. To prevent language drift, we assign the subject classes "man" or "woman" based on the gender determined by the VQA. After fine-tuning DreamBooth, the subject-specific distinctive appearance is encoded within a unique identifier token "$[V]$". We insert "$[V]$" into the prompt P_{VQA}, to construct the final text prompt P used by the personalized DreamBooth $\mathcal{D}'$. In Fig. 4, you can see how these individual prompts contribute to the final appearance.

(a) Detailed ablation results of textual guidance

*"[V] (~~white~~ **grey**) turtleneck sweater"*   *"[V] mini (~~pink~~ **red**) skirt"*   *"[V] (~~red~~ **blue**) and black striped shirt"*

(b) Editing garment colors with subject-specific token

Figure 5. **(a)** Top depicts the impact of specific elements within the textual guidance, such as garment styles & colors, hairstyle, facial features, and the placement & inclusion of "[V]". **(b)** Bottom demonstrates that TeCH facilities text-guided garment color editing.

Deeper analysis of description P. In Fig. 5 (a), we first show the impact of individual elements within the text prompt, including garment styles & colors, hairstyle, and face, which guide the model to recover the appearance of each attribute of the clothed human. The first column shows that a basic class description alone cannot effectively guide the reconstruction process. However, in the subsequent columns, text guidance incorporating detailed descriptions of clothing proves successful in accurately reconstructing the structure of clothed humans. Furthermore, with additional information regarding colors and hairstyles, the characters reconstructed by TeCH_{vqa} exhibit greater semantic consistency with respect to the input view. However, merely relying on VQA descriptions is insufficient for generating a "convincingly fake" appearance.

Only using the DreamBooth guidance (TeCH_{db}), helps to recover original garment patterns, which demonstrates that DreamBooth has a high-level understanding of texture patterns. However, it sometimes will *diffuse* the patterns to the entire human. By combining "[V]" with the VQA parsing text prompts P_{VQA}, TeCH produces remarkably realistic texture with consistent color and intricate patterns.

In Fig. 5 (b), we also demonstrate some text-guided garment color editing examples based on a fine-tuned DreamBooth model D' and subject-specific token "[V]".

3.2. Hybrid 3D Representation

To efficiently represent the 3D clothed human at a high resolution, we embed DMTet [32, 106] around the SMPL-X body mesh [86]. Specifically, we construct a compact

tetrahedral grid $(V_{\text{shell}}, T_{\text{shell}})$ within an outer shell M_{shell}, shown in Fig. 2-(c). Compared to the DMTet cubic-based tetrahedral grid, the outer shell tetrahedral grid is more computationally efficient for high-resolution geometry modeling of a human. Using PIXIE [26], we estimate an initial body $\mathcal{M}_{\text{body}}$. To create M_{shell}, a series of mesh dilation, down-sampling, and up-sampling steps are applied to the body mesh M_{body} (see details Sec. C of Appx.).

We use two MLP networks $\Psi_{\text{g}}, \Psi_{\text{c}}$ with hash encoding [83], parameterized by ψ_g and ψ_c to learn the geometry and color separately. The geometry network Ψ_{g} predicts the SDF value $\Psi_{\text{g}}(v_i) = s(v_i; \psi_g)$ of each DMTet vertex v_i. It is initialized by fitting it to the SDF of M_{shell}:

$$\mathcal{L}_{\text{init}} = \sum_{p_i \in \mathbf{P}} \|s(p_i; \psi_{\text{g}}) - \text{SDF}(p_i)\|_2^2, \qquad (1)$$

where $\mathbf{P} = \{p_i \in \mathbb{R}^3\}$ is a point set randomly sampled near M_{shell}, and $\text{SDF}(p_i)$ is the pre-computed pointwise SDF. Triangular meshes can be extracted from this efficient hybrid 3D representation by Marching Tetrahedra (MT) [24]:

$$M = \text{MT}(V_{\text{shell}}, T_{\text{shell}}, s(V_{\text{shell}}; \psi_{\text{g}})). \qquad (2)$$

Given the camera parameters $\mathbf{k}$, the generated mesh is rendered through differentiable rasterization $\mathcal{R}$ [62], to get the back-projected 3D locations $\mathcal{P}(M, \mathbf{k})$, rendered mask $\mathcal{M}(M, \mathbf{k})$, and rendered normal image $\mathcal{N}(M, \mathbf{k})$

$$\mathcal{R}(M, \mathbf{k}) = (\mathcal{P}(M, \mathbf{k}), \mathcal{M}(M, \mathbf{k}), \mathcal{N}(M, \mathbf{k})) \qquad (3)$$

5

The albedo of each back-projected pixel is predicted by the color network $\Psi_{\mathbf{c}}$, where ψ_c represents the parameters:

$$\mathcal{I}'(M, \psi_c, \mathbf{k}) = \Psi_{\mathbf{c}}(\psi_c, \mathcal{P}(M, \mathbf{k})). \quad (4)$$

As detailed in Section 3.3, we optimize this 3D representation using a coarse-to-fine strategy by applying successive subdivisions on the tetrahedral grids. Specifically, a more detailed surface $M_{\text{subdiv}}(\psi_{\text{g}})$ can be obtained by applying volume subdivision on the surface tetrahedral grids $(V_{\text{surface}}, T_{\text{surface}})$ that intersect with $M(\psi_{\text{g}})$. Note that the SDF values of the refined vertices are still inferred by Ψ_{g}.

3.3. Multi-stage Optimization

We adopt a multi-stage, coarse-to-fine optimization process to sequentially recover the subject's geometry and texture. In the initial stage, we utilize the tetrahedral representation to model the subject's geometry (Sec. 3.3.1). Next, the appearance is recovered using the mesh that is extracted from the tetrahedral grid (Sec. 3.3.2). Both stages are leveraging SDS-based losses using the personalized DreamBooth model which provides multi-view supervision by sampling new camera views as described in Sec. 3.3.3.

3.3.1 Geometry Stage

We optimize the geometry based on a silhouette loss $\mathcal{L}_{\text{sil}}$ using the orig. image, a text-guided SDS loss on rendered normal images $\mathcal{L}_{\text{SDS}}^{\text{norm}}$, and geometric regularization $\mathcal{L}_{\text{reg}}$ based on pred. normals $\mathcal{L}_{\text{norm}}$ and surface smoothness $\mathcal{L}_{\text{lap}}$:

$$\mathcal{L}_{\text{geometry}} = \lambda_{\text{sil}}\mathcal{L}_{\text{sil}} + \lambda_{\text{SDS}}\mathcal{L}_{\text{SDS}}^{\text{norm}} + \mathcal{L}_{\text{reg}} \quad (5)$$
$$\mathcal{L}_{\text{reg}} = \lambda_{\text{norm}}\mathcal{L}_{\text{norm}} + \lambda_{\text{lap}}\mathcal{L}_{\text{lap}},$$

where λ represents the weights to balance the losses. During optimization of this loss, we perform a coarse-to-fine subdivision on DMTet, to robustly produce a high-resolution mesh for the clothed body. Specifically, the optimization is first performed w/o subdivision for $t_{\text{coarse}} = 5000$ iters, and then with subdivision for $t_{\text{fine}} = 5000$ iters.

Pixel-aligned silhouette loss. The silhouette loss [125, 133] enforces pixel-alignment with the foreground mask $\mathcal{S}$ of the input image $\mathcal{I}$ under the input camera view $\mathbf{k}$:

$$\mathcal{L}_{\text{sil}} = \|\mathcal{S} - \mathcal{M}(M, \mathbf{k})\|_2^2$$
$$+ \sum_{x \in \text{Edge}(\mathcal{M}(M, \mathbf{k}))} \min_{\hat{x} \in \text{Edge}(\mathcal{S})} \|x - \hat{x}\|_1. \quad (6)$$

It consists of (1) a pixel-wise L2 loss over the foreground mask $\mathcal{S}$ and the rendered silhouette $\mathcal{M}$, and (2) an edge distance loss, based on the distance of each silhouette boundary pixel $x \in \text{Edge}(\mathcal{M}(M, \mathbf{k}))$ to the nearest foreground mask boundary pixel $\hat{x} \in \text{Edge}(\mathcal{S})$.
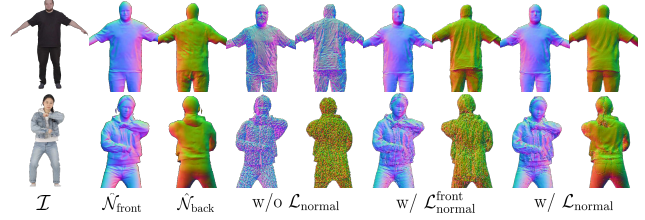

Figure 6. **The effects of normal regularization.** $\mathcal{L}_{\text{norm}}$ regularizes the surface with predicted normal images $\hat{\mathcal{N}}_{\text{front}}, \hat{\mathcal{N}}_{\text{back}}$.

SDS loss on normal images. Inspired by Fantasia3D [16], our approach integrates normal renderings with the SDS loss [94]. It enables TeCH to effectively capture intricate geometric details without rendering the color image. Given the surface normals $\mathbf{n} = \mathcal{N}(M, \mathbf{k})$, $\mathcal{L}_{\text{SDS}}^{\text{norm}}$ is defined as:

$$\mathcal{L}_{\text{SDS}}^{\text{norm}} = \nabla_{\psi_{\text{g}}}\mathcal{L}_{\text{SDS}}^{\text{norm}}(\mathbf{n}, \mathbf{c}^{P_{\text{norm}}})$$
$$= \mathbb{E}_{\text{t},\epsilon}\left[w_t\left(\hat{\epsilon}_{\phi'}(\mathbf{z}_t^{\mathbf{n}}; \mathbf{c}^{P_{\text{norm}}}, t) - \epsilon\right)\frac{\partial \mathbf{n}}{\partial \psi_{\text{g}}}\frac{\partial \mathbf{z}^{\mathbf{n}}}{\mathbf{n}}\right], \quad (7)$$

where $\mathbf{c}^{P_{\text{norm}}}$ is the text condition with an augmented prompt P_{norm}. We construct P_{norm} from P by adding an extra description "a detailed sculpture of" to better reflect the intrinsic characteristics of normal maps.

Geometric regularization. We found that relying solely on silhouette and SDS losses may lead to the generation of noisy surfaces, which is particularly evident for subjects wearing complex clothing. To address this, we leverage normal estimations as an additional constraint to regularize the reconstructed surface (see Fig. 6):

$$\mathcal{L}_{\text{norm}}(\hat{\mathcal{N}}_{\mathbf{k}}, \mathbf{n}) = \lambda_{\text{MSE}}^{\text{norm}}\left\|\hat{\mathcal{N}}_{\mathbf{k}} - \mathbf{n}\right\|_2^2 + \text{LPIPS}(\hat{\mathcal{N}}_{\mathbf{k}}, \mathbf{n})), \quad (8)$$

where $\hat{\mathcal{N}}_{\mathbf{k}}$ are the front and back normal maps *estimated* using ICON [120] indexed by the view $\mathbf{k}$ ($\mathbf{k} \in \{\text{front, back}\}$). $\mathbf{n}$ are the corresponding *differentiably rendered* normal images of the 3D shape Ψ_{g}. We use a combination of LPIPS and MSE loss to enhance the similarity between $\hat{\mathcal{N}}_{\mathbf{k}}$ and $\mathbf{n}$. Furthermore, we utilize a regularization loss based on Laplacian smoothing [6], represented as $\mathcal{L}_{\text{lap}}$.

Mesh extraction. We use Marching Tetrahedra [24] to extract the mesh from the tetrahedral grid. Like ECON [121], we register SMPL-X to this mesh which allows us to transfer skinning weights for reposing (see Fig. 9). In addition, we replace the hands with SMPL-X ones which effectively mitigates any potential artifacts introduced during reposing which is needed in the subsequent texture generation stage.

3.3.2 Texture Stage

Given the triangular mesh from the geometry stage, we optimize the full texture. To recover the consistent details and color, even for self-occluded regions, we render both the input pose (M_{in}) and the A-pose (M_{A}) during optimization. The textures of M_{in} and M_{A} are modeled by Ψ_{color} in the 3D space of M_{A}. We optimize the texture from scratch with ψ_c randomly initialized. In Fig. 7, we show the effect of this multi-pose training. We utilize an occlusion-aware reconstruction loss $\mathcal{L}_{\text{recon}}$ on the input view of M_{in}, an SDS loss $\mathcal{L}_{\text{SDS}}^{\text{color}}$ with text guidance on rendered color images of both M_{in} and M_{A}, and a color consistency regularization $\mathcal{L}_{\text{CD}}$, with respective weights λ to balance the individual losses:

$$\mathcal{L}_{\text{texture}} = \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{SDS}}\mathcal{L}_{\text{SDS}}^{\text{color}} + \lambda_{\text{CD}}\mathcal{L}_{\text{CD}}, \quad (9)$$

Note that $\mathcal{L}_{\text{CD}}$ is only utilized after the full-body texture convergence (5000 iters), in an additional optimization phase of 2000 iterations for enforcing color consistency.

Occlusion-aware reconstruction loss. To enforce pixel-alignment, we apply an input view reconstruction loss to minimize the difference between input image $\mathcal{I}$ and the albedo-rendered image $\mathcal{I}'(M, \psi_c, \mathbf{k}_{\mathcal{I}})$. Additionally, we have observed that applying $\mathcal{L}_{\text{recon}}$ to self-occluded areas may lead to incorrect texture due to geometry misalignment. Therefore, an occlusion-aware mask m_{occ} is introduced to selectively exclude the $\mathcal{L}_{\text{recon}}$ in occluded regions.

$$\mathcal{L}_{\text{recon}} = m_{\text{occ}}(\lambda_{\text{MSE}} \|\mathcal{I} - \mathcal{I}'(M, \psi_c, \mathbf{k}_{\mathcal{I}})\|_2^2$$
$$+ \text{LPIPS}(\mathcal{I}, \mathcal{I}'(M, \psi_c, \mathbf{k}_{\mathcal{I}}))), \quad (10)$$

where $\mathbf{k}_{\mathcal{I}}$ denotes the input view camera, and λ_{MSE} is a weight to balance the two loss terms.

SDS loss on color images. To recover the full-body texture, including unseen regions, we update ψ_c via SDS loss $\mathcal{L}_{\text{SDS}}^{\text{color}}$ with text guidance. This loss is calculated based on random-view color renderings $\mathbf{x} = \mathcal{I}'(\psi_g, \psi_c, \mathbf{k})$, and DreamBooth $\mathcal{D}'$ parameterized by ϕ' and guided by text prompt P.

$$\mathcal{L}_{\text{SDS}}^{\text{color}} = \nabla_{\psi_c} \mathcal{L}_{\text{SDS}}^{\text{color}}(\mathbf{x}, \mathbf{c}^P)$$
$$= \mathbb{E}_{t,\epsilon}\left[w_t\left(\hat{\epsilon}_{\phi'}(\mathbf{z}_t^{\mathbf{x}}; \mathbf{c}^P, t) - \epsilon\right) \frac{\partial \mathbf{x}}{\partial \psi_c} \frac{\partial \mathbf{z}^{\mathbf{x}}}{\mathbf{x}} \right], \quad (11)$$

where $\mathbf{k}$ is the camera pose, $\mathbf{c}^P$ is the text embedding of P.

Chamfer-based color consistency loss. As mentioned in DreamFusion [94], the SDS loss may result in over-saturated colors which will cause a noticeable color disparity between visible and invisible regions. To mitigate this issue, we incorporate a color consistency loss to ensure that the rendered novel views align closely with the color distributions observed in the input view. We quantify the disparity between the color distributions using a chamfer Distance
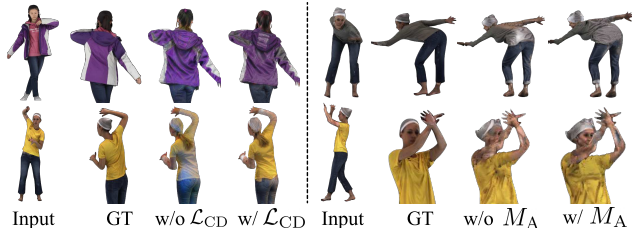

Figure 7. **The effects of color consistency loss $\mathcal{L}_{\text{CD}}$ and multi-pose training** (M_{A}) **for texture optimization.** $\mathcal{L}_{\text{CD}}$ corrects the over-saturated back-side color generated by SDS, while M_{A} improves the texture quality under self-occlusion or extreme poses.

(CD) by treating the pixels from both views as point clouds within the RGB color space:

$$\mathcal{L}_{\text{CD}} = \sum_{x \in \mathbf{F_x}} \min_{y \in \mathbf{F}_{\mathcal{I}}} \|x - y\|_2^2 + \sum_{y \in \mathbf{F}_{\mathcal{I}}} \min_{x \in \mathbf{F_x}} \|x - y\|_2^2,$$
$$(12)$$

where $\mathbf{F_x}$ and $\mathbf{F}_{\mathcal{I}}$ respectively represent the foreground pixels of the novel-view albedo rendering $\mathbf{x}$, and the input view $\mathcal{I}$. The improvement using $\mathcal{L}_{\text{CD}}$ is shown in Fig. 7.

3.3.3 Camera sampling during optimization

To optimize the 3D shape and texture using multi-view renderings, cameras are randomly sampled in a way that ensures comprehensive coverage of the entire body by adjusting various parameters. To mitigate the occurrence of mirrored appearance artifacts (*i.e.*, Janus-head), we incorporate view-aware prompts ("`front/side/back/overhead view`") w.r.t. the viewing angle in the diffusion-based generation process, whose effectiveness has been demonstrated in DreamBooth [94]. In order to improve facial details, we also sample cameras positioned around the face, together with the additional prompt "`face of`". More details about the camera sampling strategy are in Sec. D of Appx.

4. Experiments

We compare TeCH with state-of-the-art image-based 3D clothed human reconstruction methods, including body-agnostic methods, such as PIFu [103], PIFuHD [104] and PHORHUM [5], as well as methods that utilize SMPL-(X) [78, 92] body prior, such as PaMIR [136], ICON [120] and ECON [121]. For a fair comparison, all methods (*i.e.*, PIFu, PaMIR, ICON, ECON) utilize the same normal estimator from ICON. Official PIFu, PaMIR and PHORHUM are used to evaluate the quality of texture. For ECON, we use ECON$_{\text{EX}}$, due to its superior performance on both "OOD poses" and "OOD outfits" cases, as reported in the original paper [121]. Note that PHORHUM uses a different camera model which is not compatible with our testing

7

Method	3D Metrics						2D Image Quality Metrics					
	CAPE			THuman2.0			CAPE			THuman2.0		
	Chamfer ↓	P2S ↓	Normal ↓	Chamfer ↓	P2S ↓	Normal ↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
	w/o SMPL-X body prior											
PIFu	1.9683	1.6236	0.0623	1.9305	1.8031	0.0802	27.0994	0.9362	0.0987	23.5068	0.9296	0.1083
PIFuHD	3.2018	2.9930	0.0758	2.4613	2.3605	0.0924	-	-	-	-	-	-
	w/ SMPL-X body prior											
PaMIR	1.3756	1.1852	0.0526	1.2979	**1.2188**	0.0676	27.7279	0.9456	0.0904	22.5466	0.9266	0.1082
ICON	0.8689	0.8397	0.0360	**1.1382**	1.2285	0.0623	-	-	-	-	-	-
ECON	0.9186	0.9227	0.0330	1.2585	1.4184	**0.0612**	-	-	-	-	-	-
TeCH	**0.7416**	**0.6962**	**0.0306**	1.2364	1.2715	0.0642	**28.3601**	**0.9490**	**0.0639**	**25.2107**	**0.9363**	**0.0835**

Table 1. **Quantitative evaluation against SOTAs.** TeCH surpasses SOTA baselines in terms of both 3D metrics and 2D image quality metrics. This demonstrates its superior performance in accurately reconstructing clothed human geometry with intricate details, as well as producing high-quality textures with consistent appearance.

data, thus, we use PHORHUM only for qualitative comparisons. More implementation details about network structure and optimization setting can be found at Sec. E of Appx.

4.1. Models and Datasets

Off-the-shelf models. TeCH relies on multiple off-the-shelf pre-trained models and does not need any additional training data. Specifically, we use officially released stable-diffusion-v1.5[*] as T2I diffusion model, which is trained on LAION-5B, the VQA model BLIP [65] pretrained on 129M images from multiple datasets [15, 61, 74, 84, 87, 105] and fine-tuned on VQA2.0 [35], Seg-Former[*] [117] pretrained from [10, 20, 23, 137] and fine-tuned on ATR[69], PIXIE [26] trained on human images from multiple datasets [19, 74, 90, 115, 140], and the normal predictor of ICON [120] trained on AGORA [91].

Datasets for evaluation. Based on the high-fidelity 3D textured scans from CAPE [79] and THuman2.0 [126], we perform quantitative evaluations. We follow ICON [120] to analyze the robustness of reconstructions under both simple and complex poses (150 scans from CAPE). An additional 150 THuman2.0 scans are included, which comprises 100 subjects that were manually selected to represent a diverse range of clothing styles (*e.g.*, open jackets, long coats, garments with intricate patterns, *etc.*), and 50 randomly sampled subjects. The images are rendered at a resolution of 512×512. For qualitative comparison, we selected the SHHQ dataset [28] due to its wide range of textures, outfits, and gestures. From this dataset, we randomly sampled 90 images with official mask annotations.

4.2. Quantitative Comparison

We quantitatively evaluate the reconstruction quality of geometry and appearance, using the **Chamfer** (bi-directional point-to-surface) and **P2S** (1-directional point-to-surface) distance, to measure the difference between the reconstructed and ground-truth meshes. Additionally, we re-

---

[*] `runwayml/stable-diffusion-v1-5`
[*] `matei-dorian/segformer-b5-finetuned-human-parsing`

port the L2 **Normal** error between normal images rendered from both meshes, to measure the consistency and fineness of local surface details, by rotating the camera by $\{0°, 90°, 180°, 270°\}$ w.r.t. to the input view. To evaluate the quality of the texture, we report 2D image quality metrics, on the multi-view colored images rendered in the same way as the normal images, including **PSNR** (Peak Signal-to-Noise Ratio), **SSIM** (Structural Similarity) and **LPIPS** (learned perceptual image path similarity).

As shown in Tab. 1, TeCH demonstrates superior performance across all 2D metrics and 3D metrics on CAPE. This reveals that TeCH can accurately reconstruct both geometry and texture, even for subjects with challenging poses (CAPE) or loose clothing (THuman2.0). However, on THuman2.0, it achieves comparable reconstruction accuracy to prior-based methods. This can be attributed to the fact that the hallucinated back-side may differ from the ground truth while still appears realistic. A perceptual study Tab. 2 was conducted for additional clarification. See Sec. 4.4 of Appx. for more results on these datasets.

4.3. Perceptual Evaluation

To assess the generalization of TeCH on in-the-wild images and evaluate the perceptual quality of our results, we conducted a perceptual study using 90 randomly sampled images from the SHHQ dataset [28]. Participants were shown videos showcasing rotating 3D humans reconstructed by TeCH, as well as the baselines (PaMIR [136], PIFu [103], ICON [120], ECON [121] and PHORHUM [5]). They were asked to choose the more realistic and consistent result based on the input image. We gathered a total of 3,150 pairwise comparisons from 63 participants, uniformly covering 90 SHHQ subjects. The results in Tab. 2 show that TeCH is preferred, both, in terms of geometry and texture. As illustrated in Fig. 8, unlike other methods that tend to reconstruct overly smooth surfaces and blurry textures, TeCH shows remarkable generalizability when applied to in-the-wild images featuring diverse clothing styles and gestures. It produces more realistic clothing, haircut, and facial details, even for unseen back-side views.
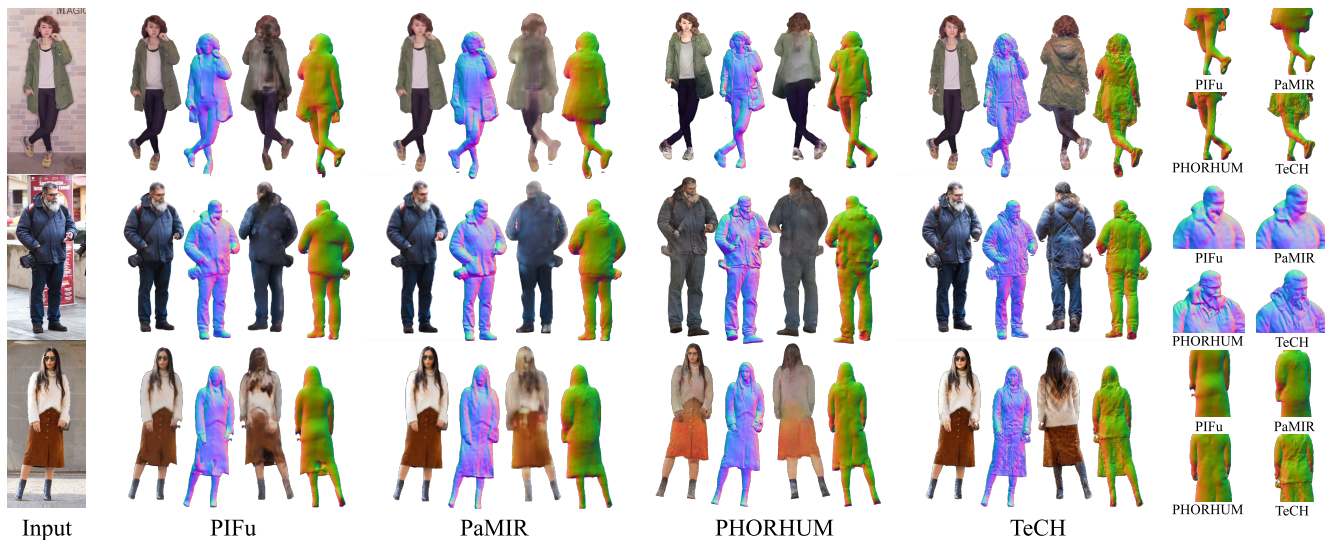
Figure 8. Qualitative comparison on SHHQ images. TeCH generalizes well on in-the-wild images with diverse clothing styles and textures. It successfully recovers the overall structure of the clothed body with text guidance, and generates realistic full-body texture which is consistent with the colored pattern and the material of the clothes. 🔍 **Zoom in** to see the geometric details.

Preference (%, ↑)	PIFu	PaMIR	PHORHUM	ICON	ECON
Geometry	88.6	87.0	81.7	97.94	90.48
Colored Rendering	95.1	93.7	93.0	-	-

Table 2. **Perceptual study**. The percentages of user preference to TeCH compared to other baselines are reported. Most participants preferred TeCH in both geometry and colored rendering (texture).

	Experiment settings						3D Metrics			2D Image Quality Metrics		
	VQA	DreamBooth	$\mathcal{L}_{norm}$	$\mathcal{L}_{CD}$	M_A	multi-stage	Chamfer ↓	P2S ↓	Normal ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	✓	✓	✓	✓	✓	✓	0.9794	0.9779	0.0466	26.7565	0.9428	0.0741
A.	✓	✗	✓	✓	✓	✓	0.9959	1.0192	0.0454	26.2078	0.9405	0.0813
	✗	✓	✓	✓	✓	✓	1.0032	1.0218	0.0470	26.9602	0.9428	0.0785
	✓	✓	✓	✓	✓	✗	0.9957	0.9963	0.0468	26.0465	0.9395	0.0775
B.	✓	✓	✗	✗	✗	✓	1.0882	0.9203	0.0870	-	-	-
C.	✓	✓	✓	✗	✓	✓	-	-	-	26.6500	0.9427	0.0746
	✓	✓	✓	✓	✗	✓	-	-	-	26.6506	0.9425	0.0786

Table 3. **Ablation study.** We quantitatively evaluate the effectiveness of each component. Top two results are colored as first second. All the factors are grouped w.r.t. their influence: A. geometry+texture, B. geometry only, C. texture only.

4.4. More Qualitative Results

In addition to Fig. 8, we show more qualitative comparisons between TeCH and other baselines (PIFu [103], PIFuHD [104], PaMIR [136], PHORHUM [5], ICON [120], ECON [121]) on CAPE, THuman2.0, and SHHQ [28] images (Figs. 12 to 14 of Appx.), by visualizing multi-view surface normals, color renderings, and zoomed-in details. For subjects in CAPE and THuman2.0, TeCH precisely recover the human shape and generate high-quality details of garments and facial features, regardless of hard poses, complex texture, loose clothing, or self-occlusion. Also, Fig. 14 demonstrates the strong generalizability of TeCH on in-the-wild images, more rotating 3D humans are provided in video.

4.5. Ablation Studies

To assess the effectiveness of key designs in TeCH, we perform ablation studies on a 10% subset of the test set, consisting of 15 subjects from THuman2.0 and 15 from CAPE. The detailed analysis on these results is as follows:

Text guidance. Table 3-A shows that either the "VQA-only" or "DreamBooth-only" guidance exhibit a decrease in performance w.r.t. reconstruction accuracy (Chamfer,

P2S) and texture quality (LPIPS). Figure 4 shows that VQA prompts help to recover the overall structure of clothing, while DreamBooth enhances the fine details of the texture pattern. Combining both text guidance sources yields the best results. A detailed analysis of individual descriptive texts (*e.g.*, garments, hairstyles, *etc*.) is in Fig. 5

Geometric regularization. As shown in Fig. 6, using only $\mathcal{L}_{SDS}^{norm}$ to optimize the geometry will produce noisy artifacts, particularity noticeable in loose clothes. The significant increase in "Normal" error shown in Tab. 3-B echos this. This issue can be mitigated by incorporating $\mathcal{L}_{norm}$ at the beginning of the optimization.

Consistent texture recovery. The results presented in Fig. 7 demonstrate that $\mathcal{L}_{CD}$ notably enhances color consistency between the frontal and back sides, and "multi-pose" training (M_A) improves texture quality when dealing with self-occlusion scenarios. This improvement is further supported by Tab. 3-C, across all 2D image quality metrics.

Multi-stage optimization. As shown in Tab. 3-A, compared to the decoupled two-stage optimization (Ours), the joint optimization results in a performance drop across both

3D and 2D metrics. This may be attributed to the entanglement of the gradients from the geometry and texture branches during optimization. Notably, in the separate texture stage, a colored image is rendered from the extracted mesh, saving 20% of the run time compared to joint optimization, which involves rendering from the DMTet mesh.

5. Applications

5.1. Avatar animation

Following the geometry optimization phase, TeCH aligns the clothed body mesh with the SMPL-X model, enabling us to animate the reconstructed avatar with SMPL-X motions [80], as shown in Fig. 9 and video.

Input   Animation video results

Figure 9. **Animate TeCH with SMPL-X motions.**

5.2. Avatar editing

The text-guided texture generation feature also allows us to edit the texture of the generated avatars. Here, we show stylization results with different painting styles, like "`pop art, pixel art, van gogh`". The resulting texture not only features the desired styles but also preserves the inherent appearance traits of the original character.
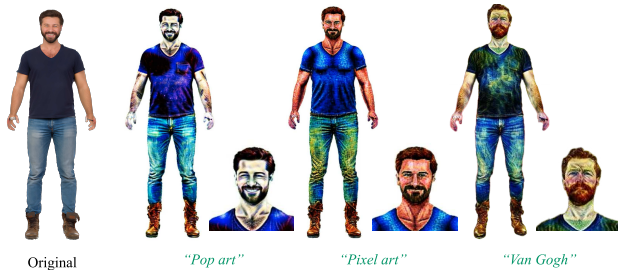

Original   *"Pop art"*   *"Pixel art"*   *"Van Gogh"*

Figure 10. **Text-guided stylization.**

6. Discussion

Limitations. Despite achieving impressive results on diverse datasets, some failures cases still exist, see Fig. 11: **A.** TeCH occasionally fails for extremely loose clothing, this may relate to the constraint from SMPL-X-based initialization. **B.** mismatched pattern may occur as tattoo. **C.** TeCH relies on robust SMPL-X pose estimation, which is still an unsolved problem, especially for challenging poses.

A. Extremely loose clothes   B. Mismatched patterns   C. SMPL-X Failure

Figure 11. The proposed method might exhibit noisy surfaces for extremely loose clothing, or mismatched patterns. If PIXIE [26] is predicting a wrong initial pose, the error propagates to TeCH.

Efficiency. For each subject, training DreamBooth takes 20 min, DMTet SMPL-X initialization takes 20 min, geometry stage (coarse-50 min, fine-50 min), mesh post-processing takes 10 min (remeshing, SMPL-X registration, hand replacement), texture stage takes 140 min, 270 min in total. Thus, our per-subject optimization process remains time-consuming, requiring approximately 4.5 hours per subject on a V100 GPU. Addressing these limitations is crucial to facilitate broader applications.

Future work. Leveraging controllable T2I models [52, 82, 96, 134] may help to improve the controllability and stability of generation process. Also, how to compositionally generate the separate components, such as haircut [107], accessories [31], and decoupled outfits [27], is still an unsolved problem. We leave these for future research.

Broader impact. TeCH has many potential applications Sec. 5. However, as the technique advances, it has the potential to facilitate deep-fake avatars and raise IP concerns. Regulations should be established to address these issues alongside its benefits in the entertainment industry.

7. Conclusion

We have proposed TeCH to reconstruct a lifelike 3D clothed human from a single image, with detailed full-body geometry and high-quality, consistent texture. The core insight is that we can leverage descriptive text prompts and personalized Text-to-Image diffusion models to optimize the 3D avatar including parts that are not visible in the input. Extensive experiments validate the superiority of TeCH over existing methods in terms of geometry and rendering quality. We believe that this paradigm of using image and textual descriptions for 3D body reconstruction is a stepping stone also for reconstruction tasks beyond human bodies.

Appendices

We provide an additional introduction to the preliminaries (Sec. A) of TeCH. We list the VQA questions P_{VQA} (Sec. B). Additional implementation details to construct the outer shell around SMPL-X (Sec. C), as well as details on the camera sampling strategy (Sec. D) are given. Implementation details of network structure and optimization setting (Sec. E). Based on the benchmark datasets (CAPE, THuman2.0) and in-the-wild photos used in the perceptual studies, we present more qualitative results (Figs. 12 to 14).

A. Preliminaries

DreamBooth. Pretrained text-to-Image diffusion models [99, 100, 102] lack the ability to mimic the appearance of subjects in a given reference set and synthesize novel renditions of them in different contexts. To enable subject-driven image generation, DreamBooth [101] *personalizes* the pretrained diffusion model through few-shot tuning.

Specifically, for a pre-trained image diffusion model $\hat{\mathbf{x}}_\phi$, the model takes an initial noise $\epsilon \sim \mathcal{N}(0,1)$, and a text embedding $\mathbf{c} = \Gamma(P)$, generated by the text encoder Γ and a text prompt P, to produce an image $\mathbf{x}_{\text{gen}} = \hat{\mathbf{x}}_\phi(\epsilon, \mathbf{c})$. DreamBooth uses 3∼5 images of the same subject to fine-tune the diffusion model using MSE denoising losses:

$$\mathbb{E}_{\mathbf{x}, \mathbf{c}, \epsilon, \epsilon', t} = \big[w_t \| \hat{\mathbf{x}}_\phi(\alpha_t \mathbf{x}_{\text{gt}} + \sigma_t \epsilon, \mathbf{c}) - \mathbf{x}_{\text{gt}} \|_2^2$$
$$+ \lambda w_{t'} \| \hat{\mathbf{x}}_\phi(\alpha_{t'} \mathbf{x}_{\text{prior}} + \sigma_{t'} \epsilon', \mathbf{c}_{\text{prior}}) - \mathbf{x}_{\text{prior}} \|_2^2 \big] \quad (13)$$

Where $\mathbf{x}_{\text{gt}}$ represents ground-truth images, and $\mathbf{c}$ is the embedding of a text prompt with a rare token as the unique identifier, and α_t, σ_t, w_t controls the noise schedule and sample quality of the diffusion process at time $t \sim \mathcal{U}([0,1])$. The second term is the prior-preservation loss weighted by λ, which is supervised by self-generated images $\mathbf{x}_{\text{prior}}$ conditioned with the class-specific embedding $\mathbf{c}_{\text{prior}} = \Gamma(\text{"a man/woman"})$. This loss mitigates the phenomenon of language drift, where the model collapses into a single mode by associating the class name with a particular instance, thus augmenting the output diversity.

Score Distillation Sampling (SDS). DreamFusion [94] introduces Score Distillation Sampling (SDS) loss, to perform Text-to-3D synthesis by using pretrained 2D Text-to-Image diffusion model ϕ. Instead of sampling in pixel space, SDS optimizes over the 3D volume, which is parameterized with θ, with the differential renderer g, so the generated image $\mathbf{x} = g(\theta)$ closely resembles a sample from the frozen diffusion model. Here is the gradient of $\mathcal{L}_{\text{SDS}}$:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \mathbf{x} = g(\theta))$$
$$= \mathbb{E}_{t, \epsilon} \left[w_t \left(\hat{\epsilon}_\phi(\mathbf{z}_t^{\mathbf{x}}; \mathbf{c}, t) - \epsilon \right) \frac{\partial \mathbf{x}}{\partial \theta} \frac{\partial \mathbf{z}^{\mathbf{x}}}{\partial \mathbf{x}} \right] \quad (14)$$

where $\hat{\epsilon}_\phi(\mathbf{z}_t^{\mathbf{x}}; \mathbf{c}, t)$ denotes the noise prediction of the diffusion model with condition $\mathbf{c}$ and latent $\mathbf{z}_t^{\mathbf{x}}$ of the generated image $\mathbf{x}$. Such SDS-guided optimization is performed with random camera poses to improve the multi-view consistency. In contrast to DreamFusion, the 3D shape here is parameterized with an improved DMTet instead of NeRF.

Deep Marching Tetrahedra (DMTet). DMTet [32, 106] is a hybrid 3D representation designed for high-resolution 3D shape synthesis and reconstruction. It incorporates the advantages of both explicit and implicit representations, by learning Signed Distance Field (SDF) values on the vertices of a deformable tetrahedral grid. For a given DMTet, represented as (V_T, T), where V_T are the vertices in the tetrahedral grid T, comprising K tetrahedrons $T_k \in T$, with $k \in \{1, \dots, K\}$. Each tetrahedron is defined by four vertices $\{v_k^1, v_k^2, v_k^3, v_k^4\}$. The objective of the model is firstly to estimate the SDF value $s(v_i)$ for each vertex, then to iteratively refine the surface and subdivide the tetrahedral grid by predicting the position offsets Δv_i and SDF residual values $\Delta s(v_i)$. A triangular mesh can be extracted through Marching Tetrahedra [24]. As noted by Magic3D [73], DMTet offers two advantages over NeRF, **fast-optimization** and **high-resolution**. It achieves this by efficiently rasterizing a triangular mesh into high-resolution image patches using a differentiable renderer [62], enabling interaction with pre-trained high-resolution latent diffusion models, such as eDiff-I [7], and Stable Diffusion [100].

B. VQA Questions Q

To construct the descriptive prompt P_{VQA}, we designed a series of questions to parse clothed human attributes. First, we use BLIP [65] and a series of general questions Q_{general} to parse genders, facial appearance, hair colors, hairstyles, facial hairs, and body poses. Secondly, we use SegFormer [117] to parse human garments, consisting of 10 categories {hat, sunglasses, upper-clothes, skirt, pants, dress, belt, shoes, bag, scarf}, denoted as G, and use another group of questions Q_{garments} to parse the attribute of each garment $g \in G$. All the questions are listed in Tab. 4.

Empirically, we found that the BLIP [65] VQA model tends to use $1 \sim 3$ words to answer these questions, so we simply concatenate all the answers and remove repeated words to construct P_{VQA}. Note that for the CAPE dataset, we add the dataset-specific description "hairnet" to the guidance as it is hard to be recognized by BLIP.

C. Construction of the Outer SMPL-X Shell

To construct a compact tetrahedral grid $(V_{\text{shell}}, T_{\text{shell}})$, we calculate a coarse outer shell M_{shell} from SMPL-X estimated body mesh M_{body}. Specifically, we dilate M_{body} with an offset of $\Delta M_{\text{body}} = 0.1$ and simplify the mesh

Groups	Quetions
Q_{general}	Is this person a man or a woman?
	What is this person wearing?
	What is the hair color of this person?
	What is the hairstyle of this person?
	Describe the facial appearance of this person.
	Does this person have facial hair?
	How is the facial hair of this person?
	Describe the pose of this person.
Q_{garments}	Is this person wearing g?
	What g is the person wearing? $\rightarrow d$
	What is the color of the $d + g$?
	What is the style of the $d + g$?

Table 4. **Predefined questions for parsing clothed human attributes.** g is the segmentation category of a part of the garments, and d is the recognized garment category from the answer to the second question in Q_{garments}.

by reducing triangle numbers by $r_{\text{decimate}} = 90\%$ using quadric decimation [42]. The we generate the tetrahedral grid $(V_{\text{shell}}, T_{\text{shell}})$ of this outer shell by TetGen [37] with a maximum volume size of 5×10^{-8}.

D. Camera Sampling

To ensure full coverage of the entire body and the human face, during optimization process, we sample virtual camera poses into two groups: 1) $\mathbf{K}_{\text{body}}$ cameras with a field of view (FOV) covering the full body or the main body parts, and 2) zoom-in cameras $\mathbf{K}_{\text{face}}$ focusing the face region.

The ratio $\mathcal{P}_{\text{body}}$ determines the probability of sampling $\mathbf{k} \in \mathbf{K}_{\text{body}}$, while the height h_{body}, radius r_{body}, elevation angle ϕ_{body}, and azimuth ranges θ_{body} are adjusted relative to the SMPL-X body scale. Empirically, we set $\mathcal{P}_{\text{body}} = 0.7$, $h_{\text{body}} = (-0.4, 0.4)$, $r_{\text{body}} = (0.7, 1.3)$, $\theta_{\text{body}} = [-180°, 180°)$, $\phi_{\text{body}} = \{0°\}$, with the M_{body} proportionally scaled to a unit space with xyz coordinates in the range $[-0.5, 0.5]$. To mitigate the occurrence of mirrored appearance artifacts (*i.e.*, Janus-head), we incorporate view-aware prompts, "`front/side/back/overhead view`", w.r.t. the viewing angle during generation process, whose effectiveness has been demonstrated in DreamBooth [94].

In order to enhance facial details, we sample additional virtual cameras positioned around the face $\mathbf{k} \in \mathbf{K}_{\text{face}}$, together with the additional prompt "`face of`". With a probability of $\mathcal{P}_{\text{face}} = 1 - \mathcal{P}_{\text{body}} = 0.3$, the sampling parameters include the view target c_{face}, radius range r_{face}, rotation range θ_{face}, and azimuth range ϕ_{face}. Empirically, we set c_{face} to the 3D position of SMPL-X head keypoint, $r_{\text{face}} = [0.3, 0.4]$, $\theta_{\text{face}} = [-90°, 90°]$ and $\phi_{\text{face}} = \{0°\}$.

E. Implementation Details

E.1. Network Structure

We use two networks Ψ_{g} and Ψ_{c} to predict the SDF for geometry modeling and to predict the RGB value for albedo texture modeling, respectively. For Ψ_{g}, we use a 2-layer MLP network with a hidden dimension of 32 and a hash positional encoding with a maximum resolution of 1028 and 16 resolution levels. During the forward process, we use coordinates of V_{shell} in the normalized unit space, the vertices of the tetrahedral grid as the input of Ψ_{g} to query SDF value for each vertex.

For Ψ_{c}, we use a similar network with 1-layer MLP and a hash positional encoding with a maximum resolution of 2048. We model the albedo texture in the canonical A-pose 3D space. Specifically, for the post-processed result mesh $M_{\text{in}} = (V_{\text{in}}, F)$, we register the model with SMPL-X, and repose it with the standard A-pose $M_{\text{A}} = (V_{\text{A}}, F)$. During rendering, if a target pixel is projected onto a triangle $(v_{\text{in}}^i, v_{\text{in}}^j, v_{\text{in}}^j)$, where $(i, j, k) \in F$ of the M_{in}. We query the pixel color with its corresponding 3d position in the A-pose space, calculated by interpolation of the triangle $(v_{\text{A}}^i, v_{\text{A}}^j, v_{\text{A}}^j)$. Additionally, we use two 2-layer MLP $\Psi_{\text{bg}}^{\text{g}}, \Psi_{\text{bg}}^{\text{c}}$ conditioned by camera $\mathbf{k}$ to learn adaptive 3D background colors for both normal map rendering $\mathcal{N}(M, \mathbf{k})$ and color rendering $\mathcal{I}'(M, \psi_{\text{c}}, \mathbf{k})$.

E.2. Optimization Details

In both stages of our multi-stage optimization pipeline, we use an Adam optimizer with a base learning rate of $\eta = 1 \times 10^{-3}$, and weight decay of $\lambda_{\text{WD}} = 5 \times 10^{-4}$

Geometry-stage optimization. We optimize Ψ_{g} in a coarse-to-fine manner, with $t_{\text{coarse}} = 5000$ steps w/o mesh subdivision and $t_{\text{fine}} = 5000$ steps w/ mesh subdivision. We use a loss weight setting of $\lambda_{\text{sil}} = 1 \times 10^4$, $\lambda_{\text{SDS}} = 1$, $\lambda_{\text{lap}} = 1 \times 10^4$, and a base loss weight $\lambda_{\text{norm}}^{\text{base}} = 1 \times 10^4$. For λ_{norm}, to ensure robust convergence of the geometry, we start with a higher value of λ_{norm} during each stage and gradually decrease it using a two-round cosine annealing, where $\lambda_{\text{norm}}(t)$ is the weight of $\mathcal{L}_{\text{norm}}$ at the t-th iteration:

$$\lambda_{\text{norm}}(t) = \begin{cases} 0.5\lambda_{\text{norm}}^{\text{base}}\left(1 + \cos\left(\frac{t}{t_{\text{coarse}}}\pi\right)\right) & \text{if } t < t_{\text{coarse}} \\ 0.5\lambda_{\text{norm}}^{\text{base}}\left(1 + \cos\left(\frac{t - t_{\text{coarse}}}{t_{\text{fine}}}\pi\right)\right) & \text{if } t \geq t_{\text{coarse}} \end{cases},$$
(15)

Texture-stage optimization. We optimize Ψ_{c} for $t_{\text{texture}} = 7000$ steps, with $\lambda_{\text{recon}} = 2 \times 10^4$ and $\lambda_{\text{SDS}} = 1$. Besides, we set $\lambda_{\text{CD}} = 0$ at the beginning of the training, and $\lambda_{\text{SDS}} = 1 \times 10^6$ at the last $t_{\text{CD}} = 2000$ iterations to enforce color consistency.
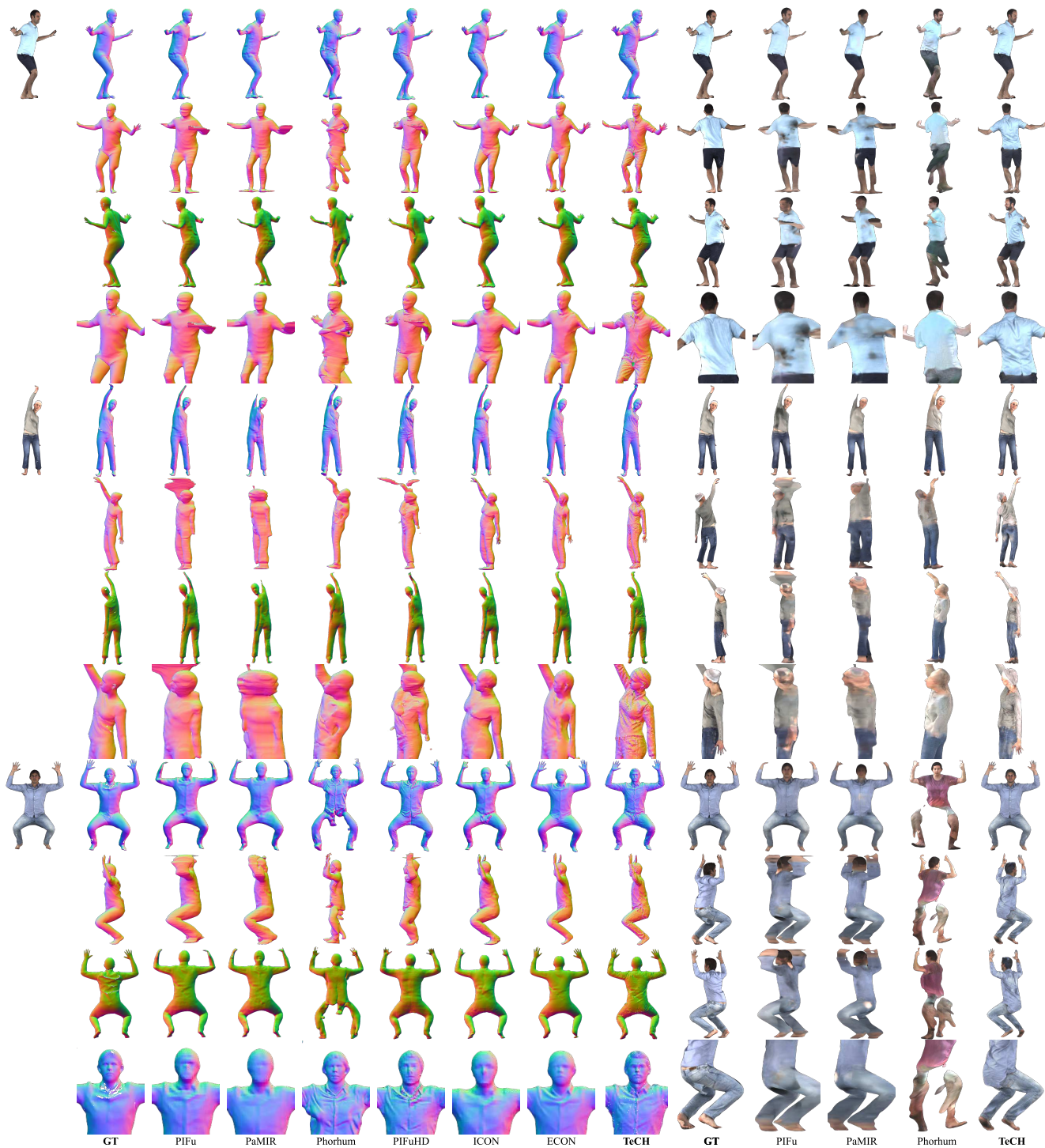
| **GT** | PIFu | PaMIR | Phorhum | PIFuHD | ICON | ECON | **TeCH** | **GT** | PIFu | PaMIR | Phorhum | **TeCH** |

Figure 12. **Qualitative comparison on CAPE**. TeCH performs better on subjects with challenging poses.
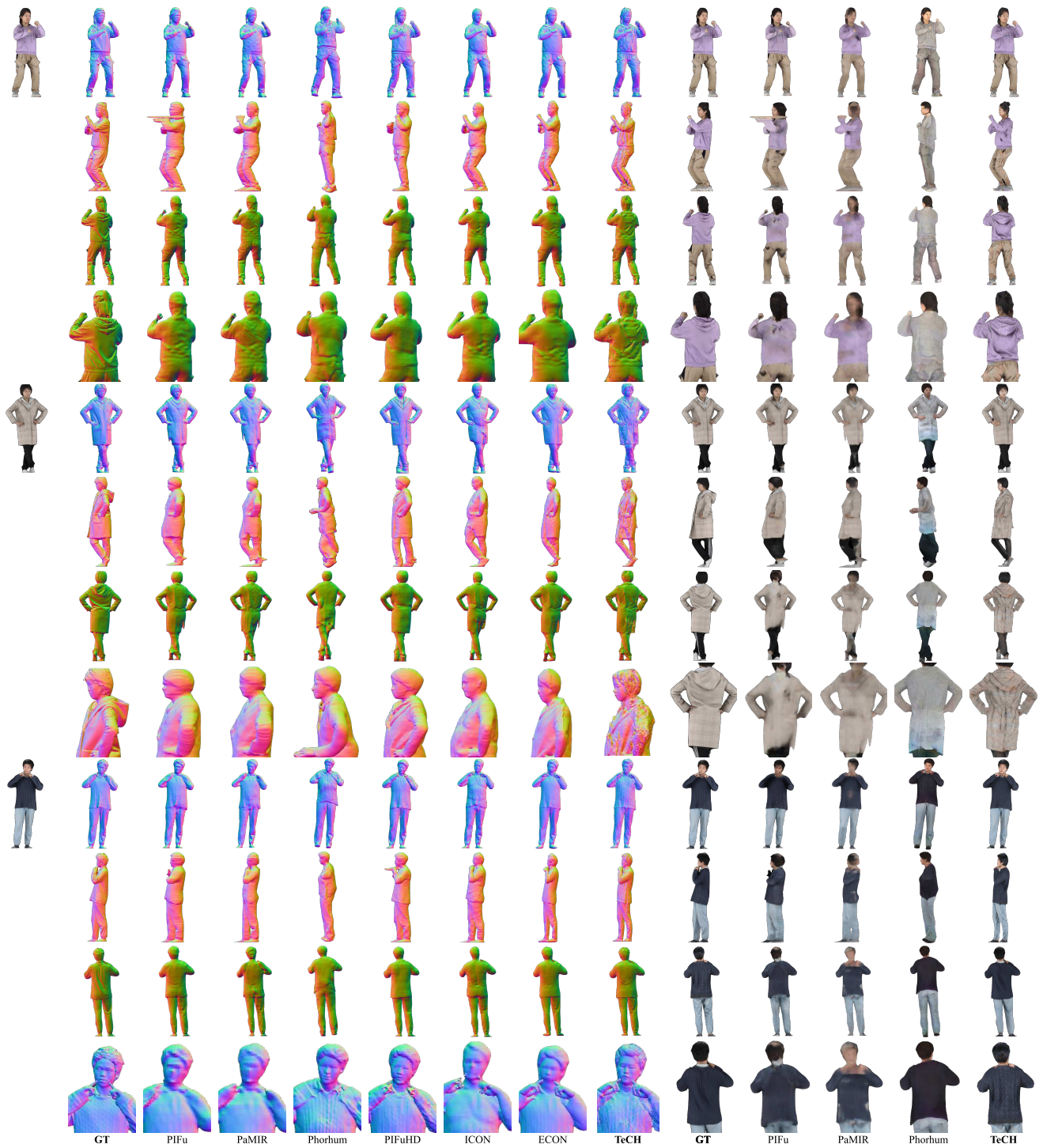
Figure 13. **Qualitative comparison on THuman2.0**. TeCH performs better regardless of hard pose, complex texture, or loose clothing.

Input	PIFu	PaMIR	PHORHUM	TeCH

Figure 14. **Qualitative comparison on SHHQ images.** TeCH generalizes well on in-the-wild images with diverse clothing styles and textures. It successfully recovers the overall structure of the clothed body with text guidance, and generates realistic full-body texture which is consistent with the colored pattern and the material of the clothes. **🔍 Zoom in** to see the geometric details.

References

[1] Thiemo Alldieck, Marcus A. Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *International Conference on 3D Vision (3DV)*, 2018. 2

[2] Thiemo Alldieck, Marcus A. Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] Thiemo Alldieck, Marcus A. Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2Shape: Detailed Full Human Body Geometry From a Single Image. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[5] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 7, 8, 9

[6] Rie Ando and Tong Zhang. Learning on graph with laplacian regularization. *Conference on Neural Information Processing Systems (NeurIPS)*, 2006. 6

[7] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-Image Diffusion Models with Ensemble of Expert Denoisers. *arXiv preprint:2211.01324*, 2022. 11

[8] Alexander Bergman, Petr Kellnhofer, Wang Yifan, Eric Chan, David Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3

[9] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-Garment Net: Learning to dress 3D people from images. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[10] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1209–1218, 2018. 8

[11] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[12] Yukang Cao, Guanying Chen, Kai Han, Wenqi Yang, and Kwan-Yee K. Wong. JIFF: Jointly-aligned Implicit Face Function for High Quality Single View Clothed Human Reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[13] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models. *arXiv preprint:2304.00916*, 2023. 3

[14] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In *arXiv*, 2023. 4

[15] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 8

[16] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. In *International Conference on Computer Vision (ICCV)*, 2023. 6

[17] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. gDNA: Towards generative detailed neural avatars. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[18] Wei Cheng, Ruixiang Chen, Wanqi Yin, Siming Fan, Keyu Chen, Honglin He, Huiwen Luo, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. DNA-Rendering: A Diverse Neural Actor Repository for High-Fidelity Human-centric Rendering. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[19] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, pages 20–40, 2020. 8

[20] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset. In *CVPR Workshop on the Future of Datasets in Vision*. sn, 2015. 8

[21] Enric Corona, Mihai Zanfir, Thiemo Alldieck, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Structured 3d features for reconstructing relightable and animatable avatars. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[22] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. NeRDi: Single-View NeRF Synthesis with Language-Guided Diffusion as General Image Priors. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009. 8

[24] Akio Doi and Akio Koide. An efficient method of triangulating equi-valued surfaces by using tetrahedral cells. *IEICE TRANSACTIONS on Information and Systems*, 74(1): 214–224, 1991. 5, 6, 11

[25] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. AG3D: Learning to Gen-

erate 3D Avatars from 2D Image Collections. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[26] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, pages 792–804, 2021. 2, 5, 8, 10

[27] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 10

[28] 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. *European Conference on Computer Vision (ECCV)*, 2022. 3, 8, 9

[29] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3D human shape estimation from single images. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[30] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations (ICLR)*, 2023. 4

[31] Daiheng Gao, Yuliang Xiu, Kailin Li, Lixin Yang, Feng Wang, Peng Zhang, Bang Zhang, Cewu Lu, and Ping Tan. DART: Articulated Hand Model with Diverse Accessories and Rich Textures. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 10

[32] Jun Gao, Wenzheng Chen, Tommy Xiang, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Learning deformable tetrahedral meshes for 3d reconstruction. *Conference on Neural Information Processing Systems (NeurIPS)*, 33:9936–9947, 2020. 2, 5, 11

[33] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. GET3D: A Generative Model of High Quality 3D Textured Shapes Learned from Images. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3

[34] Yuying Ge, Ruimao Zhang, Lingyun Wu, Xiaogang Wang, Xiaoou Tang, and Ping Luo. A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[35] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[36] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. Stylepeople: A generative model of fullbody human avatars. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5151–5160, 2021. 3

[37] Si Hang. Tetgen, a delaunay-based quality tetrahedral mesh generator. *ACM Trans. Math. Softw*, 41(2):11, 2015. 12

[38] Tong He, John P. Collomosse, Hailin Jin, and Stefano Soatto. Geo-PIFu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2

[39] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. ARCH++: Animation-Ready Clothed Human Reconstruction Revisited. In *International Conference on Computer Vision (ICCV)*, pages 11046–11056, 2021. 2, 3

[40] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *Transactions on Graphics (TOG)*, 2022. 3

[41] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. EVA3D: Compositional 3D Human Generation from 2D Image Collections. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[42] Hugues Hoppe. New quadric metric for simplifying meshes with appearance attributes. In *Proceedings Visualization'99 (Cat. No. 99CB37067)*, pages 59–510. IEEE, 1999. 12

[43] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[44] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. DreamWaltz: Make a Scene with Complex 3D Animatable Avatars. *arXiv preprint:2305.12529*, 2023. 3

[45] Yangyi Huang, Hongwei Yi, Weiyang Liu, Haofan Wang, Boxi Wu, Wenxiao Wang, Binbin Lin, Debing Zhang, and Deng Cai. One-shot implicit animatable avatars with model-based priors. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[46] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable Reconstruction of Clothed Humans. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3093–3102, 2020. 2, 3

[47] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. *Transactions on Graphics (TOG)*, 2023. 3

[48] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis. In *International Conference on Computer Vision (ICCV)*, pages 5885–5894, 2021. 3

[49] 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)*, 2020. 2

[50] Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[51] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[52] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. HumanSD: A Native Skeleton-Guided Diffusion Model for Human Image Generation. In *International Conference on Computer Vision (ICCV)*, 2023. 10

[53] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 2

[54] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[55] Byungjun Kim, Patrick Kwon, Kwangho Lee, Myunggi Lee, Sookwan Han, Daesik Kim, and Hanbyul Joo. Chupa: Carving 3D Clothed Humans from Skinned Shape Priors using 2D Diffusion Probabilistic Models. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[56] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5252–5262, 2020. 2

[57] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021.

[58] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *International Conference on Computer Vision (ICCV)*, pages 11035–11045, 2021.

[59] 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 *International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019. 2

[60] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. DreamHuman: Animatable 3D Avatars from Text. *arXiv preprint:2306.09329*, 2023. 3

[61] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123:32–73, 2017. 8

[62] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *Transactions on Graphics (TOG)*, 39(6), 2020. 5, 11

[63] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-Degree textures of people in clothing from a single image. In *International Conference on 3D Vision (3DV)*, 2019. 2

[64] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3383–3393, 2021. 2

[65] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pages 12888–12900. PMLR, 2022. 2, 4, 8, 11

[66] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. NIKI: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[67] Ruilong Li, Kyle Olszewski, Yuliang Xiu, Shunsuke Saito, Zeng Huang, and Hao Li. Volumetric human teleportation. In *ACM SIGGRAPH 2020 Real-Time Live*, 2020. 2

[68] 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 *European Conference on Computer Vision (ECCV)*, pages 590–606. Springer, 2022. 2

[69] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. Deep human parsing with active template regression. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(12):2402–2414, 2015. 8

[70] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. Deep human parsing with active template regression. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(12):2402–2414, 2015. 4

[71] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *International Conference on Computer Vision (ICCV)*, pages 1386–1394, 2015. 4

[72] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[73] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3D: High-Resolution Text-to-3D Content Creation. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 11

[74] 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 *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 8

[75] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund T, Zexiang Xu, and Hao Su. One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization. *arXiv preprint*, 2023. 4

[76] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot One Image to 3D Object. In *International Conference on Computer Vision (ICCV)*, 2023. 4

[77] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[78] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015. 2, 3, 7

[79] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 8

[80] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of Motion Capture as Surface Shapes. In *International Conference on Computer Vision (ICCV)*, pages 5442–5451, 2019. 10

[81] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. RealFusion: 360 Reconstruction of Any Object from a Single Image. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[82] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint:2302.08453*, 2023. 10

[83] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 5

[84] Edwin G. Ng, Bo Pang, Piyush Kumar Sharma, and Radu Soricut. Understanding guided image captioning performance across domains. In *Conference on Computational Natural Language Learning*, 2020. 8

[85] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *European Conference on Computer Vision (ECCV)*, pages 597–614. Springer, 2022. 3

[86] Hayato Onizuka, Zehra Haiyrci, Diego Thomas, Akihiro Sugimoto, Hideaki Uchiyama, and Rin-Ichiro Taniguchi. TetraTSDF: 3D human reconstruction from a single image with a tetrahedral outer shell. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[87] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2011. 8

[88] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. NPMs: Neural Parametric Models for 3D Deformable Shapes. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[89] Pablo Palafox, Nikolaos Sarafianos, Tony Tung, and Angela Dai. Spams: Structured implicit parametric models. *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[90] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference (BMVC)*, 2015. 8

[91] Priyanka Patel, Chun-Hao Paul Huang, Joachim Tesch, David Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, 2021. 8

[92] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 3, 7

[93] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. ClothCap: Seamless 4D Clothing Capture and Retargeting. *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 36(4), 2017. Two first authors contributed equally. 2

[94] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 3, 4, 6, 7, 11, 12

[95] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors. *arXiv preprint:2306.17843*, 2023. 4

[96] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling Text-to-Image Diffusion by Orthogonal Finetuning. *arXiv preprint:2306.07280*, 2023. 10

[97] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 3

[98] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Ben Mildenhall, Nataniel Ruiz, Shiran Zada, Kfir Aberman, Michael Rubenstein, Jonathan Barron, Yuanzhen Li, and Varun Jampani. DreamBooth3D: Subject-Driven Text-to-3D Generation. *arXiv preprint: 2303.13508*, 2023. 4

[99] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *International Conference on Machine Learning (ICML)*, 2021. 11

[100] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 11

[101] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 4, 11

[102] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2, 11

[103] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019. 2, 3, 7, 8, 9

[104] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *Computer Vision and Pattern Recognition (CVPR)*, pages 81–90, 2020. 2, 7, 9

[105] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 3, 8

[106] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Conference on Neural Information Processing Systems (NeurIPS)*, 34: 6087–6101, 2021. 2, 5, 11

[107] Vanessa Sklyarova, Jenya Chelishev, Andreea Dogaru, Igor Medvedev, Victor Lempitsky, and Egor Zakharov. Neural Haircut: Prior-Guided Strand-Based Hair Reconstruction. In *International Conference on Computer Vision (ICCV)*, 2023. 10

[108] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. FACSIMILE: Fast and accurate scans from an image in less than a second. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[109] Jiang Suyi, Jiang Haoran, Wang Ziyu, Luo Haimin, Chen Wenzheng, and Xu Lan. HumanGen: Generating Human Radiance Fields with Explicit Priors. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[110] David Svitov, Dmitrii Gudkov, Renat Bashirov, and Victor Lemptisky. Dinar: Diffusion inpainting of neural textures for one-shot human avatars. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[111] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-It-3D: High-Fidelity 3D Creation from A Single Image with Diffusion Prior. In *International Conference on Computer Vision (ICCV)*, 2023. 4

[112] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[113] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[114] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel View Synthesis with Diffusion Models (3DiM). In *International Conference on Learning Representations (ICLR)*, 2023. 4

[115] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10957–10966, 2019. 8

[116] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica K. Hodgins. MonoClothCap: Towards temporally coherent clothing capture from monocular RGB video. In *International Conference on 3D Vision (3DV)*, 2020. 2

[117] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Conference on Neural Information Processing Systems (NeurIPS)*, 34:12077–12090, 2021. 2, 4, 8, 11

[118] Zhangyang Xiong, Di Kang, Derong Jin, Weikai Chen, Linchao Bao, and Xiaoguang Han. Get3DHuman: Lifting StyleGAN-Human into a 3D Generative Model using Pixel-aligned Reconstruction Priors. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[119] Yuliang Xiu, Ruilong Li, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *European Conference on Computer Vision (ECCV)*, pages 49–67, 2020. 2

[120] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6, 7, 8, 9

[121] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 6, 7, 8, 9

[122] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. NeuralLift-360: Lifting An In-the-wild 2D Photo to A 3D Object with 360° Views. *Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[123] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6183–6192, 2020. 2, 3

[124] Xueting Yang, Yihao Luo, Yuliang Xiu, Wei Wang, Hao Xu, and Zhaoxin Fan. D-IF: Uncertainty-aware Human

Digitization via Implicit Distribution Field. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[125] Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J. Black. Human-Aware Object Placement for Visual Environment Reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 6

[126] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 8

[127] Ilya Zakharkin, Kirill Mazur, Artur Grigorev, and Victor Lempitsky. Point-based modeling of human clothing. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[128] Yifei Zeng, Yuanxun Lu, Xinya Ji, Yao Yao, Hao Zhu, and Xun Cao. AvatarBooth: High-Quality and Customizable 3D Human Avatar Generation. *arXiv preprint:2306.09864*, 2023. 3

[129] 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 *International Conference on Computer Vision (ICCV)*, 2021. 2

[130] Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Kang Du, and Min Zheng. AvatarVerse: High-quality & Stable 3D Avatar Creation from Text and Pose. *arXiv preprint:2308.03610*, 2023. 3

[131] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. PyMAF-X: Towards Well-aligned Full-body Model Regression from Monocular Images. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 2

[132] Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. Avatargen: a 3d generative model for animatable human avatars. In *European Conference on Computer Vision Workshops (ECCVw)*, pages 668–685. Springer, 2023. 3

[133] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3D Human-Object Spatial Arrangements from a Single Image in the Wild. In *European Conference on Computer Vision (ECCV)*, pages 34–51, Cham, 2020. Springer International Publishing. 6

[134] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint:2302.05543*, 2023. 10

[135] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. DeepHuman: 3D Human Reconstruction From a Single Image. In *International Conference on Computer Vision (ICCV)*, pages 7738–7748, 2019. 3

[136] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. PaMIR: Parametric Model-conditioned Implicit Representation for image-based human reconstruction. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44 (6):3170–3184, 2021. 2, 3, 7, 8, 9

[137] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Computer Vision and Pattern Recognition (CVPR)*, pages 633–641, 2017. 8

[138] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[139] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[140] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *International Conference on Computer Vision (ICCV)*, 2019. 8

21