SIFU: Side-view Conditioned Implicit Function for Real-world Usable Clothed Human Reconstruction

Zechuan Zhang   Zongxin Yang†   Yi Yang
ReLER, CCAI, Zhejiang University
{zechuan, yangzongxin, yangyics}@zju.edu.cn

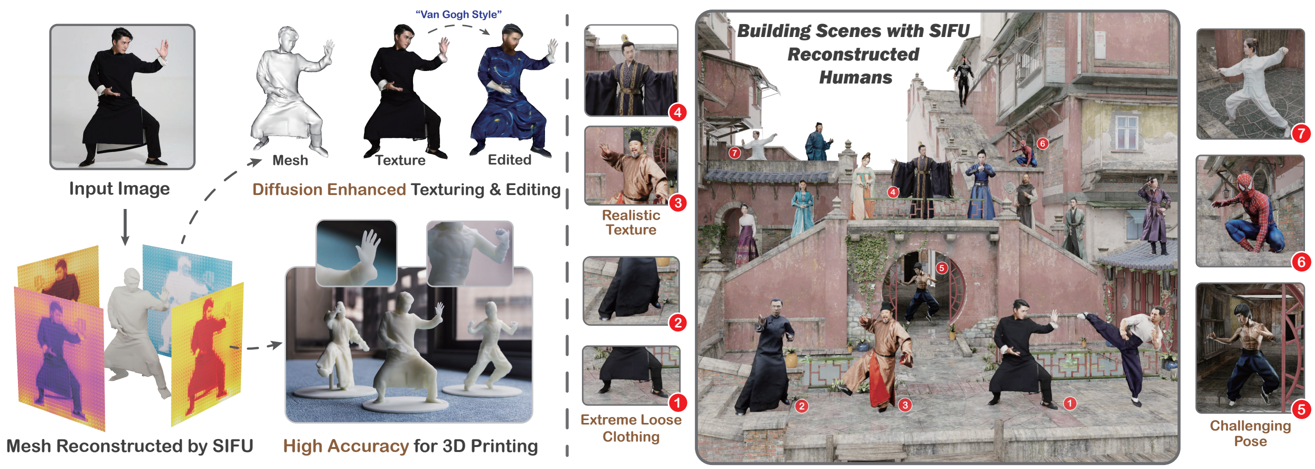


Figure 1. With just a single image, SIFU is capable of reconstructing a high-quality 3D clothed human model, making it well-suited for practical applications such as 3D printing and scene creation. At the heart of SIFU is a novel **Side-view Conditioned Implicit Function**, which is key to enhancing feature extraction and geometric precision. Furthermore, SIFU introduces a **3D Consistent Texture Refinement** process, greatly improving texture quality and facilitating texture editing with the help of text-to-image diffusion models. Notably proficient in dealing with complex poses and loose clothing, SIFU stands out as an ideal solution for real-world applications.

Abstract

*Creating high-quality 3D models of clothed humans from single images for real-world applications is crucial. Despite recent advancements, accurately reconstructing humans in complex poses or with loose clothing from in-the-wild images, along with predicting textures for unseen areas, remains a significant challenge. A key limitation of previous methods is their insufficient prior guidance in transitioning from 2D to 3D and in texture prediction. In response, we introduce **SIFU** (**S**ide-view Conditioned **I**mplicit **F**unction for Real-world **U**sable Clothed Human Reconstruction), a novel approach combining a Side-view Decoupling Transformer with a 3D Consistent Texture Refinement pipeline. SIFU employs a cross-attention mechanism within the transformer, using SMPL-X normals as queries to effectively decouple side-view features in the process of mapping 2D features to 3D. This method not only improves the precision of the 3D models but also their robustness, especially when SMPL-X estimates are not perfect. Our texture refinement process leverages text-to-image diffusion-based prior to generate realistic and consistent textures for invisible views. Through extensive experiments, SIFU surpasses SOTA methods in both geometry and texture reconstruction, showcasing enhanced robustness in complex scenarios and achieving an unprecedented Chamfer and P2S measurement. Our approach extends to practical applications such as 3D printing and scene building, demonstrating its broad utility in real-world scenarios.*

1. Introduction

High-quality 3D models of clothed humans are crucial in diverse sectors, including augmented and virtual reality (AR/VR), 3D printing, scene assembly, and filmmaking. The traditional process of creating these models not only

†Zongxin Yang is the corresponding author.

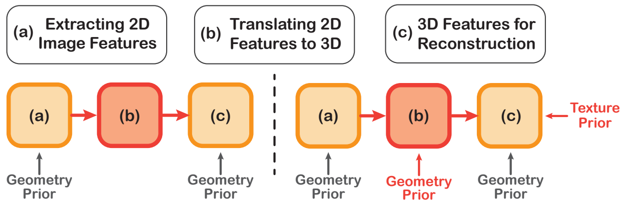


Figure 2. **Contrast between previous methods (Left) and ours (Right):** Our approach improves the reconstruction process by incorporating additional guidance on geometry and texture priors.

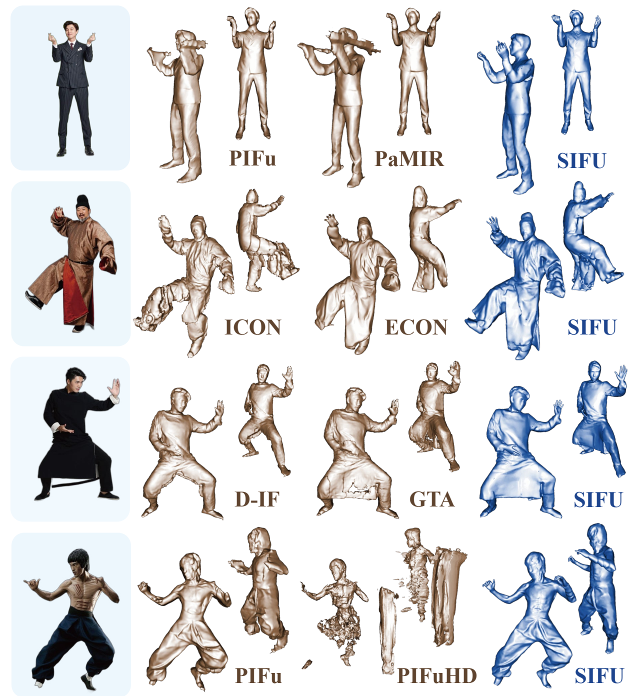


Figure 3. **Comparison of SIFU with State-of-the-Art (SOTA) Methods in 3D Human Inference from In-the-Wild Images.** Existing SOTA methods often struggle with complex poses and loose clothing, leading to a range of artifacts. These issues include the absence of human shapes (PIFu, PaMIR, PIFuHD), missing body parts (ECON), disrupted clothing (ICON, D-IF), and a lack of fine details (GTA). In contrast, SIFU effectively addresses these challenges, delivering high-quality, detailed results.

requires a considerable amount of time but also specialized equipment capable of capturing multi-view photographs, in addition to the reliance on skilled artists. Contrasting this, in everyday situations, we most often have access to monocular images of individuals, easily obtained through phone cameras or found on various web pages. Thus, a method that accurately reconstructs 3D human models from a single image could significantly cut costs and simplify the process of independent creation. While existing deep learning models [6, 11, 30, 69, 70, 81, 82, 85, 92, 93] show promise in this area, they struggle with complex poses and loose clothing, as illustrated in Fig. 3. Furthermore, these models fail to correctly texture hidden areas, resulting in less realistic outcomes. Therefore, there's a significant need for models that can generalize across various scenarios and efficiently produce realistic, real-world applicable 3D clothed humans.

Through analyzing existing methods, we pinpointed two key challenges in this field: **(i) Insufficient Prior Guidance in Translating 2D Features to 3D:** The reconstruction of 3D objects from 2D images typically involves three main steps: (a) *extracting 2D image features*, (b) *translating 2D features to 3D*, and (c) *3D features for reconstruction*. As shown by Fig. 2, current approaches often add geometric prior (like SMPL-X [61]) to the first and last steps, focusing on techniques such as normal map prediction [81, 82, 85], SMPL-guided SDF [81, 85, 92], or volume features [10, 93]. While the use of priors for improving the transition from 2D image features to 3D is crucial, it remains underexplored. Currently, this transition is typically achieved by projecting features onto 3D points [6, 10, 11, 69, 70, 81, 85, 93] or by employing fixed learnable embeddings to generate 3D features [92]. These methods, however, do not fully harness the potential of priors in enhancing accuracy of 3D reconstruction. **(ii) Lack of Texture Prior:** While current methods [6, 11, 69, 70, 92] attempt to predict vertex colors, they struggle to accurately predict textures for unseen views, particularly with limited training data. This limitation highlights a need for additional texture priors in 3D human reconstruction.

In response to the challenges we've identified, we propose two refined strategies to enhance 3D human reconstruction. **Firstly**, we believe that enhancing the process of translating 2D features to 3D with additional guidance could significantly improve both the accuracy and efficiency of 3D reconstructions. To more effectively integrate prior guidance, such as SMPL-X [61], with image features, we utilize the cross-attention mechanism of the transformer [77]. This approach aims to optimize the fusion of geometry and image data, potentially leading to more precise and realistic 3D human models. **Secondly**, considering the impressive generative capabilities of pretrained diffusion models, as shown in recent studies [12, 13, 25, 57, 73] and their proficiency in learning rich 3D priors [48–51, 64, 65, 76], we suggest their incorporation as priors to enhance texture prediction, particularly for invisible regions. Besides, maintaining 3D consistency from different angles and matching the style of the input image is also crucial for creating realistic textures.

In this paper, we present **SIFU** (**S**ide-view Conditioned **I**mplicit **F**unction for Real-world **U**sable Clothed Human Reconstruction), a novel approach employing a **Side-view Conditioned Implicit Function** (§3.2) with a **3D Consistent Texture Refinement** (§3.3) pipeline for precise geometry and realistic texture reconstruction. Our approach employs normals from SMPL-X as queries in a cross-attention mechanism with image features. This method effectively decouples side-view features in the process of mapping 2D features to 3D, thereby enhancing the accuracy and robustness of reconstruction. Moreover, our texture refinement employs text-to-image diffusion models [68] and ensures uniform diffusion features across different perspectives, resulting in detailed, consistently styled textures.

Through extensive experiments, **SIFU** outperforms existing SOTA methods in geometry and texture quality, achieving an unprecedented Chamfer and P2S measurement of **0.6 cm** on THuman2.0 [87] (Tab. 1). Additionally, SIFU shows improved robustness in geometry reconstruction (Tab. 2), even with inaccurate SMPL-X estimations. SIFU handles complex poses and loose clothing well, producing realistic textures with consistent colors and patterns (Fig. 7). Its adaptability extends to practical applications like 3D printing and scene creation (Fig. 1), showcasing its broad practical utility. Key contributions include:

- A novel **Side-view Conditioned Implicit Function** that skillfully maps 2D image features to 3D with SMPL-X guidance. This is the first instance showcasing the efficacy of using human prior information to decouple side-view 3D features from the input image, significantly advancing the field of clothed human reconstruction.
- A **3D Consistent Texture Refinement** pipeline designed to generate realistic, 3D consistent textures on clothed human meshes. This approach has notably improved the quality and uniformity of textures, offering a substantial advancement in the field.
- Our proposed model achieves state-of-the-art performance in both geometry and texture reconstruction, facilitating **real-world applications** such as 3D printing and scene building, which were challenging to achieve with previous methods.

2. Related Work

Implicit-function-based Reconstruction. Implicit representations, such as occupancy and signed distance fields, are flexible with topology and can effectively depict 3D clothed humans across a variety of scenarios, including loose garments and complex poses. A series of studies have focused on regressing the implicit surface from a single input image directly in a streamlined process [1, 6, 21, 69, 70]. Others incorporate a 3D human body prior to enhance the process of 2D feature extraction and 3D feature for reconstruction [9–11, 23, 24, 29, 31, 47, 81, 82, 85, 92, 93]. Among these, GTA [92] utilizes transformers with fixed learnable embeddings to translating image features to 3D tri-plane features. As for texture reconstruction, methods like PIFu [69], ARCH [24, 31], PaMIR [93], and GTA [92] deduce full textures from a single image. Techniques such as PHORHUM [6] and S3F [11] go further by segregating albedo and global illumination. Nevertheless, these methods lack information from other views or prior knowledge (such as diffusion models), resulting in unsatisfactory textures. HumanSGD [1] employs diffusion models for mesh inpainting but faces performance declines with mesh reconstruction inaccuracies. TeCH [29] uses diffusion-based models for visualizing unseen areas, yielding realistic results. Its limitations, however, include time-intensive per-subject optimization and dependence on accurate SMPL-X.

Explicit-shape-based Reconstruction. Recovering the human mesh from a single RGB image is a complex challenge that has received extensive attention. Many approaches [15, 16, 37–41, 43, 44, 46, 90, 91] adopt parametric body models [36, 52, 60, 83] to estimate the shape and pose of a 3D human body with minimal clothing [45, 71]. To incorporate clothing into the 3D models, methods often apply 3D clothing offsets [2–5, 42, 80, 94] or use adjustable garment templates [8, 16, 32] over the base body shape. Additionally, non-parametric forms like depth maps [17, 72], normal maps [82], and point clouds [89] are explored for creating representations of clothed humans.

Despite these advancements, explicit-shape approaches can be limited by topological constraints, which become apparent when handling diverse and complex clothing styles found in real-world settings, such as dresses, and skirts.

NeRF-based Reconstruction. The rise of Neural Radiance Fields (NeRF) has seen methods [18, 20, 33, 34, 56, 59, 62, 63, 78, 88] using videos or multi-view images to optimize NeRF for human form capture. Recent advancements like SHERF [26] and ELICIT [28] aim to generate human NeRFs from single images, with SHERF filling gaps using 2D image data and ELICIT employing a pre-trained CLIP model [66] for contextual understanding. While NeRF-based approaches are effective in creating quality images from various perspectives, they typically struggle with detailed 3D mesh generation from single images and often require extensive time for optimization.

Contrasting with these methods, SIFU stands out in reconstructing clothed human meshes across various scenarios, producing consistently realistic 3D textures suitable for real-world use. It leverages human body priors to decouple side-view features from input images during the 2D to 3D mapping process, thereby improving the accuracy of its implicit function. For texture refinement, SIFU adopts a coarse-to-fine approach, utilizing a pre-trained diffusion model, trained on a vast dataset, to predict textures in unseen areas. It also reconstructs texture from the input image for visible regions, ensuring uniform texture consistency.

3. Method

Given a single image, SIFU first reconstructs the 3D mesh and coarse textures using the Side-view Conditioned Implicit Function (Sec. 3.2). Subsequently, it employs a 3D Consistent Texture Refinement process (Sec. 3.3) to enhance textures, ensuring high quality and 3D consistency.

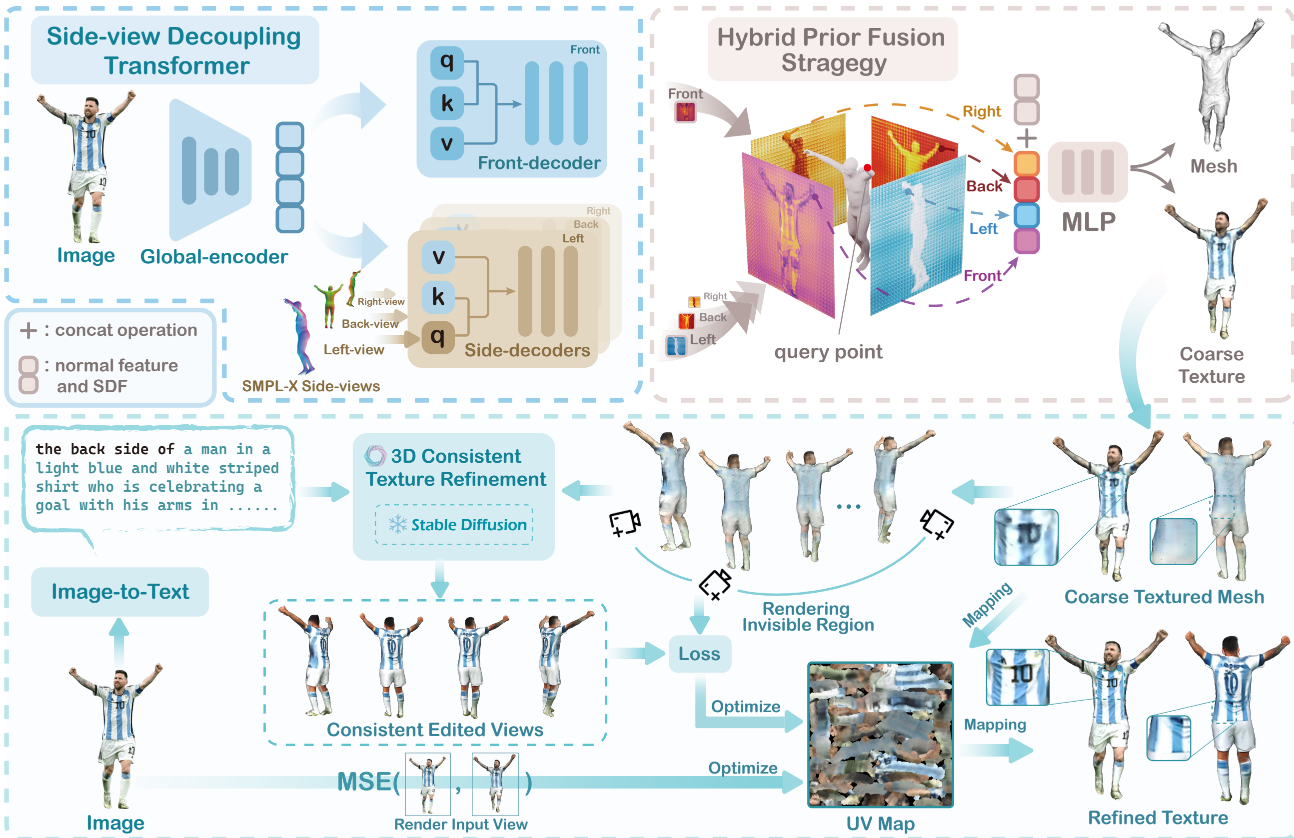


Figure 4. Given a single image, SIFU constructs a 3D clothed human mesh with coarse textures using a **Side-view Conditioned Implicit Function** (§3.2). This is followed by a step of **3D Consistent Texture Refinement** (§3.3) to generate detailed textures. Specifically, SIFU employs a side-view decoupling transformer to decouple features from the input image and the side-view normals of the SMPL-X model. Then, these features are combined at a query point through a hybrid prior fusion strategy, aiding in the reconstruction of both the mesh and its texture. Finally, the mesh with its basic textures undergoes a diffusion-based 3D consistent texture refinement, ensuring feature consistency in the latent space and resulting in high-quality textures.

Key preliminary concepts necessary for understanding our approach are briefly presented in Sec. 3.1.

3.1. Preliminary

Implicit Function is a powerful tool for modeling complex geometries and colors with neural networks. We employ implicit function to predict an occupancy field to represent 3D clothed humans. Specifically, our implicit function $\mathcal{IF}$ maps an input point $\boldsymbol{x}$ to a scalar value representing the spatial field including occupancy and color fields. Our reconstructed human surface can be represented as $\mathcal{S}_{\mathcal{IF}}$:

$$\mathcal{S}_{\mathcal{IF}} = \{\boldsymbol{x} \in \mathbb{R}^3 \mid \mathcal{IF}(\boldsymbol{x}) = (\boldsymbol{o}, \boldsymbol{c})\} \tag{1}$$

where occupancy $\boldsymbol{o} = 0.5$ and color $\boldsymbol{c} \in \mathbb{R}^3$.

SMPL and SMPL-X. The Skinned Multi-Person Linear (SMPL) model [52] is a parametric model for human body representation. It uses shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ and pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{3\times 24}$ to define the human body mesh $\mathcal{M}$:

$$\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta}) : \boldsymbol{\beta} \times \boldsymbol{\theta} \mapsto \mathbb{R}^{3\times 6890} \tag{2}$$

Here, $\boldsymbol{\beta}$ controls body size, while $\boldsymbol{\theta}$ affects joint positions and orientations. The SMPL-X model [61] builds upon SMPL, adding features for hands and face, enhancing facial expressions, finger movements, and detailed body poses.

Diffusion Models. Diffusion processes, notably represented by Diffusion Probabilistic Models (DPM) [12, 13, 25, 57, 73], are pivotal in image generation and have shown capabilities in human/avatar generation [27, 84]. These models aim to approximate a data distribution q through a progressive denoising process. Starting with a Gaussian i.i.d noisy image $\boldsymbol{x}_T \sim \mathcal{N}(0, I)$, the model denoises it until a clean image $\boldsymbol{x}_0$ from the target distribution q is obtained. DPMs can also learn a conditional distribution with additional guiding signals like text conditioning.

3.2. Side-view Conditioned Implicit Function

The Side-view Conditioned Implicit Function in our model comprises two key components: the **Side-view Decoupling Transformer** and the **Hybrid Prior Fusion Strategy**. The transformer initially uses rendered SMPL-X images from various side views as queries to perform cross-attention with the encoded input image. This process effectively decouples features conditioned on the side views. The Hybrid Prior Fusion Strategy then integrates these features at each query point, which are later input into a Multi-Layer Perceptron (MLP) for predicting occupancy and color. We detail both components in the sections below.

Side-view Decoupling Transformer. Our method draws inspiration from the shared characteristics, such as material and color, between side views (like the back or left side) and the visible front view. Despite their different perspectives, these similarities in features are crucial. Therefore, we aim to effectively separate side-view features from the front view, utilizing the SMPL-X model [61] as a guide.

The process begins with a ViT-based global encoder [14], which encodes the input image I into a latent feature h, capturing the image's globally correlated features. To decode these features, we employ two decoders: a front-view decoder, aligned with h, and a side-view decoder. The front-view decoder utilizes multi-head self-attention within a vision transformer to process the front view feature, represented as $F_{front} \in \mathbb{R}^{H\times W\times C}$.

To decouple side-view features, we render the side-view normal images N_i of SMPL-X as guidance, with $i \in \{left,\ back,\ right\}$ during the experiments. The side-view normals N_i are transformed to embeddings z_i, which then engage in a cross-attention operation as queries, with the latent feature h acting as both keys and values:

$$\mathbf{CrossAttn}(\boldsymbol{z_i}, \boldsymbol{h}) = \mathbf{SM}(\frac{(W^Q\boldsymbol{z_i})(W^K\boldsymbol{h})^T)}{\sqrt{d}})(W^V\boldsymbol{h}) \quad (3)$$

where $\mathbf{SM}$ represents $\mathbf{SoftMax}$ operation, while W^Q, W^K, and W^V are learnable parameters and d is the scaling coefficient. Following the original transformer architecture [77], our model integrates residual connections [22] and layer normalization [7] after each sub-layer. The entire side-decoder contains multiple identical layers, and we deploy three such decoders to yield feature maps $F_i \in \mathbb{R}^{H\times W\times C}$ where $i \in \{left,\ back,\ right\}$.

Hybrid Prior Fusion Strategy. In our pipeline, we incorporate the Hybrid Prior Fusion Strategy from [92] to effectively merge features at a query point, utilizing both spatial localization and human body prior knowledge. We split the feature maps F_j (for $j \in \{front, left, back, right\}$) into two groups. For the spatial query group, we project query points onto the feature map to obtain pixel-aligned features F_j^S. We then combine these features from all planes using a mix of averaging and concatenation:

$$F^S(\boldsymbol{x}) = F_f^S(\boldsymbol{x}) \oplus \mathbf{avg}(F_l^S(\boldsymbol{x}), F_b^S(\boldsymbol{x}), F_r^S(\boldsymbol{x})) \quad (4)$$

where $f,\ l,\ b,\ r$ denote the front, left, back, and right respectively. For the other group, similar to the spatial query, we project the SMPL-X [61] mesh vertices onto the four feature maps, obtaining the feature $F^S(\boldsymbol{v})$, $\boldsymbol{v} \in \mathcal{M}$, where $\mathcal{M}$ is the SMPL-X mesh. For each query point $\boldsymbol{x}$, we find its nearest triangular face $t_{\boldsymbol{x}} = [\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2] \in \mathbb{R}^{3\times 3}$ and employ barycentric interpolation to integrate features for $\boldsymbol{x}$, denoted as $F^P(\boldsymbol{x})$:

$$F^P(\boldsymbol{x}) = uF^S(\boldsymbol{v}_0) + vF^S(\boldsymbol{v}_1) + wF^S(\boldsymbol{v}_2) \quad (5)$$

where $[u, v, w]$ represents the barycentric coordinates of the query point $\boldsymbol{x}$ projected onto triangle $t_{\boldsymbol{x}}$. We concatenate these two query features as the final point feature. Moreover, we incorporate the signed distance between the query point and SMPL-X mesh $\mathcal{SDF}(\boldsymbol{x})$ and pixel-aligned normal feature $F^N(\boldsymbol{x})$ as input to a Multilayer Perceptron (MLP) for prediction of occupancy and color:

$$(\boldsymbol{o}, \boldsymbol{c}) = \mathbf{MLP}(F^S(\boldsymbol{x}), F^P(\boldsymbol{x}), \mathcal{SDF}(\boldsymbol{x}), F^N(\boldsymbol{x})) \quad (6)$$

Training Objectives. We consider two sets of points as training data, denoted as G_o and G_c. G_c is sampled uniformly with a slight perturbation along the normals of the ground-truth mesh surface, whereas G_o is sampled according to the same strategy as in [69]. For the points in G_o, we employ the following loss function:

$$\mathcal{L}_o = \frac{1}{|G_o|}\sum_{\boldsymbol{x}\in G_o} BCE(\hat{o}_{\boldsymbol{x}} - o_{\boldsymbol{x}}) \quad (7)$$

where $\hat{o}_{\boldsymbol{x}}$ denotes the model's predicted occupancy, while $o_{\boldsymbol{x}}$ is the ground-truth occupancy. For the sampled points in G_c, we apply the following loss function:

$$\mathcal{L}_c = \frac{1}{|G_c|}\sum_{\boldsymbol{x}\in G_c} |\hat{\boldsymbol{c}}_{\boldsymbol{x}} - \boldsymbol{c}_{\boldsymbol{x}}| \quad (8)$$

where $\hat{\boldsymbol{c}}_{\boldsymbol{x}}$ denotes the predicted color, and $\boldsymbol{c}_{\boldsymbol{x}}$ represents the corresponding ground-truth color. The total loss is the sum of these two separate losses, which is designed to fulfill a comprehensive training objective.

Mesh Extraction. We begin by densely sampling points in space and using our side-view conditioned implicit function to predict their occupancy values. The Marching Cubes algorithm [53] is then applied to extract the mesh, and following [82], we substitute the hands with SMPL-X models for enhanced visuals. Finally, these mesh points are processed through the implicit function again for color prediction.

3.3. 3D Consistent Texture Refinement

Upon extracting the mesh using our implicit function, we noted that color quality was coarse and areas not visible in the input were blurry, leading to a less realistic look (see Fig. 4). To address this, we developed a **3D Consistent Texture Refinement** pipeline, leveraging text-to-image diffusion priors to substantially enhance texture quality.

Pipeline. For a given input image and its reconstructed mesh M, we first utilize vision-to-text models (*e.g.*, [54, 58,

Method	Publication	CAPE-NFP Chamfer ↓	CAPE-NFP P2S ↓	CAPE-NFP Normal ↓	CAPE-FP Chamfer ↓	CAPE-FP P2S ↓	CAPE-FP Normal ↓	THuman2.0 Chamfer ↓	THuman2.0 P2S ↓	THuman2.0 Normal ↓
w/o SMPL-X body prior										
PIFu * [69]	ICCV 2019	2.5609	1.9971	0.1023	1.8139	1.5108	0.0798	1.5991	1.4333	0.0843
PIFuHD [70]	CVPR 2020	3.7670	3.5910	0.1230	2.3020	2.3350	0.0900	-	-	-
w/ SMPL-X body prior										
PaMIR * [93]	TPAMI 2021	1.6313	1.2666	0.0730	1.481	1.1631	0.0727	1.2152	1.0582	0.0730
ICON [81]	CVPR 2022	0.8846	0.8569	0.0434	0.7247	0.6979	0.0371	0.9491	0.9846	0.0621
ECON [82]	CVPR 2023	0.9462	0.9334	0.0382	0.9039	0.8938	0.0373	1.2585	1.4184	0.0612
D-IF [85]	ICCV 2023	0.8237	0.8353	0.0575	0.7625	0.769	0.0503	1.1696	1.2900	0.0936
GTA [92]	NeurIPS 2023	0.8508	0.7920	0.0424	0.6525	0.6084	0.0349	0.7329	0.7297	0.0492
Ours	-	**0.7725**	**0.7354**	**0.0378**	**0.6297**	**0.5980**	**0.0327**	**0.5961**	**0.6058**	**0.0407**

Table 1. **Quantitative evaluation against SOTA (§4.1).** All models use a resolution of 256 for marching cubes and ground-truth SMPL-X models are used during testing. *Methods are re-implemented in [81] for a fair comparison. Top two results are colored as first second.

86]) to convert the image into a textual description P, and then back-project the mesh color onto a UV texture map T, following the approach in [75]. To visualize unseen mesh areas, differentiable rendering $\mathcal{I}$ is employed on mesh M, generating images of these invisible views:

$$\boldsymbol{I} = \mathcal{I}(T, M, \boldsymbol{k}) \tag{9}$$

where $\boldsymbol{k} = \{k^1, ..., k^n\}$ represent camera views and $\boldsymbol{I} = \{I^1, ..., I^n\}$ are the corresponding rendered images.

Subsequently, a pretrained and fixed text-to-image diffusion model $\boldsymbol{\epsilon_\theta}$ refines the blurry images $\boldsymbol{I}$ into enhanced images $\boldsymbol{J}$, using P as a condition. To ensure consistency among refined images, a **consistent editing** technique $\mathcal{H}$ is applied to $\boldsymbol{\epsilon_\theta}$, preserving the original semantic layout of $\boldsymbol{I}$:

$$\boldsymbol{J} = \mathcal{H}(\boldsymbol{\epsilon_\theta}, P, \boldsymbol{I}) = \mathcal{H}(\boldsymbol{\epsilon_\theta}, P, \mathcal{I}(T, M, \boldsymbol{k})) \tag{10}$$

where $\boldsymbol{J} = \{J^1, ..., J^n\}$ corresponds to the refined views of $\boldsymbol{I}$. After obtaining $\boldsymbol{J}$, a pixel-wise Mean Squared Error (MSE) loss is computed between each J^i and I^i to optimize the texture map T. Additional losses include a perceptual loss $\mathcal{L}_{vgg}$ [35] and a Chamfer Distance loss $\mathcal{L}_{CD}$ [29], aimed at ensuring style similarity between $\mathcal{J}$ and the input image. We also compute an MSE loss $\mathcal{L}^f_{MSE}$ from the input view against the input image. These combined losses jointly optimize T, enhancing overall texture quality:

$$\min_T \lambda_1 \mathcal{L}_{MSE} + \lambda_2 \mathcal{L}_{vgg} + \lambda_3 \mathcal{L}_{CD} + \lambda_4 \mathcal{L}^f_{MSE} \tag{11}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are the weights attributed to each loss.

Consistent Editing. To achieve consistent image editing across different views, we adopt a method inspired by [19]. This involves enforcing consistency among diffusion features from various rendered views. We perform DDIM inversion [74] on the input image $\boldsymbol{I}$, extracting diffusion tokens across all layers. A set of key views is selected for joint editing [79], ensuring a unified appearance in the resultant features. These features are then propagated to all views using a nearest-neighbor approach to maintain coherence across them. Please refer to the SupMat for more detailed procedural insights and specific mechanisms.

Method	Backbone	Chamfer ↓	P2S ↓	Normal ↓
PaMIR [93]	CNN	1.3224	1.1349	0.0767
ICON [81]	CNN	1.2935	1.3949	0.0781
D-IF [85]	CNN	1.5262	1.7296	0.1191
ECON [82]	-	2.1195	1.8074	0.1029
GTA [92]	Transformer	1.0473	1.0780	0.0649
Ours	Transformer	**0.9937**	**1.0645**	**0.0599**

Table 2. **Assessing model robustness to SMPL-X (§4.1).** To evaluate the models' robustness in reconstruction, we used the THuman2.0 dataset [87] and introduced random noise to the ground-truth SMPL-X models. This approach simulates inaccuracies in poses and shapes for robustness testing.

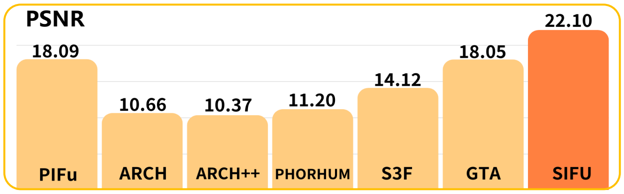
4. Experiment

Datasets. We trained our model on the THuman2.0 dataset [87], comprising 526 human scans, with 490 used for training, 15 for validation, and 21 for testing. Ground-truth SMPL-X models were used during training, and PIXIE [15] was employed for inference. Our main evaluations were conducted on the CAPE [55] and THuman2.0 datasets. To test our model's versatility with different poses, we divided the CAPE dataset into "CAPE-FP" and "CAPE-NFP" subsets. Further details on datasets and implementation are available in the SupMat.

4.1. Evaluation

Metrics. Our model's reconstruction quality for geometry is quantitatively evaluated using Chamfer and P2S distances, comparing reconstructed meshes with ground-truth. We also measure L2 Normal error between normal images from both meshes, assessing surface detail consistency by rotating the camera at $\{0°, 90°, 180°, 270°\}$ relative to the input view. For texture quality, we report the PSNR on colored images rendered similarly to normal images.

Quantitative Evaluation. In geometry evaluation, our experiments utilize the ground-truth SMPL-X model for methods using a "SMPL-X body prior," as shown in Tab. 1. SIFU establishes a new standard in all metrics, especially excelling on the THuman2.0 dataset with an unprecedented Chamfer and P2S of **0.6 cm**. This highlights SIFU's proficiency in accurate reconstructions across diverse scenarios,



(a) Quantitative comparison of texture quality on THuman2.0 [87].

(b) Qualitative results on Thuman2.0 [87]

Figure 5. **Texture comparison against SOTAs (§4.1).** We quantitatively and qualitatively compare texture quality on THuman2.0 [87]. PIXIE [15] used for SMPL-X estimation during testing. Please **zoom in** for details.

benefiting from our side-view conditioned approach.

For texture reconstruction, SIFU surpasses PIFu [69] by **22.2%** in PSNR, demonstrating its superior texture quality. For visual comparisons, refer to Fig. 5 and the SupMat.

Robustness to SMPL-X. In real-world scenarios, encountering in-the-wild images lacking precise SMPL-X parameters is common. The ability to handle SMPL-X estimation errors is crucial for high-quality reconstructions. We evaluated our model's resilience by introducing noise (scaled by 0.05) to the pose and shape parameters of the ground-truth SMPL-X models. As shown in Tab. 2, SIFU demonstrates significant robustness, indicating strong practical utility.

Qualitative Results. Our results showcase the model's strong performance on in-the-wild images. As depicted in Fig. 7, our model is capable of handling complex scenarios such as loose clothing and challenging poses with proficiency. Further examples are provided in the SupMat.

4.2. Ablation Studies

Different Backbone Analysis. In validating the effectiveness of our side-view decoupling transformer, we experimented with various alternative architectures. As per the results in Tab. 3, self-attention and learnable embeddings, without SMPL-X guidance, led to significant errors, and even convolutional networks with similar capacities were unable to effectively link input images with SMPL-X conditioned views. This ablation study clearly demonstrates that our custom transformer architecture excels, delivering superior reconstruction results.

Different Feature Plane Analysis. In assessing the effect

Method	Chamfer ↓	P2S ↓	Normal ↓
A - Different Backbone			
no cross-attention	0.9846	0.8672	0.0477
learnable embedding	0.9860	0.8538	0.0471
use convolution network	0.8699	0.8221	0.0387
B - Different Feature Plane			
only front plane	1.1165	0.9574	0.0558
front and back planes	0.9929	0.9189	0.0464
w/o left plane	0.7941	0.7576	0.0387
w/o right plane	0.8058	0.7671	0.0386
C - Different Query Strategy			
pixel-aligned	0.8111	0.7615	0.0400
Ours	**0.7725**	**0.7354**	**0.0378**

Table 3. **Ablation study (§4.2).** We quantitatively evaluate the contribution of each component in our model. The evaluation is performed on the CAPE-NFP dataset, with ground-truth SMPL-X models provided during the testing phase.

Figure 6. **Ablation on texture refinement (§4.2).** We compare our 3D consistent texture refinement with other diffusion-based methods on in-the-wild images. Please **zoom in** to see details.

of various numbers of side-view feature planes, we found, as shown in Tab. 3, that adding just the left or right side-view planes most significantly improved accuracy, reducing the Chamfer by about 0.2 cm. The inclusion of all four planes offered a smaller error reduction, approximately 0.03 cm. Considering the minor improvements from more planes against the added complexity, we chose a balanced approach with four planes, as shown in Fig. 4.

Query Strategy Efficacy. We compared the hybrid prior fusion strategy with the pixel-aligned method [69, 70, 81]. As shown in Tab. 3, the hybrid approach consistently outperforms the conventional method in all evaluation metrics.

Different Texture Refinement. In comparing our approach with diffusion-based methods like TEXTure [67] and DreamGaussian [75] (using Zero123 XL [50]), and also against our model without refinement, it is evident from Fig. 6 that our 3D Consistent Texture Refinement method excels in both texture quality and consistency.

4.3. Applications

Texture Editing. With the powerful ability of text-to-image diffusion models, we can change the text prompt to easily generate edited textures in the 3D consistent texture refinement. The edited results are shown in Fig. 1 and Fig. 8.

Scene Building and 3D Printing. The model's accurate ge-

Figure 7. **Qualitative results on in-the-Wild images (§4.1):** The first two rows present results for humans wearing loose clothing, and the subsequent two rows display outcomes for humans in challenging poses. (**Zoom in** for detailed view)

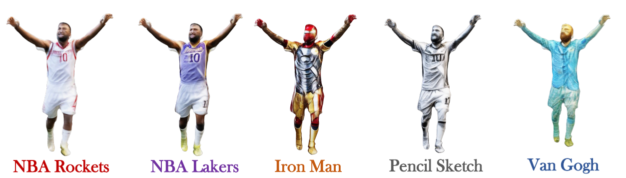


Figure 8. **Texture editing (§4.3).** We edit the texture of the individual in Fig. 4 to achieve diverse outcomes by changing the text prompt in our 3D Consistent Texture Refinement.

ometry and refined textures make it ideal for virtual scene creation and 3D printing (see Figs. 1 and 9 and SupMat). It enhances realism in simulations and games and streamlines the 3D printing process, reducing the need for complex scanning. This has potential applications in rapid prototyping, educational resources, and custom 3D figurines.

5. Conclusion

We introduce SIFU, a novel method for reconstructing high-quality 3D clothed human meshes, complete with detailed textures. Our method employs SMPL-X normals [61] as queries in a cross-attention mechanism with image features, efficiently decoupling side-view features during the conversion of 2D features to 3D. This process significantly improves geometric accuracy and robustness in our 3D reconstructions. Moreover, we design a 3D Consistent Texture

Figure 9. **Building scenes with SIFU reconstructed humans (§4.3).** We showcase examples of building impressive scenes with SIFU reconstructed humans. Please **zoom in** to see details.

Refinement process, which employs text-to-image diffusion priors while maintaining consistency among diffusion features in the latent space. This innovative approach ensures the creation of realistic textures, particularly in regions that are not visible in the initial input. SIFU distinctly outperforms existing methods in terms of both geometric and textural fidelity, showcasing exceptional capabilities in handling complex poses and loose clothing. These qualities make SIFU highly suitable for real-world applications.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (U2336212) and the Fundamental Research Funds for the Central Universities (No. 226-2022-00051).

References

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

[2] 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. 3

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

[4] 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 *CVPR*, 2019.

[5] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2Shape: Detailed Full Human Body Geometry From a Single Image. In *ICCV*, 2019. 3

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

[7] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

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

[9] 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 *CVPR*, 2022. 3

[10] Yukang Cao, Kai Han, and Kwan-Yee K. Wong. Sesdf: Self-evolved signed distance field for implicit 3d clothed human reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

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

[12] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 4

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 4

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[15] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, pages 792–804. IEEE, 2021. 3, 6, 7

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

[17] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3D human shape estimation from single images. In *ICCV*, 2019. 3

[18] Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *CVPR*, 2023. 3

[19] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 6

[20] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[21] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[23] Tong He, John P. Collomosse, Hailin Jin, and Stefano Soatto. Geo-PIFu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *NeurIPS*, 2020. 3

[24] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. ARCH++: Animation-Ready Clothed Human Reconstruction Revisited. In *ICCV*, pages 11046–11056, 2021. 3

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4

[26] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *ICCV*, 2023. 3

[27] Shuo Huang, Zongxin Yang, Liangting Li, Yi Yang, and Jia Jia. Avatarfusion: Zero-shot generation of clothing-decoupled 3d avatars using 2d diffusion. In *ACM MM*, pages 5734–5745, 2023. 4

[28] 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 *IEEE Conference on Computer Vision (ICCV)*, 2023. 3

[29] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *International Conference on 3D Vision (3DV)*, 2024. 3, 6

[30] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable Reconstruction of Clothed Humans. In *CVPR*, pages 8568–8576, 2020. 2

[31] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable Reconstruction of Clothed Humans. In *CVPR*, pages 3093–3102, 2020. 3

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

[33] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. *arXiv*, 2022. 3

[34] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 3

[35] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155, 2016. 6

[36] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 3

[37] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 3

[38] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, pages 5252–5262, 2020.

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

[40] 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 *ICCV*, pages 11035–11045, 2021.

[41] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, pages 2252–2261, 2019. 3

[42] 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. 3

[43] 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 *CVPR*, pages 3383–3393, 2021. 3

[44] 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 *CVPR*, 2023. 3

[45] Jiahao Li, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Jotr: 3d joint contrastive learning with transformers for occluded human mesh recovery. In *ICCV*, pages 9110–9121, 2023. 3

[46] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation. In *ECCV*, pages 590–606. Springer, 2022. 3

[47] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. High-Fidelity Clothed Avatar Reconstruction from a Single Image. In *CVPR*, 2023. 3

[48] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[49] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization, 2023.

[50] 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, 2023. 7

[51] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Learning to generate multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2

[52] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 34(6):248:1–248:16, 2015. 3, 4

[53] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 5

[54] Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reliable video narrator via equal distance to visual tokens. In *CVPR*, 2024. 5

[55] 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. 6

[56] Jiteng Mu, Shen Sang, Nuno Vasconcelos, and Xiaolong Wang. ActorsNeRF: animatable few-shot human rendering with generalizable nerfs. pages 18391–18401, 2023. 3

[57] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2, 4

[58] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 5

[59] Xiao Pan, Zongxin Yang, Jianxin Ma, Chang Zhou, and Yi Yang. Transhuman: A transformer-based human representation for generalizable neural human rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3544–3555, 2023. 3

[60] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 3

[61] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face,

and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 4, 5, 8

[62] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14314–14323, 2021. 3

[63] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9054–9063, 2021. 3

[64] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 2

[65] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 2

[66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 2021. 3

[67] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023. 7

[68] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[69] 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 *ICCV*, pages 2304–2314, 2019. 2, 3, 5, 6, 7

[70] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *CVPR*, pages 81–90, 2020. 2, 3, 6, 7

[71] Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Global-to-local modeling for video-based 3d human pose and shape estimation. In *CVPR*, pages 8887–8896, 2023. 3

[72] 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 *ICCV*, 2019. 3

[73] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2, 4

[74] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 6

[75] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 6, 7

[76] 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. *arXiv preprint arXiv:2303.14184*, 2023. 2

[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 5

[78] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 3

[79] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 6

[80] 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. 3

[81] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *CVPR*, 2022. 2, 3, 6, 7

[82] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *CVPR*, 2023. 2, 3, 5, 6

[83] 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 *CVPR*, pages 6183–6192, 2020. 3

[84] Yuanyou Xu, Zongxin Yang, and Yi Yang. Seeavatar: Photorealistic text-to-3d avatar generation with constrained geometry and appearance. *arXiv preprint arXiv:2312.08889*, 2023. 4

[85] Xueting Yang, Yihao Luo, Yuliang Xiu, Wei Wang, Hao Xu, and Zhaoxin Fan. D-if: Uncertainty-aware human digitization via implicit distribution field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9122–9132, 2023. 2, 3, 6

[86] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models. *arXiv preprint arXiv:2401.08392*, 2024. 6

[87] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In

IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021), 2021. 3, 6, 7

[88] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. *CVPR*, 2023. 3

[89] Ilya Zakharkin, Kirill Mazur, Artur Grigorev, and Victor Lempitsky. Point-based modeling of human clothing. In *ICCV*, 2021. 3

[90] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop. In *ICCV*, 2021. 3

[91] 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. *IEEE TPAMI*, 2023. 3

[92] Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3, 5, 6

[93] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. PaMIR: Parametric Model-conditioned Implicit Representation for image-based human reconstruction. *IEEE TPAMI*, 44(6):3170–3184, 2021. 2, 3, 6

[94] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *CVPR*, 2019. 3