Diffusion-FOF: Single-view Clothed Human Reconstruction via Diffusion-based Fourier Occupancy Field

Yuanzhen Li, Fei Luo*, Chunxia Xiao*
School of Computer Science, Wuhan University, China
yuanzhen@whu.edu.cn, luofei@whu.edu.cn, cxxiao@whu.edu.cn

Abstract

Reconstructing a clothed human from a single-view image has several challenging issues, including flexibly representing various body shapes and poses, estimating complete 3D geometry and consistent texture, and achieving more fine-grained details. To address them, we propose a new diffusion-based Fourier occupancy field method to improve the human representing ability and the geometry generating ability. First, we estimate the back-view image from the given reference image by incorporating a style consistency constraint. Then, we extract multi-scale features of the two images as conditional and design a diffusion model to generate the Fourier occupancy field in the wavelet domain. We refine the initial estimated Fourier occupancy field with image features as conditions to improve the geometry accuracy. Finally, the reference and estimated back-view images are mapped onto the human model, creating a textured clothed human model. Substantial experiments are conducted, and the experimental results show that our method outperforms the state-of-the-art methods in geometry and texture reconstruction performance.

1. Introduction

3D human reconstruction has wide applications in education, entertainment, and AR/VR [4, 5, 20]. Traditional methods can reconstruct a human's geometry and texture by extracting adequate information from multi-view images [10, 23, 40]. However, multiple-view images for a human are not available in many scenarios. For example, photos of people in the smartphone's album are usually shot on the single-view. Reconstructing a clothed human from a single-view image is a challenging task [24, 27]. Several issues need to be further dealt with, including flexibly representing various shapes and poses, robustly estimating complete 3D geometry and consistent texture appearance, and achieving more fine-grained details.

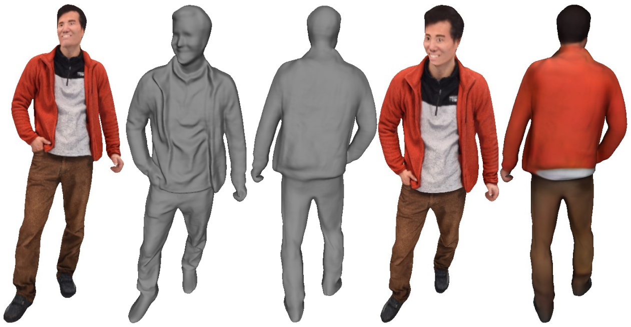


Figure 1. The 3D reconstruction example of our method for a single-view clothed human image.

Properly representing a human is one important factor, as it determines the scale of body shapes and poses that can be reconstructed. The widely used parametric model SMPL [21] is based on skinning and blend shapes and is learned from thousands of 3D body scans. Researchers introduced clothing displacement to SMPL to reconstruct clothing details [1, 2]. These methods have limitations in reconstructing loose-fitting clothes such as dresses and robes. The visual hull [24], the depth [11, 19], and the voxel [35, 46] have been also explored. They require significant computational resources. Besides explicit representation methods, researchers proposed implicit neural representation methods [27, 28], using the network to predict the geometry and texture of spatial sample points. Recently, Feng et al. [9] proposed a novel 3D geometry representation called Fourier occupancy field (FOF), representing 3D geometry as a multi-channel image. However, FOF directly utilizes a 2D Convolutional Neural Network (CNN) to estimate the occupancy field and loses high-frequency information. Such weakness may lead to geometry distortion or overly smoothing.

The powerful generative model is another important factor in determining how well the reconstruction results can approximate real humans. Diffusion model [8] has substantially succeeded in several computer vision tasks

*Chunxia Xiao and Fei Luo are co-corresponding authors

[26, 33, 43]. We propose to combine the generation capability of the image-conditioned diffusion model and the flexibility of the Fourier occupancy field representation. In particular, we introduce the wavelet mechanism to handle the weaknesses of the diffusion model and the Fourier occupancy field. Briefly, the wavelet transform decomposes the high-frequency and low-frequency information, facilitating learning intricate details. The spatial dimensions of wavelet subbands are 1/4 of the original image, reducing computational resources and inference time.

We propose a novel single-view clothed human reconstruction method via a diffusion-based Fourier occupancy field. It takes several steps to realize reconstruction. First, we estimate the back-view image to provide more human prior knowledge. Current methods have not considered the texture consistency between the back-view and reference images. When wearing a black jacket with a white T-shirt, the clothing on the back view will likely be black. Thus, we introduce a style consistency constraint between the predicted back-view and reference images. Then, we extract multi-scale features of the reference image and the estimated back-view image. We design a diffusion model to generate the Fourier occupancy field in the wavelet domain based on the image features as conditional. Meanwhile, we refine the initial estimated Fourier occupancy field with image features as conditions to improve the geometry accuracy. Finally, the reference and estimated back-view images are mapped onto the model, producing the final textured clothed human model. Extensive comparative experiments and ablation studies verify the effectiveness of our method.

In summary, the main contributions of our method are summarized as follows:

- We propose a wavelet-based diffusion model to reconstruct 3D clothed human from a single image. The wavelet transform facilitates the explicit learning of detailed information and reduces computational time.
- During the back-view image prediction, we introduce a style consistency constraint between the predicted back-view image and the reference image to enhance the style consistency of the texture.
- In the geometry prediction, we add the predicted back-view image as conditional, providing more human prior information to the geometry prediction network.

2. Related Work

Single-view human reconstruction. Only with 2D information of a single-view image, reconstructing the 3D human is inherently ill-posed. Some works introduced additional assumptions or prior knowledge. Previous research has proposed effective parametric models [3, 21] of the human body, which utilizes statistical methods to reduce the variations in human body shape and pose to a compact set of parameters. With the rapid development of deep learning, methods [17, 18] attempt to estimate and regress the model parameters from a single image by deep neural networks. Human-parameterized models lack 3D details of clothes, hairstyles, and adornments. To generate the clothing details, methods [1,2] added offsets on the top of parameterized model vertices. Although these methods can reconstruct certain clothing details, they often face challenges when dealing with loose-fitting garments, such as robes and dresses.

To address the constraints of loose-fitting clothing reconstruction, various 3D human representations have been explored, including the visual hull [24], the double depth maps [11, 13], and the voxel [35]. However, these methods need high memory requirements, limiting the spatial resolution of shape estimation.

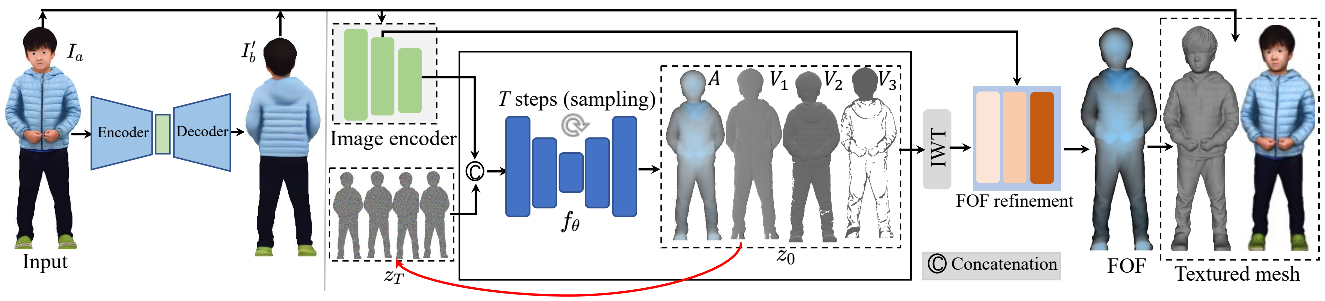
To reconstruct high-resolution 3D clothed humans, Saito et al. [27] proposed a pixel-aligned implicit function for reconstructing 3D human geometry and texture called PIFu. They used neural networks to extract 2D image features and designed a deep implicit function to estimate the geometry and texture of 3D sampling points. PIFuHD [28] designed a coarse-to-fine framework to reconstruct high-resolution 3D human geometry with the normal as prior knowledge. These two methods could generate distorted shapes in complex poses due to the lack of regularization. To improve geometry stability, methods [39, 41, 45] utilized human parameterized models as prior information to guide implicit function learning. However, inaccuracies in parameterized model estimation would decrease reconstruction accuracy.

Recently, Feng et al. [9] proposed a novel 3D representation, the Fourier occupancy field, which converts the 3D occupancy value into a 2D vector field using Fourier series expansion. Different from [9], we aim to generate the Fourier occupancy field from the view of generative modeling implemented by a wavelet-based diffusion model.

Denoising diffusion model. The Diffusion model is a generative model that employs an iterative denoising process to generate high-quality results of target tasks [14]. It has demonstrated superior performance in various computer vision tasks. In 2D computer vision applications, it has been used for tasks such as image generation [33, 34], inpainting [25], and super-resolution [26]. In 2.5D computer vision tasks, the diffusion model has been applied to monocular depth estimation [15, 29] and depth refinement [30]. In the realm of 3D content generation, it finds application in tasks such as 3D shape generation [6, 47] and completion [7, 48].

3. Method

The pipeline of our method is shown in Figure 2. Given a single human image I_a, we first predict the back-view image I'_b. Then, we design a wavelet-based diffusion model



(a) Back-view image prediction

(b) Mesh reconstruction

Figure 2. Overview of our method. Given the human image I_a, we first predict the back-view image I'_b. Then, we design a wavelet-based diffusion model to generate the Fourier occupancy field based on the two images as conditional. We transform the Fourier occupancy field into the 3D mesh. Finally, the two images I_a and I'_b are projected back to the 3D mesh, generating the textured model.

to generate the Fourier occupancy field based on the two images as conditional. We transform the Fourier occupancy field into occupancy values and use the marching cubes [22] algorithm to generate the 3D mesh. Finally, the two images I_a and I'_b are mapped onto the human model through rasterization, creating the textured model.

3.1. Preliminaries

Fourier occupancy field. It encodes the 3D object geometry to a 2D vector field [9]. Each pixel (x, y) on the image I corresponds to a line in 3D space and can be used as an occupancy function $f(\mathrm{z})$ represent:

$$f(\mathrm{z}) = \begin{cases} 1, & (\mathrm{x}, \mathrm{y}, \mathrm{z}) \text{ is inside the object} \\ 0.5, & (\mathrm{x}, \mathrm{y}, \mathrm{z}) \in S \\ 0, & (\mathrm{x}, \mathrm{y}, \mathrm{z}) \text{ is outside the object} \end{cases} \tag{1}$$

where S represents the surface of the 3D object. The occupancy function $f(\mathrm{z})$ satisfies the Dirichlet conditions and can be expanded as a convergent Fourier series:

$$f(\mathrm{z}) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos\left(nz\pi\right) + b_n \sin\left(nz\pi\right)\right). \tag{2}$$

where $\{a_n\}$ and $\{b_n\}$ are coefficients of basis functions $\{\cos(nx)\}$ and $\{\sin(nx)\}$, respectively. Approximate $f(\mathrm{z})$ by a subspace spanned by the first $2N+1$ basis functions:

$$f(\mathrm{z}) = b^\top(\mathrm{z})c, \tag{3}$$

where $b(\mathrm{z}) = [1/2, \cos(\mathrm{z}), \sin(\mathrm{z}), ..., \cos(N\mathrm{z}), \sin(N\mathrm{z})]^\top$ is the vector of the first $2N+1$ basis functions spanning the approximation subspace, and $c = [a_0, a_1, b_1, ..., a_N, b_N]^\top$ is the $2N+1$ Fourier coefficient vectors. This procedure can be extended to cover the entire xy-plane: $F(\mathrm{x}, \mathrm{y}, \mathrm{z}) = b^\top(\mathrm{z})C(\mathrm{x}, \mathrm{y})$, and C called Fourier occupancy field (FOF). In our experiment, N is settled as 15.

When FOF is estimated, the approximate occupancy field can be extracted according to Eq. 3. Then, we can use the marching cubes algorithm [22] to extract the 3D mesh from the occupancy field.

Wavelet transform. It is a classic technique extensively used in image compression. It is applied to separate the original image's low-frequency approximation and high-frequency details. The low sub-band corresponds to a down-sampled version resembling the original image, and the high sub-bands capture local statistical information on vertical, horizontal, and diagonal edges. Haar wavelet [12] is a simple wavelet transform, including discrete wavelet transform (DWT) and inverse wavelet transform (IWT).

We use Haar wavelet to decompose the Fourier occupancy field $C \in \mathbb{R}^{H \times W \times (2N+1)}$ into four wavelet subbands $\{A, V_1, V_2, V_3\} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times (2N+1)}$, representing the average of the source image and high-frequency information in the vertical, horizontal, and diagonal directions, respectively.

Diffusion model. It is divided into forward diffusion and inverse denoising phases [14]. The forward process is fixed to a Markov chain that gradually adds Gaussian noise to the target training data $z_0 = z$:

$$q(z_{1:T}) = \textstyle\prod_{t=1}^{T} q(z_t|z_{t-1}), \tag{4}$$

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t\mathbf{I}), \tag{5}$$

where β_t is a variance schedule that increases from $\beta_0 = 0$ to $\beta_T = 1$ and controls how much noise is added in each step. We can obtain the sampling z_t at even arbitrary time step t from z_0:

$$q(z_t|z_0) = \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t}z_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{6}$$

where $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s = \prod_{s=0}^{t}(1-\beta_s)$. The reverse process aims to derive the posterior distribution for the less noisy image z_{t-1} given the more noisy image z_t using the denoising network f_θ:

$$p_\theta(z_{0:T}) = p(z_T)\textstyle\prod_{t=1}^{T} p_\theta(z_{t-1}|z_t), \tag{7}$$

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \sigma_t^2\mathbf{I}), \tag{8}$$

where $p(z_T) = \mathcal{N}(z_T; \mathbf{0}, \mathbf{I})$ is the random sampling of Gaussian noise, $\mu_\theta(z_t, t)$ and σ_t^2 are the mean and variance of the parametric denoising model, respectively. The objective is to minimize the distance between a true denoising distribution $q(z_t|z_{t-1})$ and the parameterized $p_\theta(z_{t-1}|z_t)$ through Kullback-Leibler (KL) divergence. DDPM [14] utilizes the network to predict the noise ϵ.

3.2. Back-view Image Prediction

In the back-view image prediction, we propose a style consistency constraint between the predicted and reference images. We utilize a Siamese network training strategy to train the back-view image prediction network.

Given the human image I_a, we use a deep neural network to estimate the back-view image I_b'. We train the network using the L1 loss between the estimated image I_b' and the ground truth I_b:

$$\mathcal{L}_1(I_b', I_b) = \|I_b' - I_b\|_1 \,. \tag{9}$$

There is a significant correlation between the back-view and front clothes of the human. For example, when a person wears a black jacket with a white T-shirt, the clothing on the back side will be black. Consequently, the predicted back-view image I_b' and the reference image I_a have consistent style (texture). Inspired by the image style transfer [16], we propose a style consistency constraint between the predicted back-view image I_b' and the reference image I_a:

$$\mathcal{L}_s(I_b', I_a) = \|G(\Phi_1(I_b')) - G(\Phi_1(I_a))\|_2^2 \,, \tag{10}$$

where $G(.)$ denotes the Gram matrix, and Φ_1 represents the latent space feature extracted from the first layer of the pre-trained VGG19 model [31], which is effective at capturing low-level color features of the reconstructed texture.

From the above analysis, the predicted back-view image I_b' correlates with the ground truth I_b and the reference image I_a. To enhance the effectiveness of network training, we adopt a Siamese network training strategy to train the back-view image prediction network.

The training framework is shown in Figure 3. The twin networks have the same parameters and sharing weights. We input a pair of images I_a and I_b into the twin networks and output their back-view image I_b' and I_a', respectively. We establish constraints between the estimated and input images to train the network.

In summary, we train the back-view image prediction network with the following loss:

$$\begin{aligned} \mathcal{L}_{\text{color}} = & \lambda_1(\mathcal{L}_1(I_b', I_b) + \mathcal{L}_1(I_a', I_a)) + \\ & \lambda_2(\mathcal{L}_s(I_b', I_a) + \mathcal{L}_s(I_a', I_b)), \end{aligned} \tag{11}$$

where the hyper-parameters $\lambda_1 = 1.0$ and $\lambda_2 = 0.8$ control the relative weights of the different terms. We use HR-Net [36] as the back-view image prediction network.

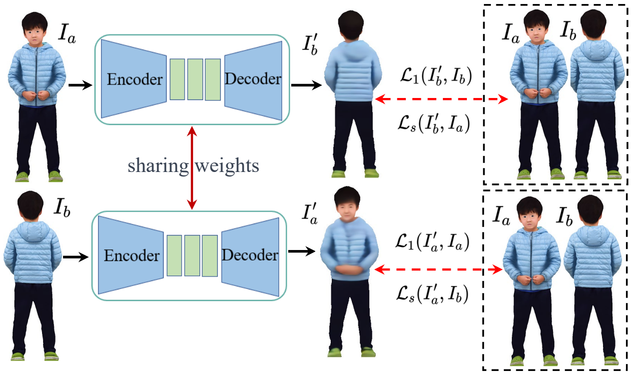

I_a
Encoder
Decoder
I_b'
I_a
I_b
$\mathcal{L}_1(I_b', I_b)$
$\mathcal{L}_s(I_b', I_a)$
sharing weights
I_b
Encoder
Decoder
I_a'
I_a
I_b
$\mathcal{L}_1(I_a', I_a)$
$\mathcal{L}_s(I_a', I_b)$

Figure 3. Training framework of the back-view image estimation.

3.3. Wavelet-based Diffusion FOF Prediction

The training framework is shown in Figure 4. We learn the conditional denoising process $p_\theta(z_{0:T}|z_0, x)$ with the reference image I_a and the predicted back-view image I_b' as the condition x. In our work, the Eq. 7 is modified to:

$$p_\theta(z_{0:T}|x) = p(z_T)\textstyle\prod_{t=1}^{T} p_\theta(z_{t-1}|z_t, x). \tag{12}$$

As our task involves optimizing Fourier occupancy fields, we train the denoising network f_θ to predict z_0 instead of the noise ϵ. Once trained, at generation time, the model f_θ can then approximate the mean $\mu_\theta(z_t, t)$ of the posterior $p_\theta(z_{t-1}|z_t, x)$ as:

$$\mu_\theta(z_t, t) \approx \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1-\alpha_t}{1-\bar{\alpha}_t}\left(z_t - \sqrt{\bar{\alpha}_t} f_\theta(z_t, t)\right)\right). \tag{13}$$

Thus, we can obtain the less noisy image z_{t-1} from the noisy image z_t by sampling the approximate posterior in each generation step.

We use DWT to decompose the FOF C into four wavelet sub-bands as the training data z_0 in the diffusion model. At even arbitrary time step t, the sampling z_t is constructed by adding Gaussian noise to the ground truth z_0 in the forward diffusion. The inverse denoising is to predict z_0 with the condition x. We estimate the initial FOF in the wavelet domain and further refine it in the pixel domain.

Firstly, we design an image encoder module to extract multi-scale feature representations from the reference images I_a and predicted back-view image I_b'. We use the low-resolution image feature as the condition x in the conditional diffusion model. Then, we concatenate z_t, x and t and feed them into the denoising network f_θ, outputs wavelet coefficient $\tilde{z}_0$. The wavelet coefficient $\tilde{z}_0$ is converted into the FOF C_{init} using IWT. Finally, We design a refinement module to refine the initial predicted FOF C_{init} with the high-resolution image feature as guidance and output a perfect FOF C_{refine}.

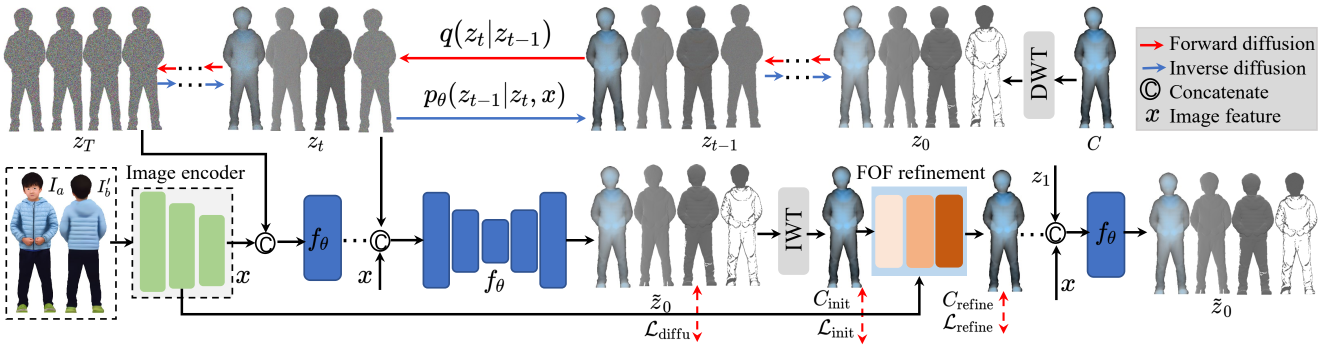


Figure 4. The training framework of the FOF prediction. FOF C is decomposed into sub-bands z_0 using DWT. At even time step t, z_t is obtained from z_0 through forward diffusion. Multi-scale features of the images I_a and I'_b are extracted using an image encoder network, with low-resolution features as condition x. The concatenation of z_t, x, and t is input into the denoising network f_θ, output $\hat{z}_0$. Subsequently, IWT is applied to $\hat{z}_0$ to convert it into FOF. Finally, a refinement module enhances the predicted FOF guided by high-resolution features. Three loss functions, $\mathcal{L}_{\text{diffu}}$, $\mathcal{L}_{\text{init}}$, and $\mathcal{L}_{\text{refine}}$, are designed to optimize the network.

We formulate three objective functions to train the model

$$\mathcal{L}_{\text{geometry}} = \mathcal{L}_{\text{diffu}}(\tilde{z}_0, z_0) + \mathcal{L}_{\text{init}}(C_{\text{init}}, C_{\text{gt}}) + \mathcal{L}_{\text{refine}}(C_{\text{refine}}, C_{\text{gt}}). \tag{14}$$

The loss $\mathcal{L}_{\text{diffu}}(\tilde{z}_0, z_0)$ represents the difference between the ground truth z_0 and predicted $\tilde{z}_0$:

$$\mathcal{L}_{\text{diffu}}(\tilde{z}_0, z_0) = \left\|\tilde{z}_0 - z_0\right\|_1. \tag{15}$$

The loss $\mathcal{L}_{\text{init}}(C_{\text{init}}, C_{\text{gt}})$ represents the difference between the ground truth $C_{\text{gt}}(C)$ and the initial predicted FOF C_{init}:

$$\mathcal{L}_{\text{init}}(C_{\text{init}}, C_{\text{gt}}) = \left\|C_{\text{init}} - C_{\text{gt}}\right\|_1. \tag{16}$$

The loss $\mathcal{L}_{\text{refine}}(C_{\text{refine}}, C_{\text{gt}})$ represents the difference between the ground truth C_{gt} and the refined FOF C_{refine}:

$$\mathcal{L}_{\text{refine}}(C_{\text{refine}}, C_{\text{gt}}) = \left\|C_{\text{refine}} - C_{\text{gt}}\right\|_1. \tag{17}$$

Figure 5 presents the details of the image encoder network and refinement network. CBAM [37] is a lightweight attention module. We use the U-Net architecture [8] as our denoise network f_θ.

In the inference process, given a human image and the predicted back-view image, we first extract the image feature as the condition. Then, we randomly sample the noisy map $\tilde{z}_T$ from the Gaussian distribution $\tilde{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively predict $\tilde{z}_0$ in the wavelet domain. We employ the DDIM update rule [32] for the sampling process. Finally, we use the inverse wavelet transform (IWT) to transform $\tilde{z}_0$ into the FOF and refine it with the image feature. We provide the pseudo-code of the geometry training and test procedures in the supplementary material.

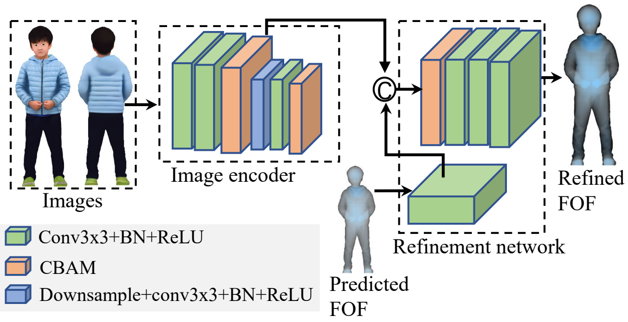


Figure 5. Image encoder network and refinement network.

Table 1. Quantitative comparison of texture on the 2K2K dataset. ↑ means the larger the better while ↓ means the smaller the better.

Methods	PSNR↑	SSIM↑	LPIPS↓
SiCloPe [24]	17.579	0.844	0.281
Ours	**26.362**	**0.929**	**0.142**
PIFu [27]	14.334	0.821	0.269
GTA [44]	18.352	0.842	0.236
Ours	**25.252**	**0.925**	**0.157**

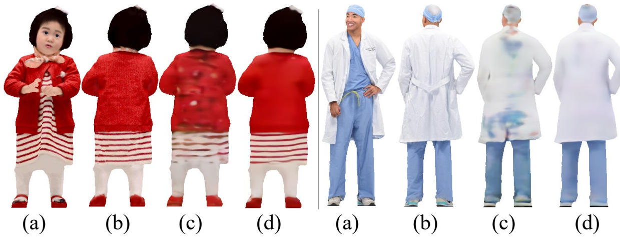


Figure 6. Qualitative comparison of back-view image estimation. (a) Input image. (b) Ground Truth. (c) SiCloPe [24]. (d) Ours.

4. Experiments

4.1. Experimental Settings

Datasets. We utilize three public 3D clothed human datasets: Thuman2.0 [42], 2K2K [13], and CLOTH4D [49], for training and evaluating our method. We randomly select 400 scans from Thuman2.0 and 1000 scans from 2K2K for our training dataset. The test set comprises 100 scans from

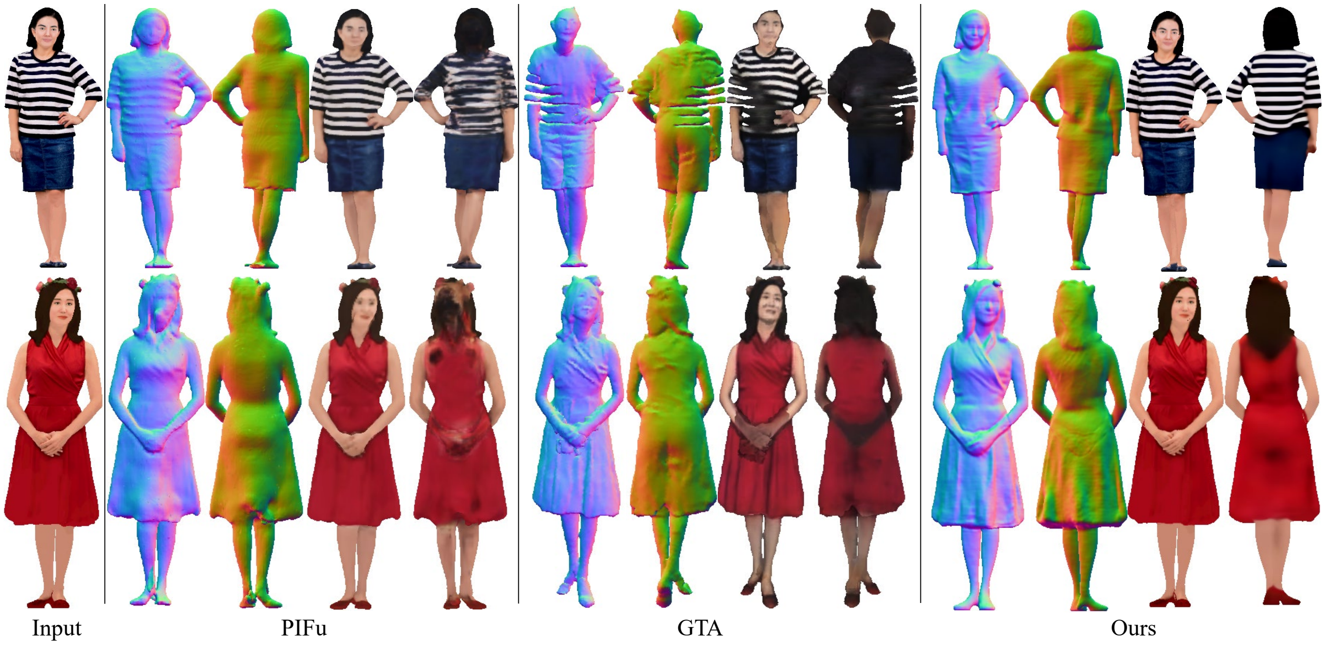


Figure 7. Qualitative comparison of geometry and texture with methods PIFu [27] and GTA [44].

Table 2. Quantitative evaluation of geometry. The best and second-best results are highlighted in **bold** and underlined on each dataset.

	THuman2.0			2K2K			CLOTH4D		
Methods	Chamfer↓	P2S↓	Normal↓	Chamfer↓	P2S↓	Normal↓	Chamfer↓	P2S↓	Normal↓
PIFu [27]	3.224	2.802	3.476	2.232	2.210	3.212	2.124	2.132	3.124
PIFuHD [28]	2.952	2.240	2.698	1.820	1.745	2.362	1.726	1.732	2.235
ICON [39]	2.378	1.820	2.792	1.142	1.012	2.242	1.114	1.198	1.854
FOF [9]	2.474	1.817	2.532	1.132	1.093	2.101	1.218	1.195	1.723
D-IF [41]	2.368	1.820	2.724	1.134	0.912	2.125	1.124	1.012	1.863
ECON [38]	2.312	1.878	2.584	1.174	1.192	1.885	1.132	1.098	1.684
2K2K [13]	2.523	1.914	2.532	1.129	1.107	1.754	1.146	1.132	1.647
GTA [44]	2.298	1.793	2.475	1.130	1.103	1.791	1.102	0.963	1.594
Ours	**2.131**	**1.635**	**2.031**	**0.934**	**0.872**	**1.502**	**0.906**	**0.927**	**1.517**

Thuman2.0, 300 from 2K2K, and 300 from CLOTH4D. For each model, we render the FOF [9], front image, and back-view image as a training pair at each viewpoint of 300 different viewpoints around it utilizing the weak perspective camera model. Each render image resolution is 512 × 512.

Metrics. We evaluate texture performance using three metrics: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) and learned perceptual image patch similarity (LPIPS). We use Chamfer distance and P2S distance [9] to evaluate the geometry quality, comparing reconstructed meshes with ground truth. We also measure L1 normal error between normal images from both meshes by rotating the camera at four fixed angles $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$ relative to the input view.

4.2. Comparison with the State-of-the-art Methods

Texture evaluation. We compare our method with methods: SiCloPe [24], PIFu [27], and GTA [44]. SiCloPe estimates the back-view image; thus, we compare the predicted back-view image. PIFu and GTA use the implicit function to estimate the color of 3D points. To evaluate our results, we render textured meshes by rotating the camera at four fixed angles (0°, 90°, 180°, 270°) relative to the input view and compare the rendered images.

Table 1 presents the quantitative comparison results. When evaluating the estimated back-view image, our result surpasses SiCloPe. Our method has also achieved the best results when evaluating rendered images. Figure 6 presents visual comparison results with method SiCloPe, where our results exhibit greater clarity. Figure 7 presents visual comparison results with methods PIFu and GTA. Our texture and geometry present greater clarity and realism in invisible areas. Quantitative and qualitative experiments verify the effectiveness of our texture estimation method.

Geometry evaluation. We compare our method with methods: PIFu [27], PIFuHD [28], ICON [39], D-IF [41],

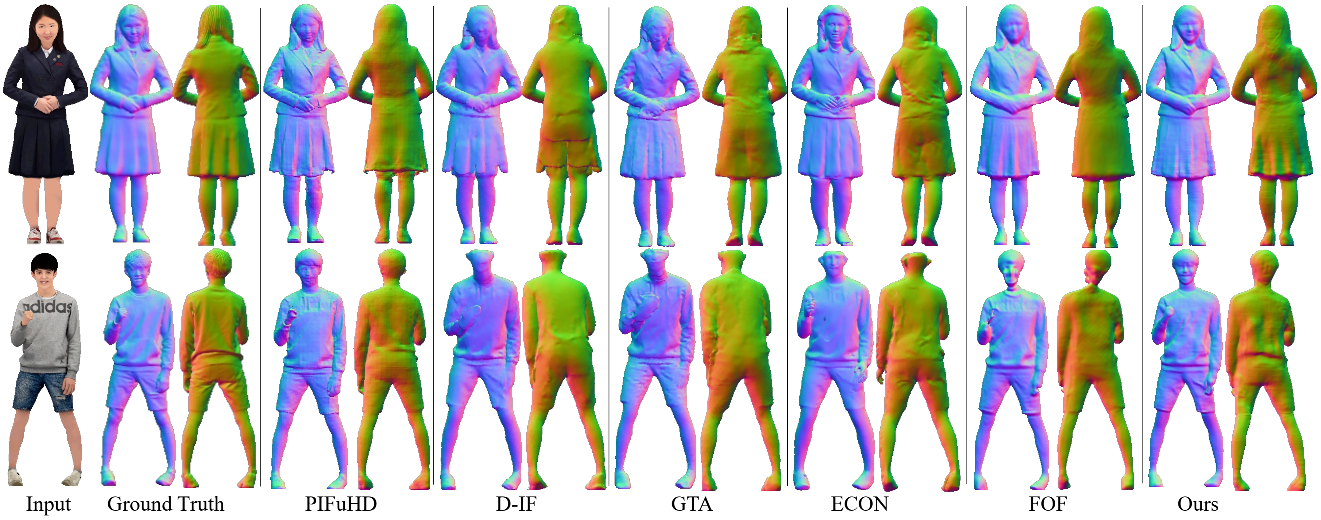


Figure 8. Qualitative comparison of geometry on the 2K2K dataset with state-of-the-art single-view human reconstruction methods: PIFuHD [28], D-IF [41], GTA [44], ECON [38], and FOF [9].

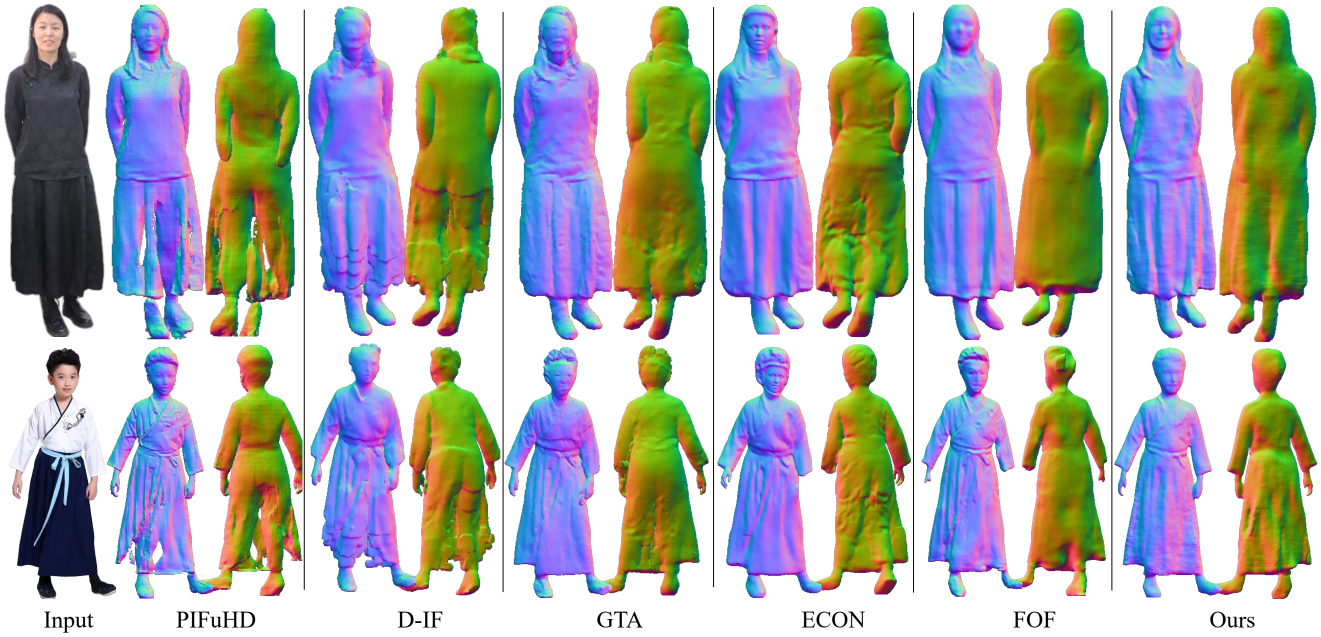


Figure 9. Qualitative comparison of geometry in-the-wild of images with state-of-the-art single-view human reconstruction methods: PIFuHD [28], D-IF [41], GTA [44], ECON [38], and FOF [9].

FOF [9], ECON [38], 2K2K [13], and GTA [44]. Table 2 reports the quantitative comparison results. We can see that our method excels over the compared methods on the three quantitative metrics.

Figure 8 and Figure 9 show qualitative comparison results. In Figure 8, the two input images are from the 2K2K dataset. Our method can accurately reconstruct the geometry and capture more realistic wrinkle details. Figure 9 shows two examples of in-the-wild images. The first image is captured using a camera under natural lighting conditions; the second is from the Internet. Our method achieves outstanding results in reconstructing loose-fitting clothing and recovering more reasonable details in invisible areas. Figure 9 validates our method has robustness and generalization. From both quantitative and qualitative results, our geometry reconstruction method is effective.

4.3. Ablation Study

Ablation study on back-view image estimation. To validate the effectiveness of the proposed style loss $\mathcal{L}_s$ and Siamese network training strategy, we design the following three variants: (1) loss $\mathcal{L}_1$: L1 loss between the predicted back-view image and the ground truth back-view image. (2) loss ($\mathcal{L}_1$+$\mathcal{L}_s$): loss $\mathcal{L}_1$ combines with the style consistency loss $\mathcal{L}_s$ between the predicted back-view image and the input image. (3) loss ($\mathcal{L}_1$+$\mathcal{L}_s$) with Siamese: loss ($\mathcal{L}_1$+$\mathcal{L}_s$) combines with the Siamese network training strategy.

We train the three variants and evaluate the results on the 2K2K dataset. The comparison results of these three variants are summarized in Table 3. The results show that the style loss $\mathcal{L}_s$ and Siamese network training strategy are two important factors for our back-view image prediction. We also present two qualitative results in Figure 10. Compared to only using the $\mathcal{L}_1$ loss, incorporating the style loss $\mathcal{L}_s$ yields more reasonable results.

Table 3. Ablation study on back-view image estimation.

Methods	PSNR↑	SSIM↑	LPIPS↓
loss ($\mathcal{L}_1$)	18.427	0.886	0.247
loss ($\mathcal{L}_1$+$\mathcal{L}_s$)	24.476	0.920	0.174
loss ($\mathcal{L}_1$+$\mathcal{L}_s$) with Siamese	**26.362**	**0.929**	**0.142**

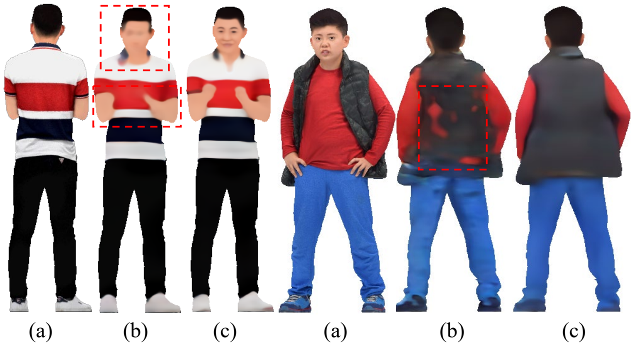


Figure 10. Ablation study for style loss $\mathcal{L}_s$ in the back-view image prediction. (a) Input image, (b) w/o style loss. (c) w style loss.

Ablation study on geometry reconstruction. To validate the effectiveness of the predicted back-view image I'_b, diffusion model, and wavelet transform, we design the following eight variants: (1) FOF-M1: baseline method with the reference image I_a as input. (2) FOF-M2: baseline method with I_a and I'_b as input. (3) FOF-M3: FOF-M2 combines with refinement module. (4) Ours-W1: training on the diffusion model with I_a as condition. (5) Ours-W2: training on the diffusion model with I_a and I'_b as condition. (6) Ours-M1: training on the wavelet diffusion model with I_a as condition. (7) Ours-M2: training on the wavelet diffusion model with I_a and I'_b as condition. (8) Ours-M3: Ours-M2 combines with refinement module.

We train the eight variants and evaluate the results. The comparison results are summarized in Table 4. The results confirm that the predicted back-view image I'_b, the diffusion model, and the wavelet transform are three important factors for our geometry prediction. We also present qualitative results in Figure 11. The geometry quality of FOF-M2 is higher than that of FOF-M1, indicating that the back-view image contributes beneficially to geometry reconstruction. Compared to the baseline method, our approach achieves higher geometry accuracy in reconstruction, particularly in capturing details such as the scarves and facial features.

Table 4. Ablation study on 3D reconstruction.

	THuman2.0			2K2K		
Methods	Chamfer↓	P2S↓	Normal↓	Chamfer↓	P2S↓	Normal↓
FOF-M1	2.474	1.817	2.532	1.132	1.093	2.101
FOF-M2	2.467	1.808	2.528	1.127	1.085	1.984
FOF-M3	2.462	1.803	2.522	1.122	1.078	1.971
Ours-W1	2.245	1.700	2.224	0.981	0.916	1.705
Ours-W2	2.243	1.695	2.221	0.978	0.912	1.700
Ours-M1	2.241	1.690	2.219	0.972	0.907	1.694
Ours-M2	2.138	1.642	2.114	0.942	0.887	1.584
Ours-M3	**2.131**	**1.635**	**2.031**	**0.934**	**0.872**	**1.502**

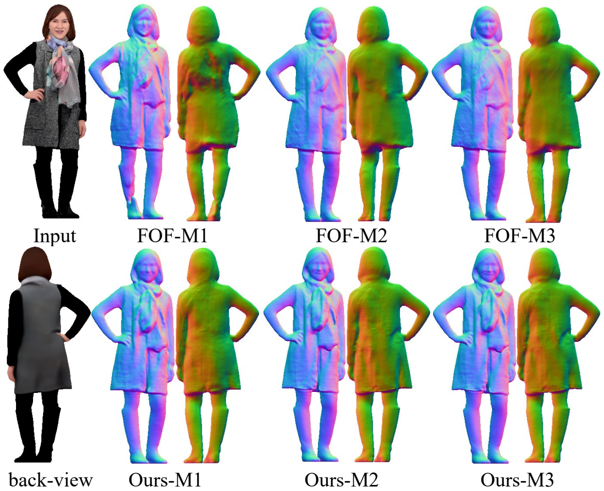


Figure 11. Qualitative comparison with FOF [9].

Limitation. Our method can reconstruct the 3D clothed human model from a single image. However, the reconstructed geometry accuracy may be lower when the human self-occlusion is severe in the image.

5. Conclusion

We have proposed a diffusion model to reconstruct a 3D human model from a single image. First, we proposed a style consistency constraint between the back-view and reference images to effectively predict the back-view image. We then proposed a wavelet-based diffusion model in the geometry prediction to generate the FOF conditional on the two images. The two images are mapped onto the human model, creating a textured clothed human model. Experimental results indicated that our texture estimation and geometry reconstruction methods are effective.

Acknowledgments This work is partially supported by the National Natural Science Foundation of China (No. 61972298 and No. 62372336) and Wuhan University-Huawei GeoInformatices Innovation Lab.

References

[1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *CVPR*, pages 1175–1186, 2019. 1, 2

[2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *ICCV*, 2019. 1, 2

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. *ACM Transactions on Graphics*, 24(3):408–416, 2005. 2

[4] Zhongyun Bao, Chengjiang Long, Gang Fu, Daquan Liu, Yuanzhen Li, Jiaming Wu, and Chunxia Xiao. Deep image-based illumination harmonization. In *CVPR*, pages 18542–18551, 2022. 1

[5] Tuo Cao, Wenxiao Zhang, Yanping Fu, Shengjie Zheng, Fei Luo, and Chunxia Xiao. Dgecn++: A depth-guided edge convolutional network for end-to-end 6d pose estimation via attention mechanism. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1

[6] Gene Chou, Yuval Bahat, and Felix Heide. Diffusion-sdf: Conditional generative modeling of signed distance functions. In *ICCV*, 2023. 2

[7] Ruihang Chu, Enze Xie, Shentong Mo, Zhenguo Li, Matthias Nießner, Chi-Wing Fu, and Jiaya Jia. Diffcomplete: Diffusion-based generative 3d shape completion. *arXiv preprint arXiv:2306.16329*, 2023. 2

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 1, 5

[9] Qiaojun Feng, Yebin Liu, Yu-Kun Lai, Jingyu Yang, and Kun Li. Fof: Learning fourier occupancy field for monocular real-time human reconstruction. In *NeurIPS*, 2022. 1, 2, 3, 6, 7, 8

[10] Yanping Fu, Qingan Yan, Jie Liao, and Chunxia Xiao. Joint texture and geometry optimization for rgb-d reconstruction. In *CVPR*, pages 5950–5959, 2020. 1

[11] 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*, pages 2232–2241, 2019. 1, 2

[12] Alfred Haar. *Zur theorie der orthogonalen funktionensysteme*. Georg-August-Universitat, Gottingen, 1909. 3

[13] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *CVPR*, 2023. 2, 5, 6, 7

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851. 2020. 2, 3, 4

[15] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction. *arXiv preprint arXiv:2303.17559*, 2023. 2

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016. 4

[17] 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. 2

[18] 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. 2

[19] Yuanzhen Li, Fei Luo, and Chunxia Xiao. Monocular human depth estimation with 3d motion flow and surface normals. *The Visual Computer*, 39(8):3701–3713, 2023. 1

[20] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *CVPR*, pages 8139–8148, 2020. 1

[21] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6), 2015. 1, 2

[22] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, pages 163–169, 1987. 3

[23] Fei Luo, Yongqiong Zhu, Yanping Fu, Huajian Zhou, Zezheng Chen, and Chunxia Xiao. Sparse rgb-d images create a real thing: a flexible voxel based 3d reconstruction pipeline for single object. *Visual Informatics*, 7(1):66–76, 2023. 1

[24] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *CVPR*, pages 4475–4485, 2019. 1, 2, 5, 6

[25] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH Conference Proceedings*, pages 1–10, 2022. 2

[26] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 2

[27] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, pages 2304–2314, 2019. 1, 2, 5, 6

[28] 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 84–93, 2020. 1, 2, 6, 7

[29] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023. 2

[30] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *ECCV*, 2022. 2

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[33] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 2

[34] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[35] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *ECCV*, pages 20–36, 2018. 1, 2

[36] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2021. 4

[37] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 5

[38] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *CVPR*, 2023. 6, 7

[39] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: implicit clothed humans obtained from normals. In *CVPR*, pages 13286–13296, 2022. 2, 6

[40] Long Yang, Qingan Yan, Yanping Fu, and Chunxia Xiao. Surface reconstruction via fusing sparse-sequence of depth images. *IEEE Transactions on Visualization and Computer Graphics*, 24(2):1190–1203, 2017. 1

[41] Xueting Yang, Yihao Luo, Yuliang Xiu, Wei Wang, Hao Xu, and Zhaoxin Fan. D-IF: Uncertainty-aware Human Digitization via Implicit Distribution Field. In *ICCV*, 2023. 2, 6, 7

[42] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, 2021. 5

[43] Zongsheng Yue and Chen Change Loy. Difface: Blind face restoration with diffused error contraction. *arXiv preprint arXiv:2212.06512*, 2022. 2

[44] Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. In *NeurIPS*, 2023. 5, 6, 7

[45] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3170–3184, 2022. 2

[46] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *ICCV*, pages 7739–7749, 2019. 1

[47] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *ICCV*, pages 5826–5835, 2021. 2

[48] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *ICCV*, pages 5826–5835, 2021. 2

[49] Xingxing Zou, Xintong Han, and Waikeung Wong. Cloth4d: A dataset for clothed human reconstruction. In *CVPR*, pages 12847–12857, 2023. 5