

MultiGO: Towards Multi-level Geometry Learning for Monocular 3D Textured Human Reconstruction

Gangjian Zhang^{1*} Nanjie Yao^{1*} Shunsi Zhang² Hanfeng Zhao²
 Guoliang Pang² Jian Shu¹ Hao Wang^{1†}

¹HKUST(GZ) ²Guangzhou Quwan Network Technology

Email: gzhang292@connect.hkust-gz.edu.cn, nanjiey@uci.edu, haowang@hkust-gz.edu.cn



Figure 1. Comparisons with the SOTA Methods in Monocular 3D Textured Human Reconstruction. Existing SOTA methods struggle with recovering correct human poses and intricate geometry details. SiFU [48] is unable to reconstruct correct human postures, such as incorrect left-hand positions. VS [15] performs poorly in fine-grained areas such as unclear finger movement and cloth wrinkles. SiTH [5] produces geometry and texture errors that occur from the generative model, such as the third arm on the back.

Abstract

This paper investigates the research task of reconstructing the 3D clothed human body from a monocular image. Due to the inherent ambiguity of single-view input, existing approaches leverage pre-trained SMPL(-X) estimation models or generative models to provide auxiliary information for human reconstruction. However, these methods capture only the general human body geometry and overlook specific geometric details, leading to inaccurate skeleton reconstruction, incorrect joint positions, and unclear cloth wrinkles. In response to these issues, we propose a multi-level geometry learning framework. Technically, we design three key components: skeleton-level enhancement, joint-level augmentation, and wrinkle-level refinement modules. Specifically, we effectively integrate the projected 3D Fourier features into a Gaussian reconstruction model, introduce perturbations to improve joint depth estimation during training, and refine the human coarse wrinkles by resembling the de-noising process of the diffusion model. Extensive quantitative and qualitative experiments on two test

sets show the superior performance of our approach compared to state-of-the-art (SOTA) methods.

1. Introduction

With the increasing popularity of virtual worlds, there is a growing demand for the realistic digital human creation. To achieve it efficiently, monocular 3D human reconstruction has been an important task. However, since single-view images cannot provide sufficient information for reconstruction, there exists substantial ambiguity in simulating the geometry and texture of the occluded human body parts.

To address this issue, existing methods have explored the introduction of pre-conditioned SMPL-X techniques to provide a 3D body mesh as geometric prior information. These methods employ various options for human body geometry priors, such as SMPL-X normal maps [15, 37, 38, 48], low-frequency and high-frequency signals of the SMPL-X body [42], triplane representation [47], and the outputs of generative models [5, 26], in an attempt to achieve more accurate human body reconstruction.

However, these traditional approaches often focus on modeling the general geometry of the human body, overlooking the multi-level structures that include the skeleton, joints, and finer details such as wrinkles around the fingers

†: Corresponding author, *: Equal contribution.
 Project page: <https://multigohuman.github.io/>.

and face. This oversimplified modeling approach results in inaccurate skeletal reconstructions, incorrect joint positions, and unclear representation of clothing wrinkles.

Moreover, these approaches often utilize 3D representation methods like occupancy grids [26], SDF [21], and NeRF [18]. While these representations are capable of accurately modeling human body geometry, they are often burdened by high computational costs and low efficiency. Recently, Szymanowicz et al. [29] perform monocular reconstruction by mapping a single object image into 3D Gaussian points using a neural network in a feed-forward process, which enables effective and efficient 3D object reconstruction. This work motivates us to explore Gaussian models for the monocular human reconstruction task.

To this end, this paper proposes a novel multi-level geometry learning framework, MultiGO, based on an existing object Gaussian reconstruction model [30]. We aim to enhance the geometry of human reconstructions across various levels of granularity, including skeletons, joints, and wrinkles, thereby largely improving 3D human reconstruction quality. The proposed MultiGO contains three key components, i.e., the Skeleton-Level Enhancement (SLE) module, Joint-Level Augmentation (JLA) strategy, and Wrinkle-Level Refinement (WLR) module, which deal with the specificity of three different levels of the human geometry respectively.

Technically, (1) the SLE module is designed to enhance the accuracy of capturing human overall posture by effectively bridging the 3D SMPL-X prior with 2D human imagery. By projecting 3D Fourier features into the same 2D space as the input image, the SLE module allows the model to fully utilize established geometric priors related to human shapes and configurations. (2) The JLA strategy tackles the significant challenge in estimating the depth of human joints in 3D space. It recognizes that, during inference, inaccuracies in the estimated SMPL-X depth prior can lead to substantial errors. To address this, the JLA strategy introduces controlled perturbations to the ground truth joint positions during training. (3) The WLR module refines geometric details, such as wrinkles on the human body, by using high-quality textures generated from Gaussian representations as a conditioning. It effectively resembles this refinement process to the final stage of diffusion de-noising.

Experiments on two test sets, CustomHuman and THuman3.0 validate that our proposed multi-level geometry learning framework achieves SOTA performance. Our contributions are as follows:

- A Skeleton-Level Enhancement module, which captures human overall posture by integrating the projected 3D Fourier features with 2D images.
- A Joint-Level Augmentation strategy, which applies perturbations to joint positions during training to improve the model’s resilience to depth prior inaccuracy in inference.

- A Wrinkle-Level Refinement module, which refines the coarse geometric wrinkles using reconstructed Gaussian texture in a de-noising manner of the diffusion model.

2. Related Work

2.1. Monocular Human Reconstruction

Monocular human reconstruction has become a popular research focus for human digitalization. The pioneering method, PIFu [26], introduced a pixel-aligned implicit function for shape and texture generation. ICON [37] improved on this by using skinned body models [17] as human body priors, while ECON [38] combined implicit representations with explicit body regularization. GTA [47] used a 3D-decoupling transformer and hybrid prior fusion for comprehensive 3D geometry and texture reconstruction. VS [15] introduced a “stretch-refine” strategy for managing large deformations in loose clothing. HiLo [42] enhanced geometry details by extracting high and low-frequency information from a parametric model, improving noise robustness as a result. SiTH [5] tackled occlusion by leveraging a 2D prior from the SD model. Lastly, SiFU [48] utilized a cross-attention mechanism in transformers to optimize geometry, employing 2D diffusion models for texture prediction.

2.2. Human Gaussian Model

Recent advancements in 3D Gaussian splatting [10] have promoted 3D human creation. Traditional representations like SDF and NeRF often struggle with balancing efficiency and rendering quality. Techniques such as HuGS [12], D3GA [50], and 3DGS-Avatar [22] utilize rich spatial and temporal data for modeling in multi-view videos, monocular video sequences, and sparse-view inputs. Animatable Gaussians enhance garment dynamics through pose projection mechanisms and 2D CNNs. Notably, Gauhuamn [8] and HUGS [12] optimize human Gaussians from monocular videos, while HiFi4G [9] employs a dual-graph mechanism for spatial-temporal consistency, and ASH [20] uses mesh UV parameterization for real-time rendering. GPS-Gaussian [49] proposes a generalizable multi-view human Gaussian model with high-quality rendering.

2.3. Large 3D Object Reconstruction Model

Recent advancements in 3D object reconstruction leverage large models for efficient 2D-to-3D conversions. LRM [7] has notably enhanced model capacity and data volume, enabling direct 3D reconstruction from a single image. Further improvements, such as instant3d and others [13, 33, 40], utilize multi-view diffusion models for better results. Key datasets like Objaverse [3] support training these models, with significant focus on 3D Gaussian representation [30, 39, 41, 44] and 3D triplane representation [35]. These methods often require extensive 3D data for pre-

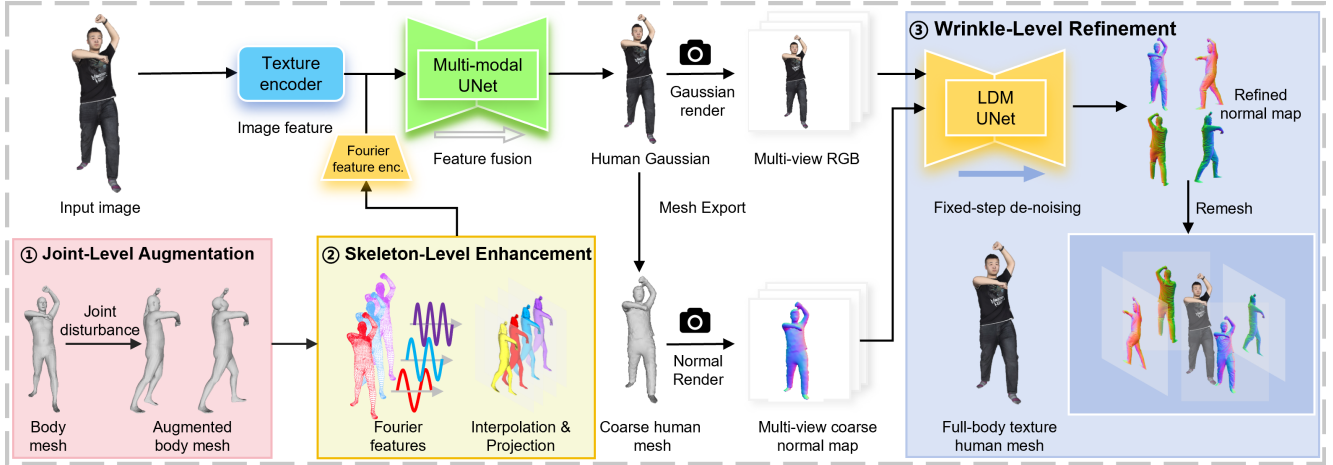


Figure 2. **Method Overview.** Our method, MultiGO, addresses monocular textured 3D human reconstruction by introducing a multi-level geometry learning framework that significantly enhances reconstruction quality. To accurately capture the human body’s posture, we propose the Skeleton-Level Enhancement (SLE) module, which projects 3D Fourier features into the 2D space of the input image, allowing the Gaussian reconstruction model to fully utilize prior human shape knowledge. For improved depth estimation of human joints, the Joint-Level Augmentation (JLA) strategy applies controlled perturbations during training, increasing the model’s robustness to depth inaccuracies during inference. To refine geometric details like body wrinkles, the (Wrinkle-Level Refinement) WLR module resembles the final de-noising steps in diffusion theory, treating coarse meshes as Gaussian noise and using the high-quality texture of reconstructed Gaussian as conditions to refine wrinkles.

training diffusion models [16, 27, 32], which are crucial for generating novel multi-view inputs.

3. Methodology

3.1. Preliminaries

Gaussian Splatting. Gaussian splatting [10], emerging as a popular 3D representation, utilize a collection of 3D Gaussians, denoted by Θ , to model 3D data. Each 3D Gaussian is characterized by a parameter set: $\theta_i = \{\mathbf{x}_i, \mathbf{s}_i, \mathbf{q}_i, \alpha_i, \mathbf{c}_i\} \in \mathbb{R}^{14}$. Here, $\mathbf{x} \in \mathbb{R}^3$ represents the geometry center, $\mathbf{s} \in \mathbb{R}^3$ the scaling factor, $\mathbf{q} \in \mathbb{R}^4$ the rotation quaternion, $\alpha \in \mathbb{R}$ the opacity value, and $\mathbf{c} \in \mathbb{R}^3$ the color feature. In our work, we set the dimension of \mathbf{c} to 3 for rendering, allowing spherical harmonics to model view-dependent effects.

SMPL Series Model. The Skinned Multi-Person Linear (SMPL) model [17] is a parametric model to represent human body. It utilizes a set of body parameters $\mathcal{B} \in \mathbb{R}^d$ to define a given human body mesh \mathcal{M} :

$$\mathcal{M}(\mathcal{B}) : \mathcal{B} \Rightarrow \mathbb{R}^{3 \times 6890}. \quad (1)$$

Each parameter in \mathcal{B} controls the position or orientation of body parts, etc. Many extensions such as SMPL-X [17], and SMPL-H [25] add more parameters for facial expressions, finger movement, and other fine-grained poses. By default, we use SMPL-X as our human body model in this paper.

Latent Diffusion Model. Diffusion model [6] proposes to generate images through a degradation process. The latent

diffusion model (LDM) [24] incorporates a pre-trained variational auto-encoder (VAE) [11], including the encoder and decoder. Starting with a sample z_0 from the VAE latent distribution z , the forward process produces a sequence of noised data $\{z_t \mid t \in (0, T)\}$, where $z_t = z_0 + \epsilon$. Here ϵ is randomly sampled noise from a Gaussian noise $\mathcal{N}(0, 1)$. Conversely, the reverse process uses an iterative de-noising way to recover z_{t-1} from z_t by predicting the noise ϵ .

3.2. The Proposed Method

3.2.1 Overview

Our method, illustrated in Figure 2, processes the human front-view RGB image and a 3D body mesh generated by the SMPL-X model. A texture encoder extracts 2D image features from the RGB input, while the 3D mesh is transformed into geometric features via our SLE module. These features are then integrated using a multimodal UNet, resulting in 3D human Gaussian points. The SLE module and fusion process are detailed in Section 3.2.2. To enhance depth estimation accuracy during inference, we introduce a JLA strategy, explained in Section 3.2.3. Additionally, to improve mesh detail post-Gaussian export, we propose the WLR module, described in Section 3.2.4.

3.2.2 Skeleton-Level Enhancement Module

In the previous analysis, the monocular setting of this task means a frontal human RGB image alone cannot provide sufficient geometric information. To overcome this, we pro-

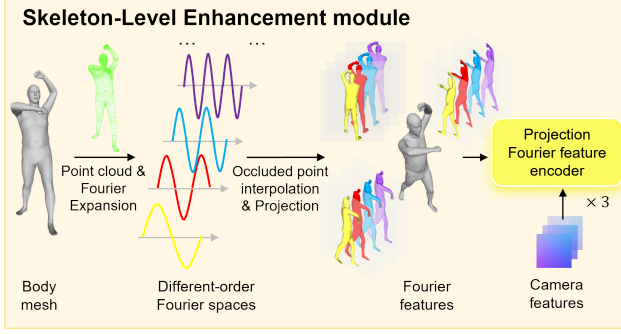


Figure 3. **Skeleton-Level Enhancement Module.** To enhance the human geometry at the skeleton level, we achieve better fusion of the heterogeneous modalities of the 3D SMPL-X body prior and 2D images. We propose interpolating the Fourier features of 3D occluded points and mapping them from three different angles into the same 2D space as the image features.

pose a novel Skeleton-Level Enhancement (SLE) module that effectively incorporates the 3D geometric prior knowledge of the SMPL-X body mesh. Recognizing that image features and 3D SMPL-X features originate from two different modalities with a huge semantic gap, our approach avoids the rigid fusion of these features. Instead, the SLE module projects 3D Fourier features into the same 2D space as the input image, facilitating better interaction and fusion between the heterogeneous features. This module enables the model to effectively learn the geometric skeleton of the human, even with a limited amount of human data.

Concretely, as shown in Figure 3, inspired by some works [14, 43] the proposed SLE module first considers all vertices of the input body mesh as points of the point cloud. The body point cloud can be represented as $\mathcal{P}_0 \in \mathbb{R}^{3 \times 10475}$. Then, the 3D Fourier expansion operation is used to enhance the expression of these points. Specifically, we extract q -order Fourier series for each point p in \mathcal{P}_0 as follows:

$$\mathcal{F}(p) = \{p\} \cup \{\cos(2^n p), \sin(2^n p) | n \in \{1, \dots, q\}\}. \quad (2)$$

Through the above operation, we have expanded the 3D space where the original SMPL-X points are located into the different Fourier spaces $\{\mathcal{S}_n | n \in \{0, \dots, 2q\}\}$. The point clouds in these spaces are denoted as $\{\tilde{\mathcal{P}}_n | n \in \{0, \dots, 2q\}\}$. Meanwhile, to make the points in these spaces denser, we have interpolated and expanded them. Specifically, we select positions on the surface of a triangular surface and average the weights of three vertices belonging to the same triangular surface. After this, denser point clouds $\tilde{\mathcal{P}}_n \in \mathbb{R}^{3 \times m}$ with different-order Fourier are obtained, where m is the point number.

To facilitate the fusion of geometric and texture features, we perform 2D projection on the occluded points in different Fourier spaces from three camera angles (orthogonally back, left, and right). By doing so, we can obtain a stack of

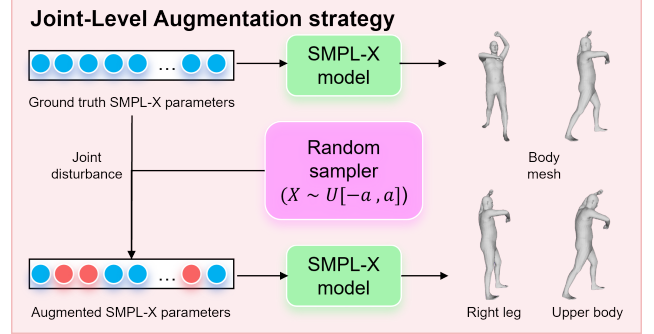


Figure 4. **Joint-Level Augmentation Strategy.** To enhance human geometry at the joint level, we augment the samples of input human SMPL-X body mesh during training. We propose to randomly perturb ground truth SMPL-X parameters associated with specific joints to increase the model robustness in inference.

Fourier features from different spaces as shown in Figure 3, which can be concatenated into $\tilde{\mathcal{F}}_1 \in \mathbb{R}^{3(2q+1) \times H \times W}$, where H and W are the projection resolution. Similarly, from the perspectives of the other two cameras, we can obtain $\tilde{\mathcal{F}}_2$ and $\tilde{\mathcal{F}}_3$. Subsequently, all of them, along with their camera feature, are fed into a Fourier feature encoder to obtain geometric features, $\mathcal{F}'_1, \mathcal{F}'_2, \mathcal{F}'_3 \in \mathbb{R}^{o \times h \times w}$. We design a one-layer convolution (Kernel size is 3, stride/padding is 1) to form the Fourier feature encoder, making it output the same dimension as the image latent $\mathcal{I}'_0 \in \mathbb{R}^{o \times h \times w}$ obtained after texture encoding the RGB image $\mathcal{I}_0 \in \mathbb{R}^{3 \times H \times W}$.

In our approach, features from two modalities and four perspectives are combined using a multimodal fusion UNet [30], enhanced with residual and self-attention layers. The self-attention layers serve to facilitate interaction between different perspectives and integrate features across modalities. These fused features are then decoded to predict the final textured Gaussian parameters, Θ , forming the ultimate feature map. Each pixel is modeled as a 3D Gaussian $\theta_i \in \mathbb{R}^{14}$. We utilize the differentiable renderer [41] to render these Gaussians. RGB and alpha images from eight views, including one input and seven novel views, are rendered. The loss is computed by comparing these rendered images to GT scans, using MSE and VGG-based LPIPS loss [46] for RGB images, and MSE loss for alpha images.

3.2.3 Joint-Level Augmentation Strategy

The SLE module aids in fitting the model to overall human pose inputs effectively. However, unlike the precise SMPL-X body parameters from human scans used during training, those used during inference are estimated from a single view, leading to potential inaccuracies in reflecting true human geometry. This results in depth discrepancies in the fitted body mesh, for example, the difference in the front and back position of the feet is obvious when viewed

from the side, but not obvious when viewed from the front. To mitigate these inaccuracies in SMPL-X depth estimation during inference, we propose a Joint-Level Augmentation (JLA) Strategy, which involves altering some body parameters in the training data, as illustrated in Figure 4.

To simplify, we represent the scan-fitted GT human body parameters as a vector, $\mathcal{B} = (\beta^0, \beta^1, \dots, \beta^d) \in \mathbb{R}^d$, where d is the total number of body parameters. Each element in this vector influences posture changes, for example, β^0 controls the forward and backward rotation of the left leg, and β^{35} manages the left and right rotation of the head. As illustrated in Figure 4, these parameters are input into a pre-trained SMPL-X model to generate a standard body mesh.

During training, instead of using these standard body meshes directly, we propose creating augmented samples to simulate inference scenarios. Specifically, we focus on parameters affecting depth information, like β^0 , which are more likely to produce errors in single-view SMPL-X estimation during inference. We construct a body depth-related mask vector $M = (\mu^1, \mu^2, \dots, \mu^d) \in \mathbb{R}^d$, where each element $\mu^j \in 0, 1$ for $j \in 1, 2, \dots, d$. For elements where $\mu^j = 1$, we generate a sample from a uniform distribution $X \sim U[-\alpha, \alpha]$ to produce an offset x^j , with α as a hyperparameter controlling the offset degree. By adding these offsets to the GT SMPL-X parameters, we obtain the augmented SMPL-X parameters $\tilde{\beta} = (\tilde{\beta}^0, \tilde{\beta}^1, \dots, \tilde{\beta}^d) \in \mathbb{R}^d$. Each $\tilde{\beta}^j$ is calculated as follows:

$$\tilde{\beta}^j = \beta^j + \mu^j * x^j \quad (3)$$

The augmented parameter vector is input into the pre-trained SMPL-X model to produce a body mesh with minor geometric joint changes from the standard mesh. Moreover, the larger the α , the less similar the augmented body mesh is to the original one. This helps the model develop correction abilities, ensuring accurate predictions even when discrepancies exist between the input SMPL-X mesh and the target human geometry. This correction ability is derived from extracting and understanding image modality.

3.2.4 Wrinkle-Level Refinement Module

While the SLE module and JLA strategy allow for the creation of high-quality 3D Gaussian point clouds of the human body, converting these into meshes using the Gaussian-to-mesh-export technique [30] often fails to capture fine details like facial expressions and clothing. To enhance the geometric details of these meshes, we introduce the Wrinkle-Level Refinement (WLR) module. This module utilizes the detailed texture of the reconstructed Gaussian as the condition to refine mesh wrinkles. Additionally, given our access to a coarse mesh and a near-ground-truth (GT) mesh, our approach innovatively resembles the refinement process with the final de-noising stage in diffusion theory. We

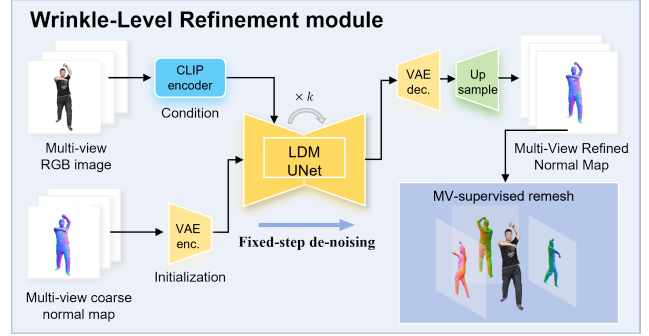


Figure 5. **Wrinkle-Level Refinement Module.** To improve the human geometry at the wrinkle level, we equate the refinement process with the last few steps of the de-noising process in the diffusion model and use a fixed number of “de-noising” steps to achieve the refined mesh predicting from initialized coarse mesh.

treat the coarse mesh as if it has been subjected to k iterations of Gaussian noise, then apply diffusion and de-noising pipeline to achieve enhanced results.

The full WLR is in Figure 5. Starting with a reconstructed human Gaussian \mathcal{G} and a coarse human mesh \mathcal{M}^c , we use differentiable renderers $R_{\mathcal{G}}$ and $R_{\mathcal{M}^c}$ to generate the Gaussian-rendered color image $I_c = \{I_c^1, \dots, I_c^N\}$ and mesh-rendered coarse normal map $I_n = \{I_n^1, \dots, I_n^N\}$ from a common set of camera views \mathbf{k} . Then, the pre-trained CLIP [23] image encoder and VAE encoder \mathcal{E} are employed to encode the color image and normal map, respectively.

Fine-tuning and inference. Training a data-driven diffusion model from scratch is challenging, especially with limited human data. To address this, we fine-tune a pre-trained latent diffusion model UNet [36], denoted as ϵ_{θ} . During fine-tuning, we define the latent of the ground truth (GT) normal map as $z_0 = \mathcal{E}(I_n^1)$ and the latent of the coarse normal map as $z_k = \mathcal{E}(I_n)$, which can be taken as the result of adding noise k times to z_0 from the traditional perspective of diffusion. The model ϵ_{θ} is to predict the distribution discrepancy ϵ between z_0 and z_k . And the time step $t = k$. The objective function for our training is as follows:

$$\min_{\theta} \mathbb{E}_{\epsilon} \|\epsilon - \epsilon_{\theta}(z_k, k, I_c)\|_2^2 \quad (4)$$

Then during inference time, we can obtain the refined normal map \hat{I}_n by:

$$\hat{I}_n = \mathcal{D}(\hat{z}_0) = \mathcal{D}(\mathcal{F}(z_0, I_c)) \quad (5)$$

where \mathcal{D} is the VAE decoder, \hat{z}_0 is the refined latent, and the \mathcal{F} represents the k -step de-noising process of our model. Additionally, we applied a fine-tuned super-resolution model [34] for upsampling, which allows a refined human normal map with intricate geometric details.

Continuous remeshing. Building on inverse rendering research [2, 19], we iteratively refine the human mesh using

Methods	Publication	CustomHuman [4]			THuman3.0 [28]		
		CD: P-to-S / S-to-P (cm) ↓	NC ↑	f-score ↑	CD: P-to-S / S-to-P (cm) ↓	NC ↑	f-score ↑
PIFu [26]	ICCV 2019	2.965/3.108	0.765	25.708	2.176/2.452	0.773	34.194
GTA [47]	NeurIPS 2023	2.404/2.726	0.790	29.907	2.416/2.652	0.768	29.257
ICON [37]	CVPR 2022	2.437/2.811	0.783	29.044	2.471/2.780	0.756	27.438
ECON [38]	CVPR 2023	2.192/2.342	0.806	33.287	2.200/2.269	0.781	33.220
VS [15]	CVPR 2024	2.508/2.986	0.779	26.790	2.523/2.943	0.758	26.340
HiLo [42]	CVPR 2024	2.280/2.739	0.794	30.275	2.385/2.862	0.765	28.119
SiFu [48]	CVPR 2024	2.464/2.782	0.790	28.574	2.480/2.822	0.762	27.929
SiTH [5]	CVPR 2024	1.800/2.188	0.816	36.148	1.763/2.002	0.787	36.230
MultiGO	-	1.620/1.782	0.850	42.425	1.408/1.633	0.834	46.091

Table 1. **Comparison of Human Geometry with SOTA methods.** The best results are highlighted with bold. Arrow ↑/↓ means higher/lower is better. Note that all the experiments are tested with the estimated SMPL/SMPL-X model instead of the GT one.

Methods	CustomHuman			THuman3.0		
	LPIPS: F/B ↓	SSIM: F/B ↑	PSNR: F/B ↑	LPIPS: F/B ↓	SSIM: F/B ↑	PSNR: F/B ↑
PIFu	0.0792/0.0966	0.8965/0.8742	18.141/16.721	0.0706/0.0849	0.9242/0.9007	20.104/17.926
GTA	0.0730/0.0891	0.9003/0.8923	18.790/18.229	0.0633/0.0770	0.9298/0.9275	21.113/20.497
ICON	0.0710/-	0.8976/-	18.613/-	0.0608/-	0.9291/-	21.127/-
ECON	0.0781/-	0.8868/-	18.454/-	0.0658/-	0.9261/-	20.961/-
SiFu	0.0692/0.0879	0.9023/0.8915	18.715/18.111	0.0597/0.0768	0.9302/0.9243	21.101/20.349
SiTH	0.0679/0.0843	0.9007/0.8870	18.417/17.608	0.0612/0.0766	0.9232/0.9107	20.326/19.355
MultiGO	0.0414/0.0643	0.9603/0.9415	22.347/20.849	0.0457/0.0616	0.9623/0.9512	23.794/22.657

Table 2. **Comparison of Human Texture with SOTA methods.** Note that only some methods predict the texture of the human body, so we render the textured 3D human reconstruction results of these methods in front view and back view, represented by “F/B” symbols. The ICON and ECON methods only predict the front view texture.

the improved human normal maps \hat{I}_n^i . In each step, we differentially render the predicted mesh’s normal maps and compare them with \hat{I}_n^i to calculate the loss and gradients. These gradients guide the adjustment of vertices and faces, allowing us to achieve a more detailed human mesh. The loss function is explicitly defined as:

$$\mathcal{L}_{normal} = \sum_{i=0}^N \|\hat{I}_n^i - R(\mathcal{M}', \mathbf{k})\|_2^2 \quad (6)$$

where the \mathcal{L}_{normal} represents the main loss, and the \mathcal{M}' is the iteratively predicted human mesh during the optimization, and $R(\cdot)$ function means the differentiable rendering.

4. Experiments

4.1. Experiment Setup

Dataset. Previous studies often rely on datasets such as RenderPeople [1], which limits reproducibility due to accessibility issues. For fair comparisons, similar to ICON [37], we conduct experiments on the latest publicly available 3D human dataset, THuman2.0. Additionally, we use the CustomHumans [4] and THuman3.0 [28] datasets for evaluation. During evaluation, all method utilizes the estimated SMPL/SMPL-X models as the body prior. Detailed information about datasets is in the supplementary material.

4.2. Quantitative Comparison of Human Geometry

Evaluation Metrics. To evaluate the reconstruction quality, we follow the prior work SiTH [5] to compute 3D metrics

Chamfer distance (CD), Normal Consistency (NC), and **f-Score** [31] on our generated results. Moreover, we select the **LPIPS** [45], **SSIM**, and **PSNR** as the 2D metrics to evaluate the texture quality of the reconstructed mesh.

Table 1 highlights the superior performance of our method in reconstructing human geometry compared to other SOTA techniques. We evaluated all approaches on two datasets, CustomHuman and THuman3.0, ensuring a fair comparison. Particularly, our method outperformed competitors, including the second-ranked method. On CustomHuman, we achieve improvements of 0.180/0.406 on CD, 0.034 on NC, and 6.277 on f-score. For THuman3.0, the gains are 0.355/0.369 on CD, 0.047 on NC, and 9.861 on f-score. We attribute our performance improvement to the introduction of our novel solution. Unlike traditional methods that focus solely on general human body geometry, our approach models human body geometry in a more detailed and multi-level manner. By enhancing human body reconstruction at three distinct levels, rather than treating it as a whole, we achieve superior results.

4.3. Quantitative Comparison of Human Texture

In Table 2, we also present a 2D quantitative comparison of human texture quality, highlighting our method’s superiority over other SOTA approaches. On CustomHuman, our method enhances LPIPS by 0.0265/0.0200 (F/B), SSIM by 0.0580/0.0492 (F/B), and PSNR by 3.557/2.620 (F/B). Similarly, on THuman3.0, it improves LPIPS by 0.0140/0.015 (F/B), SSIM by 0.0321/0.0237 (F/B), and

Methods	Components	CustomHuman			THuman3.0		
		CD: P-to-S / S-to-P (cm) ↓	NC ↑	f-score ↑	CD: P-to-S / S-to-P (cm) ↓	NC ↑	f-score ↑
w/ 3-view Proj.	SLE	1.634/1.790	0.847	42.081	1.442/1.650	0.829	45.525
w/ 2-view Proj.		1.772/2.061	0.836	38.651	1.477/1.831	0.820	43.255
w/ 1-view Proj.		1.842/2.313	0.823	36.835	1.481/1.851	0.819	42.802
w/o Fourier Proj.		2.263/2.617	0.806	31.624	1.966/2.160	0.804	35.086
w/o Shape prior		2.130/2.595	0.808	33.295	1.932/2.095	0.798	37.685
Ours ($\alpha=0.25$)	JLA	1.634/1.790	0.847	42.081	1.442/1.650	0.829	45.525
Ours ($\alpha=0.10$)		1.650/1.809	0.846	41.418	1.545/1.704	0.824	44.008
Ours ($\alpha=0.00$)		1.708/1.848	0.843	40.429	1.611/1.746	0.820	43.166
w/o WLR	WLR	1.634/1.790	0.847	42.081	1.442/1.650	0.829	45.525
Our Full Pipeline		1.620/1.782	0.850	42.425	1.408/1.633	0.834	46.091

Table 3. **Ablation Study on Three Core Components.** We evaluate the effectiveness of the SLE module for reconstructing human geometry by comparing model performance with and without the 2D projection of 3D SMPL-X features. In the “w/o Shape prior” setup, only the RGB image is used for reconstruction, serving as our baseline. In “w/o Fourier Proj.,” we encode 3D SMPL-X Fourier features as a whole using several convolution layers, instead of projecting the 3D Fourier into 2D feature first as proposed. For “1-view Proj.,” “2-view Proj.,” and “3-view Proj.,” we implement the proposed projection operation from one, two, and three camera views, respectively. We assess the JLA module’s impact by comparing performance with and without human joint disturbance during training. The hyperparameter α controls augmentation intensity: “ $\alpha=0.00$ ” indicates no disturbance, while “ $\alpha=0.10$ ” and “ $\alpha=0.25$ ” apply the strategy at varying intensities. Lastly, we evaluate the WLR module’s effectiveness by testing conditions with and without this module.

Methods	CustomHuman		
	LPIPS: F/B ↓	SSIM: F/B ↑	PSNR: F/B ↑
Ours (w/o SLE)	0.0497/0.0750	0.9414/0.9205	21.026/19.661
Ours (w/o JLA)	0.0465/0.0666	0.9496/0.9359	21.535/20.327
Our Full Pipeline	0.0414/0.0643	0.9603/0.9415	22.347/20.849
	THuman3.0		
	LPIPS: F/B ↓	SSIM: F/B ↑	PSNR: F/B ↑
Ours (w/o SLE)	0.0575/0.0676	0.9408/0.9299	22.129/21.048
Ours (w/o JLA)	0.0541/0.0657	0.9449/0.9407	22.262/21.647
Our Full Pipeline	0.0457/0.0616	0.9623/0.9512	23.794/22.657

Table 4. **Ablation Study of SLE and JLA for Texture Quality.** We demonstrate through the ablation of SLE and JLA modules that the optimization of geometry, specifically the skeleton-level and joint-level, also leads to improvements in human texture.

PSNR by 2.667/2.160 (F/B), achieving SOTA performance. We attribute the improvement in texture to our accurate reconstruction of human geometry. Once the geometric representation of the human body is enhanced, the overall quality naturally improves as well. This conclusion is demonstrated by the ablation study in Table 4.

4.4. Quantitative Ablation Study

Effectiveness of SLE. Table 3 highlights the significant impact of correctly integrating 3D features into the model. Unlike our proposed projection operation, directly encoding 3D features using methods like [43] and incorporating them into the reconstruction model yields poor results, even negatively affecting performance. This is likely due to the pre-trained model being optimized with extensive 2D RGB images, creating a modality and semantic gap between 2D and 3D features. The direct fusion of 3D SMPL-X and 2D image features is challenging with limited human data. Our

method, however, enhances model performance by effectively leveraging geometric human priors. Increasing the number of mapping surfaces reveals further performance improvements, underscoring the method’s effectiveness.

Effectiveness of JLA. Table 3 shows that JLA significantly enhances human geometry reconstruction. By fine-tuning disturbance intensity, model performance improves further. This suggests that training with augmented body meshes helps the model develop correction abilities, allowing accurate predictions despite slightly unsatisfactory estimated SMPL-X body mesh. This correction ability is rooted in effective image modality extraction and understanding.

Effectiveness of WLR. Table 3 also highlights WLR’s effectiveness in enhancing model performance across two test sets. Although the quantitative improvement is less pronounced than with SLE and JLA, WLR is crucial for refining details at the wrinkle level, such as face, and clothing, which are also essential for the task.

Effect of SLE and JLA on texture quality. In Table 4, we observe that the proposed SLE and JLA models, aimed at enhancing the geometric quality of the human body at both the skeleton and joint levels, also lead to improvements in body texture quality. This correlation is intuitive: when our reconstructed human body is geometrically aligned with the target, the rendered 2D image aligns more closely as well, resulting in better 2D performance.

4.5. Visualization

Comparison with SOTA methods. In Figures 1 and 6, we intuitively demonstrate the superiority of our algorithm over existing methods. The figure clearly shows that our approach excels in capturing human skeletons, joints, and

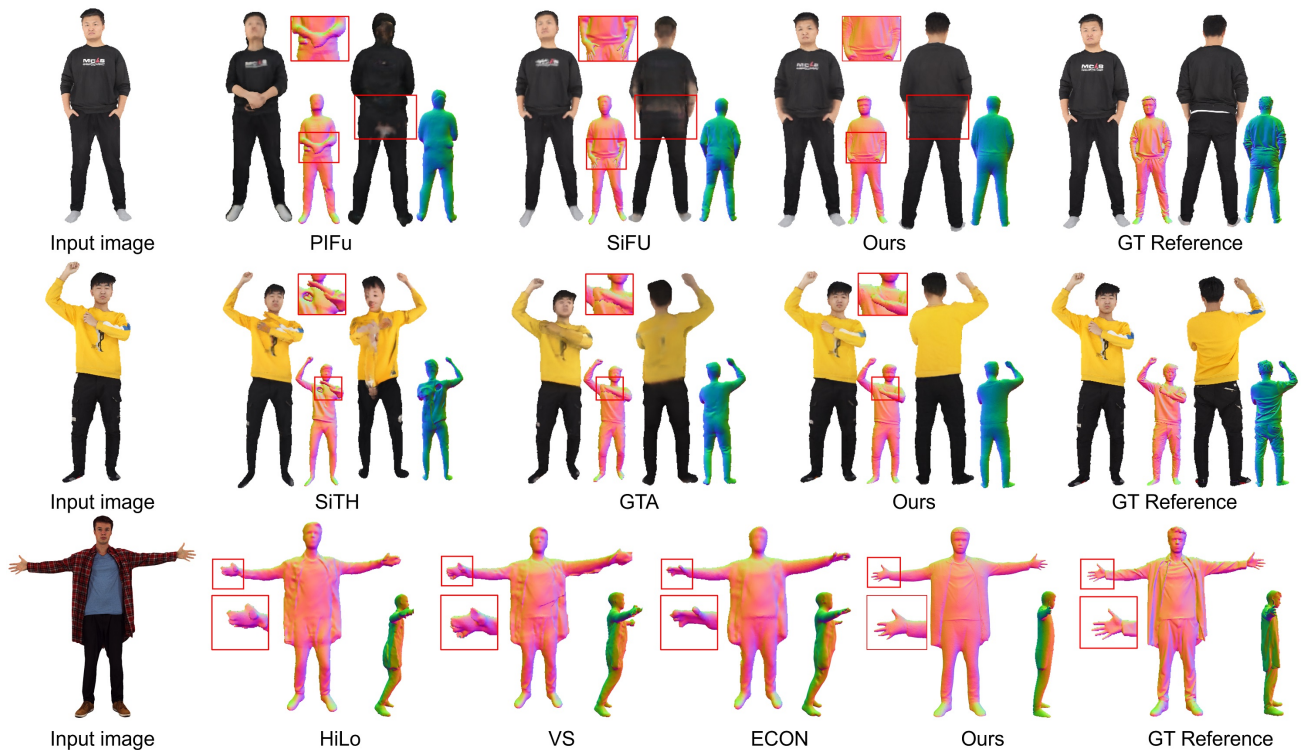


Figure 6. **Qualitative comparison with SOTA methods.** In the first two rows, we present results from methods that generate both mesh and texture. The third row showcases methods that produce mesh only. Our proposed MultiGO significantly outperforms existing state-of-the-art methods in terms of skeleton integrity, posture accuracy, and wrinkling detail. In the first row, both PIFu and SiFU fail to represent human geometry and texture accurately, resulting in incorrect palm positioning and excess color at the back. In the second row, SiTH and GTA struggle with body geometry, particularly with overly slender arms or unnecessary protrusions. In contrast, our method distinctly highlights the contours of the human hand, outperforming other approaches. Please **zoom in** for more details.

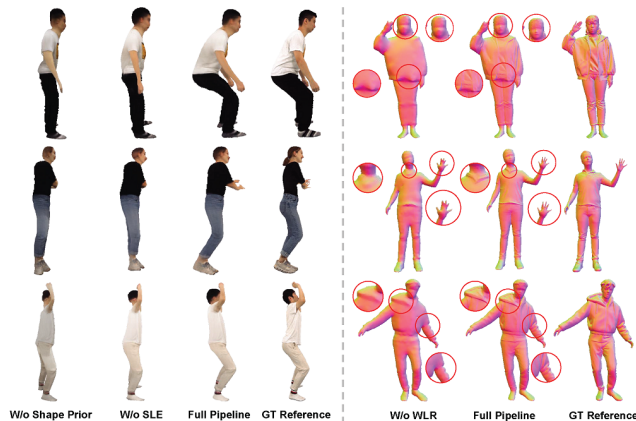


Figure 7. **Visual Ablation.** *Left:* The SLE (“Full pipeline”) makes the overall human skeleton close to the GT. *Right:* The usage of WLR makes the wrinkles clearer. Please **zoom in** for more details.

wrinkles, thanks to our multi-level modeling of human geometry. Additionally, this geometric improvement enables our method to achieve SOTA performance in texture as well.

Effect of SLE and JLA on reconstruction quality. Figure 7 illustrates the effectiveness of our proposed modules.

On the left subfigure, it is found that without any geometric prior knowledge, the generated human geometry and texture are inadequate. Although integrating 3D SMPL-X information improves the geometric quality, the enhancement is minimal. In contrast, our SLE module can effectively align the generated human body with the target quality. On the right subfigure, we observe that the generated mesh lacks fine details like wrinkles without our proposed WLR module. However, applying the WLR largely enhances the detail and realism of the human mesh.

5. Conclusion

In this paper, we introduce MultiGO, a novel multi-level geometric learning framework for monocular 3D human reconstruction. By incorporating a Skeleton-Level Enhancement module, Joint-Level Augmentation strategy, and Wrinkle-Level Refinement module, MultiGO leverages body geometric priors to enhance reconstruction quality across various granularity levels, resulting in more accurate and detailed reconstructions. Extensive experiments on two test sets, CustomHuman and THuman3.0, demonstrate that the proposed method achieves SOTA performance.

6. Acknowledgement

This research is supported by the National Natural Science Foundation of China (No. 62406267), Tencent Rhino-Bird Focused Research Program, Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2025A03J3956 & Grant No.2023A03J0008), the Guangzhou Municipal Science and Technology Project (No. 2025A04J4070), the Guangzhou Municipal Education Project (No. 2024312122) and Education Bureau of Guangzhou Municipality.

References

- [1] Renderpeople. <https://renderpeople.com/>. Accessed: 2024-11-10. 6
- [2] Mario Botsch and Leif Kobbelt. A remeshing approach to multiresolution modeling. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 185–192, 2004. 5
- [3] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2
- [4] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21024–21035, 2023. 6
- [5] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 6
- [6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 3
- [7] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2
- [8] Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20418–20431, 2024. 2
- [9] Yuheng Jiang, Zhehao Shen, Penghao Wang, Zhuo Su, Yu Hong, Yingliang Zhang, Jingyi Yu, and Lan Xu. Hifi4g: High-fidelity human performance rendering via compact gaussian splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19734–19745, 2024. 2
- [10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 2023. 2, 3
- [11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 3
- [12] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. HUGS: Human gaussian splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [13] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 2
- [14] Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner, 2024. 4
- [15] Leyuan Liu, Yuhan Li, Yunqi Gao, Changxin Gao, Yuanyuan Liu, and Jingying Chen. VS: Reconstructing clothed 3d human from single image via vertex shift. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10498–10507, 2024. 1, 2, 6
- [16] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023. 3
- [17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2, 3
- [18] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2
- [19] Werner Palfinger. Continuous remeshing for inverse rendering. *Computer Animation and Virtual Worlds*, 33(5):e2101, 2022. 5
- [20] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1165–1175, 2024. 2
- [21] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2
- [22] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. 2024. 2
- [23] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 5
- [24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3
- [25] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 3

- [26] 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 *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 6
- [27] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3
- [28] Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. Deepcloth: Neural garment representation for shape and style editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1581–1593, 2023. 6
- [29] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. *arXiv preprint arXiv:2312.13150*, 2023. 2
- [30] Jiayang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 2, 4, 5
- [31] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3405–3414, 2019. 6
- [32] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 3
- [33] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023. 2
- [34] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*. 5
- [35] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*, 2024. 2
- [36] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv preprint arXiv:2405.20343*, 2024. 5
- [37] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, 2022. 1, 2, 6
- [38] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 6
- [39] Dejia Xu, Ye Yuan, Morteza Mardani, Sifei Liu, Jiaming Song, Zhangyang Wang, and Arash Vahdat. Agg: Amortized generative 3d gaussians for single image to 3d. *arXiv preprint arXiv:2401.04099*, 2024. 2
- [40] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217*, 2023. 2
- [41] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024. 2, 4
- [42] Yifan Yang, Dong Liu, Shuhai Zhang, Zeshuai Deng, Zixiong Huang, and Mingkui Tan. Hilo: Detailed and robust 3d clothed human reconstruction with high-and low-frequency information of parametric models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10671–10681, 2024. 1, 2, 6
- [43] Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Trans. Graph.*, 42(4), 2023. 4, 7
- [44] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. *arXiv preprint arXiv:2404.19702*, 2024. 2
- [45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6
- [46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4
- [47] Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 6
- [48] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9936–9947, 2024. 1, 2, 6
- [49] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [50] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars, 2023. 2