sensors

MDPI

Article

Lightweight Explicit 3D Human Digitization via Normal Integration

Jiaxuan Liu [1,2], Jingyi Wu [1], Ruiyang Jing [1], Han Yu [1], Jing Liu [1] and Liang Song [1,2,*]

[1] Academy for Engineering and Technology, Fudan University, Shanghai 200433, China; jiaxuanliu22@m.fudan.edu.cn (J.L.); jingyiwu23@m.fudan.edu.cn (J.W.); ryjing22@m.fudan.edu.cn (R.J.)
[2] Innovation Platform for Academicians of Hainan Province, Haikou 570228, China
* Correspondence: songl@fudan.edu.cn

Abstract: In recent years, generating 3D human models from images has gained significant attention in 3D human reconstruction. However, deploying large neural network models in practical applications remains challenging, particularly on resource-constrained edge devices. This problem is primarily because large neural network models require significantly higher computational power, which imposes greater demands on hardware capabilities and inference time. To address this issue, we can optimize the network architecture to reduce the number of model parameters, thereby alleviating the heavy reliance on hardware resources. We propose a lightweight and efficient 3D human reconstruction model that balances reconstruction accuracy and computational cost. Specifically, our model integrates Dilated Convolutions and the Cross-Covariance Attention mechanism into its architecture to construct a lightweight generative network. This design effectively captures multi-scale information while significantly reducing model complexity. Additionally, we introduce an innovative loss function tailored to the geometric properties of normal maps. This loss function provides a more accurate measure of surface reconstruction quality and enhances the overall reconstruction performance. Experimental results show that, compared with existing methods, our approach reduces the number of training parameters by approximately 80% while maintaining the generated model's quality.

Keywords: three-dimensional human reconstruction; normal map estimation; a skinned multi-person linear model; deep learning

check for updates

Academic Editor: Stefano Berretti

Received: 4 January 2025
Revised: 9 February 2025
Accepted: 27 February 2025
Published: 28 February 2025

Citation: Liu, J.; Wu, J.; Jing, R.; Yu, H.; Liu, J.; Song, L. Lightweight Explicit 3D Human Digitization via Normal Integration. *Sensors* **2025**, *25*, 1513. https://doi.org/10.3390/s25051513

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Recent advancements in deep learning and computer vision have significantly heightened interest in 3D reconstruction, which has emerged as a critical technological enabler across diverse domains. 3D reconstruction involves utilizing software and hardware technologies to digitally reconstruct scenes, objects, or individuals in three dimensions. Compared with 2D data, 3D information inherently encapsulates more comprehensive and precise spatial characteristics, enabling 3D models to deliver richer contextual insights than conventional 2D imagery. This capability allows computers to preserve and perceive the three-dimensional world better, transcending the limitations of 2D data representation. Currently, 3D reconstruction has been widely adopted in fields such as autonomous driving, 3D hologram generation and reconstruction [1], 3D shape reconstruction [2], and structure-from-motion photogrammetry [3], empowering innovative applications across industries.

With the evolution of 3D reconstruction techniques, 3D human reconstruction has progressively gained prominence. Recent developments have achieved remarkable progress,

demonstrating substantial practical value and potential, thereby enabling broad applications in areas such as virtual reality (VR), human–computer interaction (HCI) [4], healthcare [5], 3D feature extraction [6], sports analytics [7], and cultural heritage conservation [8].

The widespread application potential of 3D human reconstruction across multiple domains has made the efficient reconstruction of human models from images a critical research direction in computer vision [9,10] and computer graphics [11]. However, achieving high-fidelity reconstruction remains challenging due to the human body's complex geometric structures, appearances, and articulated nature. In recent years, significant progress has been made in non-parametric reconstruction methods. PIFu (Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization) [12] and PIFuHD (Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization) [13] leverage implicit functions and utilize scanned 3D human data as supervision to achieve detailed reconstruction of clothed humans. However, they do not incorporate structural priors of the human body. Subsequently, PaMIR (Parametric Model-Conditioned Implicit Representation for Image-Based Human Reconstruction) [14] introduced the SMPL model (A Skinned Multi-Person Linear Model) [15] as a prior, significantly improving pose estimation [16] accuracy and overall reconstruction precision. However, it relies solely on implicit functions for representation, lacking more refined reconstruction techniques. Parametric models such as SMPL and SMPL-X have been widely adopted in virtual reality, game development, and virtual try-on systems. Compared with the nude human models reconstructed by SMPL and SMPL-X, our method achieves more detail and can reconstruct complex clothing.

With the growing demand for realistic human models across various fields, many methods have proposed innovative architectures and training techniques to enhance reconstruction performance. However, challenges persist, particularly the limitations of large-scale network models in practical applications. The core focus of this study is to reconstruct high-precision 3D human models from a single 2D image, emphasizing normal map estimation and reconstruction. Additionally, human pose estimation [16] and texture generation techniques are employed. While occlusion handling and dynamic scene reconstruction are significant challenges in 3D modeling, these aspects fall outside the scope of this work. We utilize the SMPL-X model as a prior for human pose estimation. By adopting the concept of normal map estimation, we generate front and back normal maps of the human body. The surface 3D information is represented by the normal maps computed by the generative network. To efficiently generate human surface normal maps, we design a lightweight and efficient generative network architecture. This architecture leverages the expanded receptive fields of Dilated Convolutions [17] and the cross-channel feature extraction capabilities of Cross-Covariance Attention [18]. During this process, the front and back of the human body are processed separately. The generated normal maps are then used to recover the mesh model through the d-BiNI (depth-aware silhouette-consistent bilateral normal integration) method [19], combined with the SMPL-X model (Expressive Body Capture: 3D Hands, Face, and Body from a Single Image) [20] to complete the mesh reconstruction. During training, we introduce a surface folds loss function [21] to generate more detailed fold variations and a spatial loss function [22] to measure the distances between points in 3D space. These two loss functions jointly constrain the normal map generation network, ensuring the network optimizes in the correct direction.

Figure 1 presents the reconstruction results of different human models. To reduce the parameter count of the generative network, we combine Dilated Convolutions with a Cross-Covariance Attention mechanism to build a lightweight network. Furthermore, we propose an innovative loss function to evaluate normal maps better. The final results

demonstrate a balance between high reconstruction quality and model efficiency. Our contributions include the following:

- We propose an innovative human reconstruction method incorporating an innovative loss function designed to optimize the training process. This loss function significantly enhances the accuracy and detail of reconstructions, as evidenced by the promising results achieved in our experiments;
- We also introduce a lightweight generative network designed to produce high-quality surface normal maps of the human body. This network employs efficient architectural designs to capture intricate geometric details effectively. Experimental results confirm the robustness and efficiency of our approach in generating accurate and realistic surface normal maps;
- Extensive experiments were conducted to evaluate our proposed method comprehensively. These assessments, comprising both quantitative and qualitative analyses, underscore the superiority of our approach. The results demonstrate consistent performance across diverse scenarios, highlighting its robustness and effectiveness in reconstructing detailed human models.



Figure 1. This figure demonstrates our method's reconstruction capability. Our approach successfully generates detailed 3D human models. The top row shows the input images, the middle row presents the generated normal maps, and the bottom row displays the reconstructed 3D human models.

The structure of this paper is organized as follows: Section 1 systematically elaborates on the research background, problem definition, and the academic significance of this work. Section 2 reviews the latest research progress in related fields, focusing on analyzing the strengths and limitations of existing methods. Section 3 provides a detailed description of the proposed algorithmic framework and network architecture design. Section 4 thoroughly investigates the effectiveness and performance of the proposed method through systematic experimental validation and comparative analysis. Finally, Section 5 summarizes the main innovative contributions of this study, objectively discusses the current limitations, and outlines potential directions for future research.

2. Related Works

In 3D human reconstruction, mainstream approaches can be categorized into "parametric" and "non-parametric" methods. Parametric methods focus on reconstructing human models by leveraging statistical human models. These methods can be further divided into those that learn body shape and pose-related corrections through statistical techniques [23,24] and those that address non-linear deformations and compensate for artifacts caused by linear blend skinning [25,26]. Parametric reconstruction relies on low-dimensional vectors to represent body shapes. This approach accurately reconstructs human shape and pose while reducing data representation complexity. Standard parametric models include SCAPE (SCAPE: shape completion and animation of people) [23], SMPL [15], and SMPL-X [20]. These models infer human parameters through optimization or regression methods. For example, SMPLify (Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image) [27] optimizes SMPL parameters (including shape and pose) by minimizing the reprojection error between detected 2D key points and a synthesized 3D pose while incorporating penetration constraints to reduce 2D-to-3D ambiguity. HMR (end-to-end recovery of human shape and pose) [28] introduces the reprojection error of human joints into the loss function to supervise SMPL pose and shape parameters. Inspired by Generative Adversarial Networks (GANs) [29], HMR also incorporates a discriminator into the loss function to ensure the validity of predicted human parameters. SPIN (Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop) [30] combines regression- and optimization-based approaches for 3D pose and shape estimation. It uses a regression network to generate SMPL parameters as initial values for an iterative fitting module. The optimization results of the iterative fitting module are then used as supervision to improve the regression network, enabling cyclic training for enhanced generation.

Non-parametric methods primarily focus on generating detailed surface models of clothed human bodies. To accommodate clothed human shapes, numerous approaches [12,19] achieve precise clothing modeling by adjusting mesh vertices. Implicit representations have also been widely adopted in non-parametric 3D reconstruction. For instance, PIFu [12] introduces pixel-aligned implicit reconstruction, while PIFuHD [13] enhances geometric details through a multi-level architecture and normal maps predicted from RGB images. However, these methods do not incorporate structural priors of the human body, leading to suboptimal reconstruction results. GeoPIFu (GeoPIFu: geometry and pixel-aligned implicit functions for single-view human reconstruction) [31] incorporates rough volumetric human shapes and methods like Self-Portrait [32], PINA (Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence) [33], and S3 (S3: Neural Shape, Skeleton, and Skinning Fields for 3D Human Modeling) [34] utilize depth or LiDAR data to regularize shapes and improve robustness to pose variations. PaMIR [14] and DeepMultiCap [35] align pixel-aligned features on posed voxelized SMPL meshes, using parametric models as priors to ensure completeness and stability. However, their reconstruction results are often coarse and unsatisfactory in many cases. ARCH (ARCH: Animatable Reconstruction of Clothed Humans) [36], ARCH++ (ARCH++: Animation-Ready Clothed Human Reconstruction Revisited) [37], and CAR (High-fidelity clothed avatar reconstruction from a single image) [38] use SMPL to map pixel-aligned query points from the posed space to a canonical space. ICON (ICON: Implicit Clothed Humans Obtained from Normals) [39] extends this to unseen poses in in-the-wild images through local feature regression, achieving impressive reconstruction quality. However, it underutilizes the generated normal maps and suffers from a large model size. ECON (ECON: Explicit Clothed Humans Optimized via Normal Integration) [19] introduces the d-BiNI technique and adopts ResNet [40] as the backbone for its generative network, achieving

promising results. Nevertheless, there remains room for optimization regarding model size and inference speed.

In generative networks, an effective architecture design is crucial. R-MSFM (R-MSFM: Recurrent Multi-Scale Feature Modulation for Monocular Depth Estimating) [41] proposes a compact feature modulation module to learn multi-scale features, achieving competitive performance using only the first three stages of ResNet-18 as the backbone. Zhang et al. [42] propose a hybrid architecture combining convolutions and transformers, effectively reducing the model size while maintaining performance. Bae et al. [43] enhance CNN features with Transformers to achieve state-of-the-art accuracy, but using multiple parallel modules slows down processing speed. ICON and ECON employ ResNet [40] as the backbone for their normal map generation networks, achieving high-quality results. However, their large model sizes and slow inference speeds limit their practicality.

3. Methods

3.1. Human Models Reconstruction

Generating a 3D human model from a single image is challenging, often resulting in abnormal outputs. To address this, we incorporate a parametric human model (SMPL-X) [20] as prior knowledge to ensure human pose accuracy. This approach has been effectively utilized in methods such as PaMIR [14], ICON [39], and ECON [19]. By estimating the pose from the input image and generating a corresponding SMPL-X model [20], we infer the surface normal map of the human body based on the input image and the SMPL-X model to produce accurate normal maps.

As illustrated in Figure 2A, we first perform SMPL-X model estimation on the reconstructed image and utilize a Differentiable Renderer to generate the surface normal map N^b from the reconstructed SMPL-X model M^b. Subsequently, the image and the normal map are fed into two independent Lite-GN networks to produce front-view and back-view normal maps. These outputs are then integrated through the d-BiNI method and a human shape completion stage to reconstruct a complete 3D human model. Lite-GN primarily consists of a Dilated Convolutions Block and a Cross-Covariance Attention Block. The detailed architecture of Lite-GN is illustrated in Figure 3, while the specific structures of the Dilated Convolutions Block and the Cross-Covariance Attention Block are shown in Figure 4. To refine the surface geometry during training, we jointly optimize two refinement loss terms (as illustrated in Figure 2B)—namely, the fold refinement loss and the spatial refinement loss—by comparing the predicted normal maps with their ground truth. This approach effectively minimizes surface irregularities and geometric discrepancies.

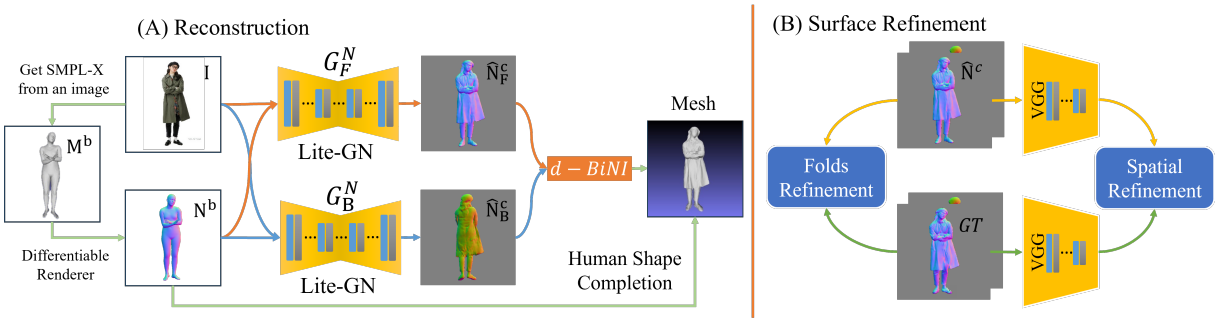


Figure 2. Overview. (**A**) Pipeline for 3D human reconstruction; (**B**) architectural framework of the loss function for normal map generation.

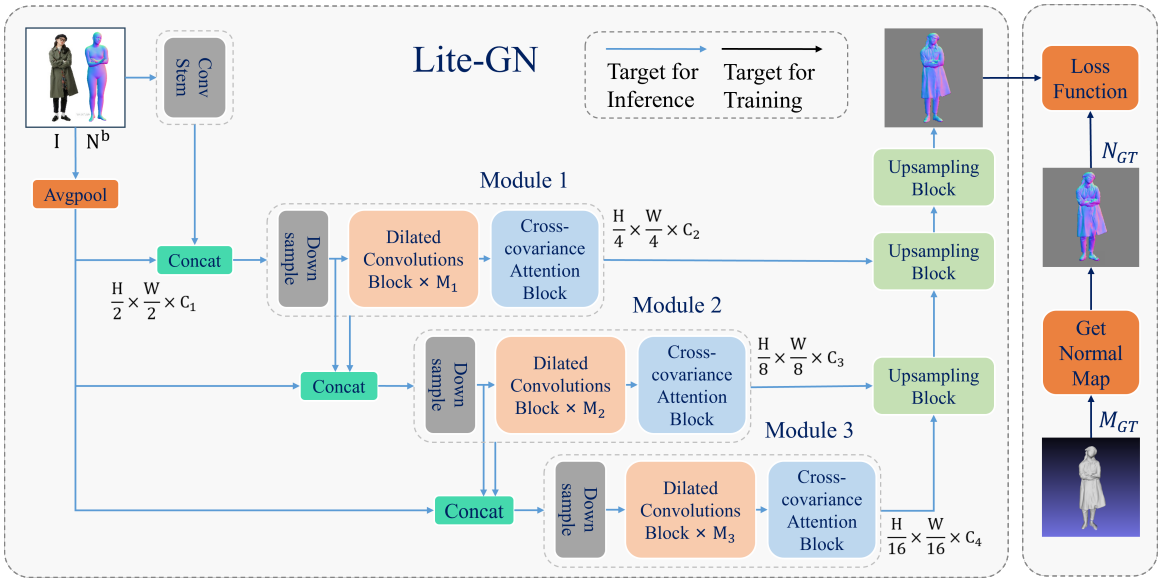


Figure 3. Overview of Lite-GN. The generative network is designed using an encoder–decoder architecture. Each module comprises M_N Dilated Convolution Blocks and a Cross-Covariance Attention Block.

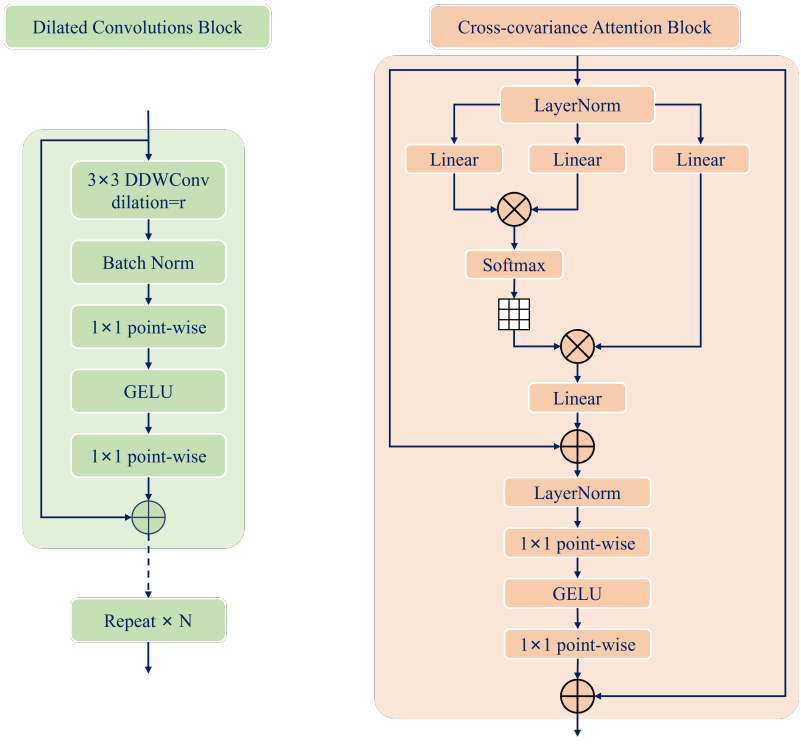


Figure 4. The detailed architectures of the Dilated Convolution Block and Cross-Covariance Attention Block are illustrated.

We first perform pose estimation on the input image and generate an SMPL-X model in the corresponding pose. Our approach generates clothed human models based on the SMPL-X model to reduce ambiguity and guide the prediction of the clothed body from both the front and back perspectives. Using PyTorch3D's (version 0.7.1, developed by Facebook AI Research, Menlo Park, CA, USA) differentiable renderer (denoted as DR), we extract surface normal maps for the SMPL-X model from both front and back views, as illustrated in Figure 2. The input image and the normal maps $\hat{N}^b = \left\{\hat{N}_F^b, \hat{N}_B^b\right\}$ are then fed into the

generative network $LiteG^N = \{G_F^N, G_B^N\}$, which outputs the normal maps of the human body for the front and back views, denoted as $\hat{N}^c = \{\hat{N}_F^c, \hat{N}_B^c\}$.

$$DR\left(M^b\right) = N^b, \tag{1}$$

$$LiteG^N\left(N^b, I\right) = \hat{N}^c, \tag{2}$$

where M^b refers to the SMPL-X model generated based on pose parameters, and $\hat{N}^c = \{\hat{N}_F^c, \hat{N}_B^c\}$ is the predicted normal map generated by G^N.

We employ a depth-aware silhouette-consistent bilateral normal integration method (d-BiNI) for the generated normal maps. This technique utilizes the prior information from the SMPL-X model. It enforces consistency along the silhouette to ensure alignment between high-frequency surface details and the predicted dense normal maps. At the same time, it maintains low-frequency surface variations, including discontinuities, by the SMPL-X model. The objective function [19] is composed of five primary components:

$$d\text{-}BiNI\left(\hat{N}_F^c, \hat{N}_B^c, Z_F^b, Z_B^b\right) \rightarrow \hat{Z}_F^c, \hat{Z}_B^c, \tag{3}$$

$$\begin{aligned} \min_{\hat{Z}_F^c, \hat{Z}_B^c} \mathcal{L}_n\left(\hat{Z}_F^c, \hat{N}_F^c\right) + \mathcal{L}_n\left(\hat{Z}_B^c, \hat{N}_B^c\right) + \lambda_d \mathcal{L}_d\left(\hat{Z}_F^c, Z_F^b\right) + \\ \lambda_d \mathcal{L}_d\left(\hat{Z}_B^c, Z_B^b\right) + \lambda_s \mathcal{L}_s\left(\hat{Z}_F^c, \hat{Z}_B^c\right), \end{aligned} \tag{4}$$

where $\mathcal{L}_n$ is the BiNI [44] loss term, $\mathcal{L}_d$ is the prior depth of the front and back depth surfaces, $\mathcal{L}_s$ is the front and back silhouette consistency term, and Z_b^* and $\hat{Z}_*^c$ represent the coarse body depth image and the clothed body depth image. The detailed calculation formula for $\mathcal{L}_d\left(\hat{Z}_i^c, Z_i^b\right)$ and $\mathcal{L}_s\left(\hat{Z}_F^c, \hat{Z}_B^c\right)$ [19] are as follows:

$$\mathcal{L}_d\left(\hat{Z}_i^c, Z_i^b\right) = \left|\hat{Z}_i^c - Z_i^b\right|_{\Omega_n \cap \Omega_z} \quad i \in \{F, B\}, \tag{5}$$

$$\mathcal{L}_s\left(\hat{Z}_F^c, \hat{Z}_B^c\right) = \left|\hat{Z}_F^c - \hat{Z}_B^c\right|_{\partial\Omega_n}, \tag{6}$$

where Ω_n and Ω_z represent the domains of the clothed and body regions, and $\partial\Omega_n$ denotes the silhouette of Ω_n.

Refining SMPL-X: Our approach utilizes the SMPL-X model as prior information; the accuracy of pose parameters extracted from images is critical for achieving precise reconstruction results. Accurate parameters provide more reliable priors, thereby enhancing the quality of the generated outcomes. To this end, we employ the PIXIE method [45] to extract SMPL-X parameters. However, the parameters directly inferred from the human model often fail to fully align with the poses and shapes depicted in the input images. To address this limitation, we further refine the SMPL-X model parameters through an optimization process [30,39]. Specifically, we optimize the shape parameter β, pose parameter θ, and translation parameter t of the SMPL-X model to ensure that the generated model closely aligns with the poses and appearances of the input images. The objective of this optimization process is to minimize the following loss function [19]:

$$L_{SMPL-X} = \min_{\theta,\beta,t}(\lambda_{N\_diff} L_{N\_diff} + L_{S\_diff}), \tag{7}$$

$$L_{N\_diff} = \left|N^b - \hat{N}^c\right|, \qquad L_{S\_diff} = \left|S^b - \hat{S}^c\right|, \tag{8}$$

where L_{N_diff} represents the normal map loss (L1), weighted by λ_{N_diff}, and L_{S_diff} is the L1 loss measuring the difference between the silhouette of the SMPL body normal map S_b and the human mask S_{bc}, which is segmented from I as described in Rembg [46].

Human shape completion: We adopt the method of ECON-EX [19,47] to complete the surface reconstruction of the human body, using the SMPL-X model to fill in the missing surface. We remove triangles of the generated mesh from the SMPL-X model M^b. The remaining triangles include side view boundaries and occluded areas, M^r. We obtain a complete human body model by using PSR [47] to merge the generated surface with the remaining surface.

3.2. The Proposed Framework: Lite-GN

Several studies have shown that a well-designed encoder can extract more effective features, leading to improved generative results [48,49]. We designed a lightweight generative network that efficiently generates normal maps from input images. The network consists of an encoder (Encoder Net) and a decoder (Decoder Net), as illustrated in Figure 3. To reduce the model size effectively, we utilize stacked Dilated Convolutions [17] to decrease the network depth while maintaining a large receptive field. Additionally, we incorporate a Transformer [50] to learn global features and employ Cross-Covariance Attention [18] to focus on inter-channel relationships. This architect significantly reduces the computational complexity.

Encoder: We achieve efficient feature encoding by introducing multi-scale information fusion. The process is as follows: First, the input human image is processed through four 3×3 convolutional downsampling layers to extract local features, resulting in a feature map of size $256 \times 256 \times 64$. In the Module 1 stage, the downsampled feature map is concatenated with the pooled input image. It is followed by two 3×3 convolutional downsampling operations, producing a feature map of size $128 \times 128 \times 64$. The features are then passed through a Dilated Convolutions Block to expand the receptive field, a process repeated six times. The resulting features are subsequently fed into a Cross-Covariance Attention Block to incorporate cross-covariance mechanisms. In the Module 2 stage, the input feature map is processed through two 3×3 convolutional operations, producing a feature map of size $64 \times 64 \times 128$. The features are then passed through a Dilated Convolutions Block, repeated six times, and subsequently into a Cross-Covariance Attention Block. In the Module 3 stage, the input feature map undergoes two 3×3 convolutional operations, generating a feature map of size $32 \times 32 \times 224$. These features are then processed through a Dilated Convolutions Block, repeated 18 times, followed by another Cross-Covariance Attention Block. It marks the completion of the encoding phase. As shown in Table 1, $[Conv3 \times 3]$ indicates using a 3×3 convolutional kernel.

Decoder: We adopt a decoder adapted from monodepth [48]. As illustrated in Figure 3, we employ multiple bilinear upsampling layers for effective upsampling, while convolutional layers connect features from the three stages of the encoder. This design ensures deep fusion of high-level and low-level features, preserving low-dimensional information while minimizing information loss. The generated normal map is output at the topmost layer.

Table 1. The network parameters of the Encoder.

Output Size	Stage	Layer
$512 \times 512 \times 6$	Input	
$256 \times 256 \times 64$	*Conv* Stem	$[Conv3 \times 3] \times 4$

Table 1. *Cont.*

Output Size	Stage	Layer
128 × 128 × 64	Stage 1	[*Conv*3 × 3] × 2 Dilated Convolutions Block × 6 Cross-Covariance Attention Block × 1
64 × 64 × 128	Stage 2	[*Conv*3 × 3] × 2 Dilated Convolutions Block × 6 Cross-Covariance Attention Block × 1
32 × 32 × 224	Stage 3	[*Conv*3 × 3] × 2 Dilated Convolutions Block × 18 Cross-Covariance Attention Block × 1

3.2.1. Dilated Convolutions Block

We employ multiple layers of Dilated Convolutions to extract features from the input image and the SMPL-X normal maps. Stacking several dilated convolution layers enhances feature integration and fusion.

For a given input feature $x[i]$, the resulting output feature $y[i]$ produced by a dilated convolution [17] can be expressed as follows:

$$y[v] = \sum_{u=1}^{U} x[v + r \cdot u]w[u], \tag{9}$$

where $w[u]$ denotes a filter of size U, and r specifies the dilation rate applied to the input $x[v]$ during the convolution operation. In the case of standard convolution (without dilation), the dilation rate r is set to 1.

By employing dilated convolution, the network can maintain the size of the output feature map while achieving an expanded receptive field. The output of our Dilated Convolutions Block is as follows:

$$\hat{X} = Linear(G(Linear(BN(DDWConv_r(X))))) + X, \tag{10}$$

where *Linear* refers to a point-wise convolution, and G represents the GELU [51] activation function. BN denotes a batch normalization layer, while $DDWConv_r(X)$ refers to a 3×3 depthwise dilated convolution with a dilation rate of r.

3.2.2. Cross-Covariance Attention Block

In the Cross-Covariance Attention Block, the input feature X is processed through point-wise convolutions to generate queries $Q = XW_q$, keys $K = XW_k$, and values $V = XW_v$, where W_q, W_k, and W_v are weight matrices. Cross-Covariance Attention [18] efficiently handles high-resolution images by combining the precision of traditional Transformers with the scalability of convolutional architectures. This approach effectively reduces the quadratic complexity of attention computation to linear complexity:

$$\tilde{X} = X + Attention(Q, K, V), \tag{11}$$

where $Attention(Q, K, V) = V \cdot Softmax(Q^T \cdot K)$ [18].

Additionally, we enhance the non-linearity of the features as follows:

$$\hat{X} = X + Linear(G(Linear(LN(\tilde{X})))), \tag{12}$$

where LN denotes the layer normalization [52] operation, *Linear* represents a point-wise convolution, and G is the GELU [53] activation function.

3.3. Surface Refinement

The VGG Loss [21] enhances the perceptual quality of generated images by performing comparisons in feature space, making it particularly suitable for tasks requiring high fidelity and structural consistency. Instead of directly comparing image differences in pixel space, VGG Loss computes the loss using intermediate feature representations extracted from a pre-trained VGG network [21]:

$$L_{VGG} = \frac{1}{h_l w_l c_l} \sqrt{\sum_{i,j,k} \left(\phi^{(l)}_{i,j,k}(\hat{N}^c) - \phi^{(l)}_{i,j,k}(N_{gt}) \right)^2}, \tag{13}$$

where h_l, w_l, and c_l represent the feature map's height, width, and number of channels at layer l, respectively. The network is denoted as ϕ, and the high-level representation extracted at layer l is $\phi^{(l)}(N)$. The Euclidean distance is computed between the high-level representations of N_{gt} and $\hat{N}^c$ at the corresponding layer.

Thus, the VGG network's feature representations are closer to the human visual system's understanding of images. This allows it to better capture high-level semantic information and perceptual quality, making it more effective in reflecting normal maps' semantic information and quality.

The VGG Loss focuses solely on perceptual differences but disregards spatial information. To address this, we employ Huber Loss [22] to measure the spatial discrepancy between two normal maps. The Huber Loss addresses the limitations of the VGG loss by directly comparing the differences between two binary vectors at each position, precisely capturing spatial information. It is because the VGG loss, based on feature maps extracted from a pre-trained network, primarily focuses on perceptual similarity (e.g., textures and styles) in images but overlooks the geometric alignment of vertices in 3D space. In contrast, the Huber Loss quantifies spatial inconsistencies by counting the number of differing elements at corresponding positions in the two vectors, forcing the model to prioritize geometric accuracy during optimization. The formula for Huber Loss [22] is as follows:

$$L_\delta(N_{gt}, \hat{N}^c) = \begin{cases} \frac{1}{2}(N_{gt} - \hat{N}^c)^2, & \left|N_{gt} - \hat{N}^c\right| \le \delta \\ \delta\left|N_{gt} - \hat{N}^c\right| - \frac{1}{2}\delta^2, & \left|N_{gt} - \hat{N}^c\right| > \delta' \end{cases} \tag{14}$$

where δ represents a hyperparameter that determines the balance between MSE and MAE. As a result, Huber Loss combines the advantages of both MSE and MAE.

So, we train the Generate network, Lite-GN, with the following loss:

$$L_N = L_\delta(N_{gt}, \hat{N}^c) + \lambda_{VGG} \mathcal{L}_{VGG}, \tag{15}$$

where λ_{VGG} denotes the weight of the VGG loss.

During the training phase (as illustrated in Figure 3), we extract the front and back normal maps N_{gt} from the ground truth mesh M_{gt} corresponding to the input image. These N_{gt} maps, along with the generated normal maps $\hat{N}_c$, are fed into the loss function L_n to train the normal map generation network, *LiteGN*. The specific details of the loss function are illustrated in Figure 2B and described by Equation (15).

4. Experiments and Results

4.1. Implementation Details

Datasets: We utilize the THuman2.0 dataset [54] as our training set to train the generative network. THuman2.0 consists of 525 high-quality human scan models and has been used to train methods such as ECON [19], ICON [39], IF-Nets (Implicit Function Networks) [19], PIFu [12], and PaMIR [14]. Additionally, we use the CAPE [55] and Render-

people datasets [56] for quantitative evaluation. The CAPE dataset assesses the robustness of reconstruction methods under complex poses, while the Renderpeople dataset evaluates robustness in handling intricate clothing.

The THuman2.0 dataset is acquired using high-precision scanners, providing high-resolution textured 3D human models. Each model is accompanied by corresponding SMPL and SMPL-X parameters for the respective poses (as illustrated in Figure 5).

Training stage: We use PIXIE (Collaborative regression of expressive bodies using moderation) [45] to extract SMPL-X parameters from the input images, which are then utilized to construct the SMPL-X model. The AdamW optimizer is employed with an initial learning rate of 10^{-3}, which is reduced by a factor of 0.1 every 10 epochs. The training process spans 50 epochs with a batch size of 4. Detailed experimental environment parameters are shown in Table 2.



Figure 5. Examples from the THuman2.0 dataset.

Table 2. Experimental environment's parameters.

Equipment	Computer Configuration Parameters
Operating system	Linux
RAM	32 G
Type of operating system	Ubuntu20.04
CPU	Intel Core i7-12700K
GPU	RTX 4090(24 GB) × 1
Development language	Python 3.8
Deep learning framework	PyTorch 1.12.1

4.2. Evaluation Metrics

Chamfer distance: Chamfer distance [57] is a widely used metric in 3D geometry processing to evaluate the similarity between two point sets. Precisely, it measures the average squared distance from each point in the ground truth scan to its nearest neighbor in the generated model. This symmetric calculation ensures that both point sets are equally accounted for in the evaluation. Chamfer distance is particularly effective for assessing the overall geometric alignment between the ground truth and the reconstructed model, making it a reliable indicator of the quality of the generated result. A lower Chamfer distance suggests the reconstructed model closely approximates the actual geometry.

P2S distance: Point-to-Surface (P2S) distance [47] is a more refined metric that measures the nearest distance from each point in the ground truth scan to the closest surface in the reconstructed model. Unlike Chamfer distance, which treats both datasets as point clouds, P2S distance incorporates the reconstructed model's surface information. It makes it especially valuable for evaluating how well the reconstructed surface conforms to the underlying geometry of the ground truth. By focusing on the proximity between points and surfaces, P2S distance provides a nuanced understanding of reconstruction accuracy, particularly in areas where precise surface details are critical.

Normals difference: Normals difference [19] assesses the angular difference between the normal vectors at corresponding points in the ground truth and reconstructed models. Normal vectors encode the local orientation of surfaces and are essential for capturing fine geometric details and high-frequency structures. By comparing these vectors, normal differences highlight discrepancies in local surface characteristics that may not be evident from positional metrics alone. This metric is crucial for evaluating the preservation of intricate details, such as sharp edges and subtle surface undulations, in the reconstruction process. Smaller differences in normals indicate a higher fidelity reconstruction of local geometric features.

4.3. Evaluation

4.3.1. Quantitative Evaluation

Our method enables the reconstruction of detailed 3D human models from a single image. It can generate complete human models for individuals of any gender or role. Even in complex poses and clothing cases, our approach maintains high reconstruction quality, as illustrated in Figure 1. The method accurately reconstructs clothing and poses while preserving critical facial features.

We conducted a comprehensive comparison of our method with several state-of-the-art approaches, including ECON [19], ICON [39], PaMIR [14], and PIFuHD [13]. To thoroughly evaluate the performance of each method, we employed four key metrics: Chamfer Distance, Point-to-Surface (P2S) Distance, Normals Difference, and Weight Size, which collectively measure the superiority of the approaches. The results of the comparison are presented in Table 3. For a fair evaluation, all methods were trained using the THuman2.0 dataset [54], which provides high-quality 3D human scans, and tested on the CAPE dataset [55], a challenging benchmark that captures diverse clothing variations. The experimental results demonstrate that our method achieves generation quality comparable to PaMIR and ICON regarding visual fidelity and structural accuracy, ensuring highly competitive results. Although some metrics are slightly inferior to ECON-EX, our approach significantly reduces the model size, highlighting its superior lightweight characteristics. As shown in Table 4, our method achieves more outstanding results regarding lightweight efficiency.

We also performed a focused comparison with other methods that utilize normal maps for 3D reconstruction. In this comparison, we evaluated the performance of the normal map generation networks across three key metrics: the number of network parameters, the size of the model weight files, and the runtime during inference. As shown in Table 4, our method substantially reduces resource consumption. Specifically, our approach minimizes network parameters and weight file size while significantly shortening the runtime, enabling faster processing and reduced computational overhead. This efficient resource utilization highlights the practical advantages of our method, especially in scenarios where computational resources are constrained or real-time performance is critical. By balancing quality and efficiency, our approach is well suited for applications requiring lightweight 3D human reconstruction.

Table 3. Quantitative comparison on CAPE dataset.

Methods	Chamfer ↓	P2S ↓	Normals ↓	Weight Size ↓
PIFuHD [13]	3.767	3.591	0.0994	1.4 GB
PaMIR [14]	0.989	0.992	0.0422	472 MB
ICON [39]	0.971	0.909	0.0409	1.38 GB
ECON-IF [19]	0.996	0.967	0.0413	1.49 GB
ECON-EX [19]	0.926	0.917	0.0367	1.35 GB
Ours	0.965	0.930	0.0472	201.6 MB

↓ indicates that lower values are better.

Table 4. Model complexity and speed evaluation. We evaluated and compared the parameters, model weight file size, and inference speed of different methods.

Methods	Model Size ↓	Weight Size ↓	Speed (ms) ↓
ICON [39]	345.6 M	1.35 GB	22.8 ms
ECON [19]	345.6 M	1.35 GB	23.6 ms
Ours	50 M	201.6 MB	12.8 ms

↓ indicates that lower values are better.

Additionally, we evaluated our approach on the Renderpeople dataset to test its robustness to various types of clothing under complex apparel conditions. We also adopted Chamfer Distance, P2S Distance, and Normals Difference as key evaluation metrics. The results indicate that our method achieves high-quality generation outcomes on the RenderPeople dataset, which is consistent with its performance on the CAPE dataset. While the generation quality surpasses most comparative methods, it is slightly inferior to ECON-EX, as shown in Table 5. However, our approach significantly reduces the model size, demonstrating its superior lightweight efficiency.

Table 5. Quantitative comparison on Renderpeople dataset.

Methods	Chamfer ↓	P2S ↓	Normals ↓
PIFuHD [13]	1.946	1.983	0.0658
PaMIR [14]	1.296	1.430	0.0518
ICON [39]	1.373	1.522	0.0566
ECON-IF [19]	1.401	1.422	0.0516
ECON-EX [19]	1.342	1.458	0.0478
Ours	1.353	1.430	0.0452

↓ indicates that lower values are better.

4.3.2. Qualitative Comparison

We compared our method against other human reconstruction approaches. The test images were divided into two "complex poses" and "complex clothing". As shown in Figure 6, our method effectively reconstructs the visual characteristics of the human body. The first row presents the input images, serving as the input for monocular reconstruction. On the left are images with complex human poses, showcasing our method's capability to handle challenging poses. On the right are examples of intricate clothing, highlighting the robustness of our approach to complex apparel. The qualitative comparison results align with the quantitative findings in Table 3. The results demonstrate that our method significantly outperforms PIFuHD [13] and PaMIR [14] while achieving comparable performance to ECON [19] and ICON [39], with the added advantage of a more lightweight network.

User preference: To complement our quantitative and qualitative evaluations, we conducted a user study to assess participants' subjective preferences regarding reconstruction results based on the CAPE dataset. In this study, participants were presented

with reconstruction results from two methods: our proposed approach and a comparative method. They were instructed to select the reconstruction they perceived as superior in quality, considering aspects such as pose accuracy, clothing details, and the precision of facial and hand features. Data were collected from 43 participants, each evaluating 10 pairs of reconstructed human models for each comparison. As shown in Figure 7, our method was consistently preferred over the comparative method, demonstrating its superiority.

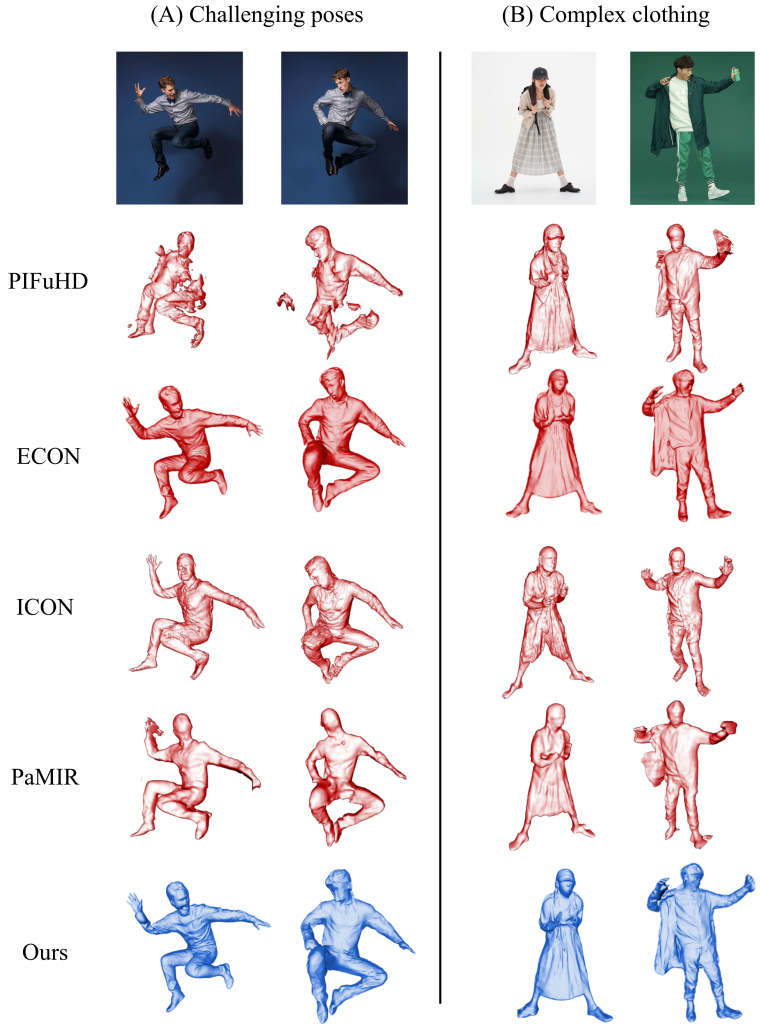


Figure 6. Existing methods exhibit varying limitations in 3D human reconstruction: PIFuHD [13] demonstrates competent clothing reconstruction capabilities but encounters challenges with complex pose estimation. While ICON [39] achieves reasonable overall performance, its ability to reconstruct intricate clothing details remains constrained. Similarly, PaMIR [14] shows insufficient reconstruction fidelity, particularly in capturing fine-grained surface details. In contrast, both ECON [19] and our proposed method demonstrate superior reconstruction quality, effectively addressing these limitations through advanced architectural designs.

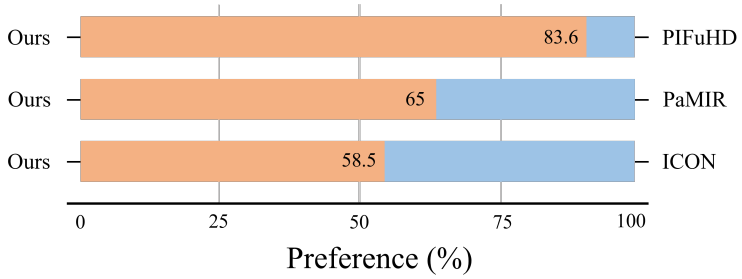


Figure 7. User preference. Our method demonstrated higher user preference compared with the baseline approach. The user study results further validate our approach's effectiveness in meeting the demand for high-quality human body reconstruction. In the figure, the orange color represents our method, while the blue color indicates the comparative method.

4.4. Ablation Study

To further validate the effectiveness of the proposed model, we conducted a series of ablation studies to evaluate the importance and contribution of different modules and designs in our method. Specifically, we systematically removed or modified certain key modules in the network and analyzed the impact of these changes on model performance. The experiments were conducted on the CAPE dataset [55], which features diverse pose variations, making it suitable for testing the model's performance under complex postural conditions. By comparing the experimental results under different configurations (as shown in Table 6), we analyzed the specific contributions of each module to the quality of generated results, runtime efficiency, and overall model performance.

Benefits of the Dilated Convolutions Block: To evaluate the impact of Dilated Convolutions, we conducted an ablation study by setting the dilation rate of all Dilated Convolutions in the Dilated Convolutions Block to 1, turning off the dilation mechanism. Experimental results showed a decrease in accuracy, indicating that incorporating the Dilated Convolutions Block expands the receptive field, addressing the limitations of shallow networks with constrained receptive fields, as shown in Table 6.

Benefits of the Cross-Covariance Attention Block: To assess the role of the Cross-Covariance Attention Block, we performed experiments by removing all instances of this block. Results demonstrated a decline in accuracy, confirming that the Cross-Covariance Attention Block is essential for extracting cross-channel feature information. This mechanism compensates for the limitations of CNNs, which extract local features.

Benefits of VGG Loss: The ablation studies conducted on our network architecture revealed a noticeable reduction in reconstruction accuracy when VGG Loss was excluded. It underscores the critical role of VGG Loss in enhancing the performance of normal map generation. By capturing high-level semantic information and improving perceptual quality, VGG Loss ensures that the generated normal maps align closely with human visual perception. Its ability to emphasize finer details and contextual relationships significantly contributes to the overall quality and fidelity of the reconstructions, making it an indispensable component of our loss function design.

Benefits of Huber Loss: The experimental results demonstrated a decline in accuracy when Huber Loss was omitted from the training process, indicating its crucial contribution to the network's performance. Huber Loss excels in balancing the sensitivity to outliers while maintaining stability during optimization. By penalizing large deviations less aggressively than traditional L1 or L2 losses, it enables the network to extract spatial information effectively. This loss function leads to improved robustness and precision in handling variations in geometric structures, particularly in areas with subtle transitions or intricate details. Its inclusion ensures the generation of accurate and realistic surface normal maps, further enhancing the overall reconstruction quality.

Table 6. Ablation study on CAPE dataset.

Methods	Chamfer	P2S	Normals
w/o Cross-Covariance Attention Block	1.074	0.994	0.0535
w/o Dilated Convolutions Block	1.139	1.087	0.0569
w/o VGG Loss	1.560	1.607	0.0662
w/o Huber Loss	1.917	1.973	0.0546

4.5. Facial and Hand Refinement Attempts

We further explored enhancing the facial and hand details of the reconstructed characters by replacing their original features with those generated using the SMPL-X face model. This approach was designed to achieve greater precision and refinement in facial features and hand articulation. The results, illustrated in Figure 8, reveal that this replacement significantly enhances the overall realism and intricacy of facial expressions and hand details. However, it also introduces a notable trade-off: the characters' unique and distinctive traits are diminished, leading to a loss of individuality. This finding highlights the challenge of balancing detail and realism with preserving personalized attributes in human reconstruction tasks.

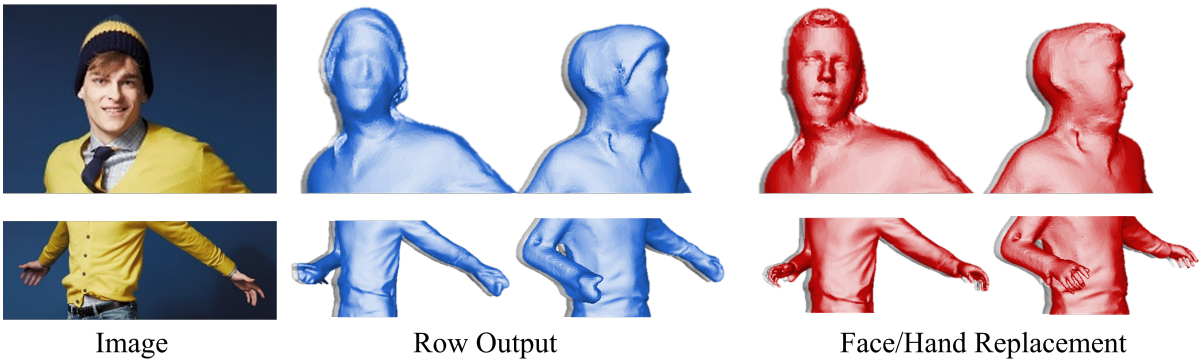


Figure 8. To enhance the reconstruction quality, the facial and hand regions of the generated model are replaced with the corresponding components from the SMPL-X model.

4.6. Text-Driven Texture Generation

We draw inspiration from the texture method proposed by TEXTure [58] to generate surface textures for human models. Utilizing a pre-trained depth-to-image diffusion model [59] guided by textual prompts, we render the 3D model from various viewpoints. Subsequently, we employ a trimap partitioning of the rendered images into three progression states. This process leverages the trimap representation to create seamless textures from different views, ultimately completing the mapping for the entire model.

By providing different textual descriptions, such as "Napoleon", "a 30-year-old man wearing a red shirt, blue pants, and brown shoes", or "an 18-year-old boy wearing a red jacket, red pants, and black shoes", we generated various results, as shown in the Figure 9, and obtained outstanding experimental results.



Figure 9. This figure demonstrates the generated texture maps from various viewpoints using different text prompts.

4.7. Animatable Avatar Creation

We further evaluated the animatability of the 3D human models generated by our method. Specifically, we reconstructed a series of images depicting the same individual in various poses and input the resulting 3D models and their corresponding SMPL models into SCANimate [60]. By leveraging our generated results, SCANimate effectively learns and represents the deformations of human models under different poses. This process significantly facilitates the construction of animatable avatars by providing high-quality input data, enabling better capture of complex geometric variations in human bodies. As shown in Figure 10, the experimental results demonstrate that our method contributes to generating high-quality animatable avatars, highlighting the effectiveness of our reconstructed outputs.

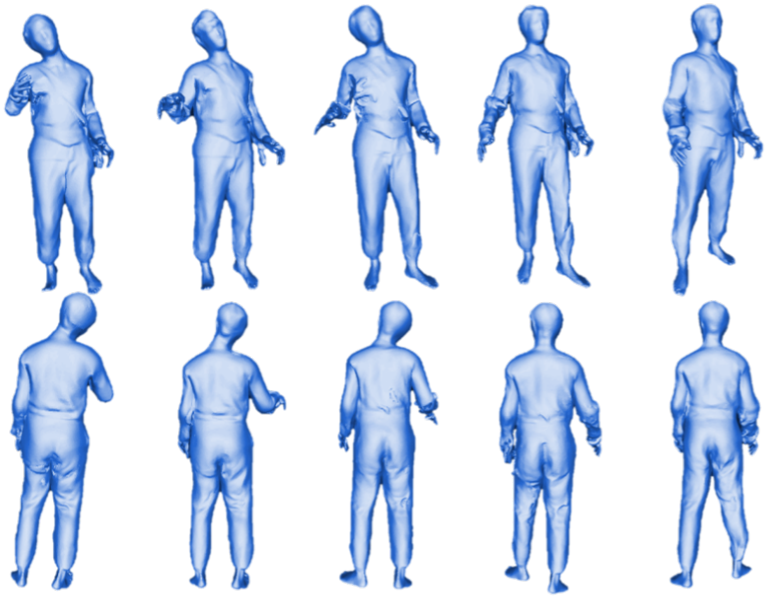


Figure 10. We developed a pose-parameter-driven avatar using generated human models based on the SCANimate [60] framework.

5. Conclusions

This paper introduces an innovative framework for 3D human reconstruction, achieving significant advancements in accuracy, efficiency, and realism. Our approach leverages SMPL-X as a prior to generating surface normal maps, which are subsequently integrated to reconstruct high-quality 3D human models. We introduce a custom loss function to refine the quality of the models, while SMPL-X priors are utilized to recover missing details. Additionally, we employ a diffusion model to generate surface textures and utilize SCANimate to produce an animatable avatar. To further enhance efficiency, we propose a lightweight normal map generation network incorporating Dilated Convolutions and Cross-Covariance Attention mechanisms, significantly reducing computational complexity without compromising performance. Experimental results demonstrate that our method achieves superior reconstruction quality while optimizing network size and parameters. However, challenges remain in handling complex poses and clothing, where arteficts such as missing limbs or inconsistent back-side reconstructions may occur. Future work will focus on improving robustness to extreme poses, refining clothing reconstruction, and enhancing overall consistency, with the aim of advancing monocular 3D human reconstruction for broader real-world applications.

Author Contributions: Conceptualization, J.L. (Jiaxuan Liu), J.W. and H.Y.; methodology, J.L. (Jiaxuan Liu) and J.W.; software, J.L. (Jiaxuan Liu); validation, J.L. (Jiaxuan Liu); formal analysis, J.L. (Jiaxuan Liu); investigation, J.L. (Jiaxuan Liu) and R.J.; resources, L.S.; data curation, J.L. (Jiaxuan Liu), J.W. and J.L. (Jing Liu); writing, original draft preparation, J.L. (Jiaxuan Liu); writing—review and editing, J.L. (Jiaxuan Liu), L.S., J.L. (Jing Liu), J.W. and H.Y.; visualization, J.L. (Jiaxuan Liu) and R.J.; supervision, L.S.; project administration, L.S.; funding acquisition, L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China, Project No. 2024YFE0200700, Subject No. 2024YFE0200703. This work was also supported in part by the Specific Research Fund of the Innovation Platform for Academicians of Hainan Province under Grant YSPTZX202314, in part by the Shanghai Key Research Laboratory of NSAI and China FAW Joint Development Project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are included in the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rymov, D.A.; Svistunov, A.S.; Starikov, R.S.; Shifrina, A.V.; Rodin, V.G.; Evtikhiev, N.N.; Cheremkhin, P.A. 3D-CGH-Net: Customizable 3D-hologram generation via deep learning. *Opt. Lasers Eng.* **2025**, *184*, 108645. [CrossRef]
2. Wu, Z.; Wang, H.; Che, F.; Chen, X.L.Z.; Zhang, Q. Dynamic 3D shape reconstruction under complex reflection and transmission conditions using multi-scale parallel single-pixel imaging. *Light Adv. Manuf.* **2024**, *5*, 373. [CrossRef]
3. Puliti, C.S. Structure from Motion Photogrammetry in Forestry: A Review. *Curr. For. Rep.* **2019**, *5*, 155–168.
4. Rossol, N.; Cheng, I.; Basu, A. A Multisensor Technique for Gesture Recognition Through Intelligent Skeletal Pose Analysis. *IEEE Trans. Hum.-Mach. Syst.* **2016**, *46*, 350–359. [CrossRef]
5. Chua, J.; Ong, L.Y.; Leow, M.C. Telehealth Using PoseNet-Based System for In-Home Rehabilitation. *Future Internet* **2021**, *13*, 173. [CrossRef]
6. Dai, Q.; Ma, C.; Zhang, Q. Advanced Hyperspectral Image Analysis: Superpixelwise Multiscale Adaptive T-HOSVD for 3D Feature Extraction. *Sensors* **2024**, *24*, 4072. [CrossRef]
7. Sharma, P.; Shah, B.B.; Prakash, C. A Pilot Study on Human Pose Estimation for Sports Analysis. In *Pattern Recognition and Data Analysis with Applications*; Gupta, D., Goswami, R.S., Banerjee, S., Tanveer, M., Pachori, R.B., Eds.; Springer: Singapore, 2022; pp. 533–544.

8. Rabosh, E.V.; Balbekin, N.S.; Petrov, N.V. Analog-to-digital conversion of information archived in display holograms: I. discussion. *J. Opt. Soc. Am. A* **2023**, *40*, B47–B56. [CrossRef] [PubMed]
9. Wang, L.; Liu, Z.; Gu, X.; Wang, D. Three-Dimensional Reconstruction of Road Structural Defects Using GPR Investigation and Back-Projection Algorithm. *Sensors* **2025**, *25*, 162. [CrossRef]
10. Liu, H.; Hellín, C.J.; Tayebi, A.; Calles, F.; Gómez, J. Vertex-Oriented Method for Polyhedral Reconstruction of 3D Buildings Using OpenStreetMap. *Sensors* **2024**, *24*, 7992. [CrossRef]
11. Pan, H.; Cai, Y.; Yang, J.; Niu, S.; Gao, Q.; Wang, X. HandFI: Multilevel Interacting Hand Reconstruction Based on Multilevel Feature Fusion in RGB Images. *Sensors* **2025**, *25*, 88. [CrossRef] [PubMed]
12. Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Li, H.; Kanazawa, A. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2304–2314. [CrossRef]
13. Saito, S.; Simon, T.; Saragih, J.M.; Joo, H. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; pp. 81–90. [CrossRef]
14. Zheng, Z.; Yu, T.; Liu, Y.; Dai, Q. PaMIR: Parametric Model-Conditioned Implicit Representation for Image-Based Human Reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3170–3184. [CrossRef]
15. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; SMPL, M.B. SMPL: A skinned multi-person linear model. *ACM Trans. Graph. (TOG)* **2015**, *34*, 1–16. . [CrossRef]
16. Kovács, L.; Bódis, B.M.; Benedek, C. LidPose: Real-Time 3D Human Pose Estimation in Sparse Lidar Point Clouds with Non-Repetitive Circular Scanning Pattern. *Sensors* **2024**, *24*, 3427. [CrossRef]
17. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2016**, arXiv:1511.07122.
18. Ali, A.; Touvron, H.; Caron, M.; Bojanowski, P.; Douze, M.; Joulin, A.; Laptev, I.; Neverova, N.; Synnaeve, G.; Verbeek, J.; et al. XCiT: Cross-Covariance Image Transformers. *arXiv* **2021**, arXiv:2106.09681.
19. Xiu, Y.; Yang, J.; Cao, X.; Tzionas, D.; Black, M.J. ECON: Explicit Clothed humans Optimized via Normal integration. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 512–523. [CrossRef]
20. Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A.A.; Tzionas, D.; Black, M.J. Expressive body capture: 3d hands, face, and body from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10975–10985.
21. Johnson, J.; Alahi, A.; Li, F.-F. *Perceptual Losses for Real-Time Style Transfer and Super-Resolution*; Springer: Cham, Switzerland, 2016.
22. Huber, P.J. Robust Estimation of a Location Parameter. *Ann. Math. Stat.* **1964**, *35*, 73–101. [CrossRef]
23. Anguelov, D.; Srinivasan, P.; Koller, D.; Thrun, S.; Rodgers, J.; Davis, J. SCAPE: Shape completion and animation of people. *ACM Trans. Graph. (TOG)* **2005**, *24*, 408–416. [CrossRef]
24. Hasler, N.; Stoll, C.; Sunkel, M.; Rosenhahn, B.; Seidel, H. A Statistical Model of Human Pose and Body Shape. *Comput. Graph. Forum* **2009**, *28*, 337–346. [CrossRef]
25. James, D.L.; Twigg, C.D. Skinning mesh animations. *ACM Trans. Graph.* **2005**, *24*, 399–407. [CrossRef]
26. Allen, B.; Curless, B.; Popović, Z. Articulated body deformation from range scan data. *ACM Trans. Graph. (TOG)* **2002**, *21*, 612–619. [CrossRef]
27. Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.V.; Romero, J.; Black, M.J. Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In Proceedings of the Computer Vision-ECCV 2016-14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Volume 9909, pp. 561–578. [CrossRef]
28. Kanazawa, A.; Black, M.J.; Jacobs, D.W.; Malik, J. End-to-end recovery of human shape and pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7122–7131.
29. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
30. Kolotouros, N.; Pavlakos, G.; Black, M.J.; Daniilidis, K. Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop. In Proceedings of the ICCV, Seoul, Republic of Korea, 27 October–2 November 2019.
31. He, T.; Collomosse, J.; Jin, H.; Soatto, S. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9276–9287.
32. Li, Z.; Yu, T.; Pan, C.; Zheng, Z.; Liu, Y. Robust 3d self-portraits in seconds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1344–1353.
33. Dong, Z.; Guo, C.; Song, J.; Chen, X.; Geiger, A.; Hilliges, O. Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20470–20480.

34. Yang, Z.; Wang, S.; Manivasagam, S.; Huang, Z.; Ma, W.; Yan, X.; Yumer, E.; Urtasun, R. S3: Neural Shape, Skeleton, and Skinning Fields for 3D Human Modeling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021; pp. 13284–13293. [CrossRef]
35. Zheng, Y.; Shao, R.; Zhang, Y.; Yu, T.; Zheng, Z.; Dai, Q.; Liu, Y. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 6239–6249.
36. Huang, Z.; Xu, Y.; Lassner, C.; Li, H.; Tung, T. ARCH: Animatable Reconstruction of Clothed Humans. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; pp. 3090–3099. [CrossRef]
37. He, T.; Xu, Y.; Saito, S.; Soatto, S.; Tung, T. ARCH++: Animation-Ready Clothed Human Reconstruction Revisited. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 11–17 October 2021; pp. 11026–11036. [CrossRef]
38. Liao, T.; Zhang, X.; Xiu, Y.; Yi, H.; Liu, X.; Qi, G.J.; Zhang, Y.; Wang, X.; Zhu, X.; Lei, Z. High-fidelity clothed avatar reconstruction from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 8662–8672.
39. Xiu, Y.; Yang, J.; Tzionas, D.; Black, M.J. ICON: Implicit Clothed humans Obtained from Normals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 13286–13296. [CrossRef]
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
41. Zhou, Z.; Fan, X.; Shi, P.; Xin, Y. R-MSFM: Recurrent Multi-Scale Feature Modulation for Monocular Depth Estimating. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 10–17 October 2021; pp. 12757–12766. [CrossRef]
42. Zhang, N.; Nex, F.; Vosselman, G.; Kerle, N. Lite-Mono: A Lightweight CNN and Transformer Architecture for Self-Supervised Monocular Depth Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, 17–24 June 2023; pp. 18537–18546. [CrossRef]
43. Bae, J.; Moon, S.; Im, S. Deep Digging into the Generalization of Self-Supervised Monocular Depth Estimation. In Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, 7–14 February 2023; pp. 187–196. [CrossRef]
44. Deng, B.; Lewis, J.P.; Jeruzalski, T.; Pons-Moll, G.; Hinton, G.; Norouzi, M.; Tagliasacchi, A. Nasa neural articulated shape approximation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 612–628.
45. Feng, Y.; Choutas, V.; Bolkart, T.; Tzionas, D.; Black, M.J. Collaborative regression of expressive bodies using moderation. In Proceedings of the 2021 International Conference on 3D Vision (3DV), London, UK, 1–3 December 2021; pp. 792–804.
46. Rembg: A Tool to Remove Images Background. 2022. Available online: https://github.com/danielgatis/rembg (accessed on 26 January 2024).
47. Kazhdan, M.; Bolitho, M.; Hoppe, H. Poisson surface reconstruction. In Proceedings of the Fourth Eurographics Symposium on Geometry Processing, Cagliari, Italy, 26–28 June 2006; Volume 7.
48. Godard, C.; Aodha, O.M.; Firman, M.; Brostow, G.J. Digging Into Self-Supervised Monocular Depth Estimation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3827–3837. [CrossRef]
49. Zhou, H.; Greenwood, D.; Taylor, S. Self-Supervised Monocular Depth Estimation with Internal Feature Fusion. *arXiv* **2021**, arXiv:2110.09482.
50. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
51. Hendrycks, D.; Gimpel, K.; Gimpel, K. Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units. *arXiv* **2016**, arXiv:abs/1606.08415. Available online: https://api.semanticscholar.org/CorpusID:2359786 (accessed on 20 February 2025).
52. Ba, L.J.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450.
53. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
54. Yu, T.; Zheng, Z.; Guo, K.; Liu, P.; Dai, Q.; Liu, Y. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5746–5756.

55. Ma, Q.; Yang, J.; Ranjan, A.; Pujades, S.; Pons-Moll, G.; Tang, S.; Black, M.J. Learning to dress 3d people in generative clothing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6469–6478.
56. RenderPeople. 2018. Available online: https://renderpeople.com (accessed on 20 October 2024).
57. Charles, R.Q.; Su, H.; Kaichun, M.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 77–85. [CrossRef]
58. Richardson, E.; Metzer, G.; Alaluf, Y.; Giryes, R.; Cohen-Or, D. TEXTure: Text-Guided Texturing of 3D Shapes. In Proceedings of the ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023, Los Angeles, CA, USA, 6–10 August 2023; pp. 54:1–54:11. [CrossRef]
59. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
60. Saito, S.; Yang, J.; Ma, Q.; Black, M.J. SCANimate: Weakly Supervised Learning of Skinned Clothed Avatar Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021; pp. 2886–2897. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.