HumanPro: Single-view 3D Clothed Human Reconstruction with Progressive Normal Guidance

Jianchi Sun[1], Fei Luo[1], Wenzhuo Fan[1], Yu Jiang[1], Chunxia Xiao[1*],

[1]School of Computer Science, Wuhan University, Wuhan, China
sunjc0306@whu.edu.cn, luofei@whu.edu.cn, fwz0818@whu.edu.cn, jiangyu1181@whu.edu.cn, cxxiao@whu.edu.cn

Abstract

Reconstructing fine-grained geometry of clothed human from single-view image is a challenging task, particularly in accurately recovering complex shapes and generating clothes details. To address these limitations, we propose a novel approach named HumanPro, which estimates high-quality human normals via a generative model, and progressively deforms a parametric body into the final clothed human mesh guided by these normals. First, we propose a geometry-aware latent diffusion model with a normal enhancer to estimate high-quality human normals from four views. Then, we propose a progressive mesh optimization consisting of shape-aware deformation alignment and global-to-patch detail refinement for human mesh reconstruction. The shape-aware deformation alignment applies image morphing to learn the shape-level gap of normals, addressing large-scale deformation of complex clothes. It can recover the overall silhouette of a clothed human, and serves as an initialization for the global-to-patch detail refinement. Our detail refinement combines global and patch-wise optimization strategies to iteratively address fine-scale deformations by minimizing the pixel-level difference of normals. This way effectively recovers fine-grained details while avoiding local minima. Extensive experiments demonstrate that HumanPro can deal with various challenging scenarios and outperforms state-of-the-art methods.

Introduction

Reconstructing human mesh from single-view image is one of the fundamental topics in computer graphics. The reconstruction results have wide applications, including virtual-real fusion (Bao et al. 2025; Zheng et al. 2025b; Shen et al. 2025), telepresence (Zhao et al. 2025), film production (Zheng et al. 2025a), etc. Compared to traditional methods that rely on bulky and expensive devices, single-view 3D reconstruction can significantly reduce time consumption. However, since the depth value is inherently ambiguous in a single-view 2D image, it is an ill-posed problem to reconstruct a detailed 3D human mesh from a single-view image.

Current methods usually reconstruct clothed humans by either learning implicit fields (He et al. 2020; Alldieck, Zanfir, and Sminchisescu 2022; Corona et al. 2023; Chan et al. 2022; Huang et al. 2020; Song et al. 2023) or deforming explicit shapes (Alldieck et al. 2019; Zhu et al. 2019, 2021; Liu et al. 2024; Kim et al. 2023; Li et al. 2025; Zhang et al. 2025). Although they are effective in some cases, these approaches struggle with fine-grained reconstruction of clothed human. Implicit-field-based methods rely on features from the parametric body and normals to map 2D images to 3D fields, but the lack of clothes cues from the naked body results in incomplete geometries. Explicit-shape-based methods deform the parametric body mesh directly, but the significant disparity between the naked body and the clothed human increases computational burden and compromises reconstruction quality. These limitations make a request for a gradual refinement process to bridge this gap and improve reconstruction quality.

Figure 1: Meshes reconstructed by HumanPro. HumanPro applies human normals as guidance to progressively deform an explicit parametric body mesh, achieving complete and fine-grained human reconstruction.

To address these problems, we propose HumanPro, a single-view 3D human reconstruction method with progressive normal guidance. HumanPro first estimates accurate human normals via a generative model, and then progressively deforms a parametric body into the final clothed hu-

*Chunxia Xiao is the corresponding author.

man mesh. Specifically, HumanPro consists of human normal estimation and human mesh reconstruction. In the human normal estimation, we propose a geometry-aware latent diffusion model with a normal enhancer to estimate human normals. The latent diffusion model also conditions on the image and the parametric body. Unlike previous methods (Li et al. 2025; Kim et al. 2023; He et al. 2025), we directly encode point-based high-frequency signals extracted from the parametric body to guide the denoising. Additionally, to further preserve fine-grained details, we introduce a normal enhancer to refine and upsample the denoised results from the diffusion model. In this way, we generate high-quality and high-resolution human normals from four views, facilitating the detailed human reconstruction.

In the human mesh reconstruction, we propose a progressive optimization framework that addresses large-scale and fine-scale deformations. The large-scale deformation originates from the geometric disparity between the naked parametric body and the target clothed human. We capture this disparity through the shape-level gap between the estimated human normals and the rendered normals from the parametric body. To bridge this gap, we introduce a shape-aware deformation alignment that learns warp maps via image morphing techniques (Liu et al. 2024; Truong et al. 2021). These learned maps enable rapid deformation of the parametric body into a rough shape, ensuring the deformed vertices cover the entire surface silhouette. Building upon this initialization, we further deal with fine-scale deformations through a global-to-patch detail refinement. This refinement iteratively optimizes the mesh by minimizing pixel-level normal differences. It first performs global mesh adjustment to capture holistic structure, then applies patch-wise optimization to recover local details. This progressive strategy not only effectively generates fine-grained details but also mitigates local minima issues.

The contributions can be summarized as follows:

- Propose a geometry-aware latent diffusion model with a normal enhancer for human normal estimation. This module can generate high-quality human normals from four views guided by the high-frequency signals of the parametric body, facilitating detailed clothed human reconstruction.
- Propose a progressive normal guided approach for human mesh reconstruction. This approach applies shape-aware deformation alignment and global-to-patch detail refinement to deform the mesh in a coarse-to-fine manner.

Extensive experiments on three benchmarks and in-the-wild images validate that the proposed method outperforms SOTA methods. Our method can deal with various challenging scenarios and produce a complete human with fine-grained details.

Related Work

Implicit-field-based Human Reconstruction

Implicit-field-based methods reconstruct human mesh by predicting field representations such as voxel (Varol et al. 2018), unsigned/signed distance field (SDF) (Park et al. 2019; Luo et al. 2024), and Fourier occupancy field (FOF) (Feng et al. 2022; Li, Luo, and Xiao 2024). Early methods (Zheng et al. 2019; Varol et al. 2018) regressed 3D voxel from single-view image to reconstruct mesh. They require huge GPU memory and are limited in resolution. Later, PIFu (Saito et al. 2019) employed an implicit function to predict the SDF value for each point without resolution limitation. Based on PIFu, a series of studies (Zheng et al. 2021; Xiu et al. 2022; Ho et al. 2024; Yang et al. 2023; He et al. 2020) employed different priors as implicit guidance for improving reconstruction, such as normal, depth, and parametric body. Further advancements like GTA (Zhang et al. 2024b), SiFU (Zhang, Yang, and Yang 2024), and HiLo (Yang et al. 2024) explored more complex networks to encode image features for inference. Due to the free-form nature of implicit functions, they sometimes generate non-human shapes. Unlike implicit functions, FOF representation (Zhang et al. 2024a; Feng et al. 2022) regresses the complete shape of a human from a global perspective. However, their results always suffer from thickness errors in side views.

Explicit-shape-based Human Reconstruction

Explicit-shape-based methods employ point clouds or mesh to represent humans, allowing them to regularize 3D shapes with human topology explicitly. Some methods (Gabeur et al. 2019; Xiu et al. 2023) estimated multi-view depth maps, and then projected them into 3D point clouds. The 3D human mesh can be obtained from point clouds by Poisson reconstruction (Kazhdan and Hoppe 2013). Since depth images only provide partial and incomplete 3D point clouds without body topology, they cannot reconstruct a full human with a consistent surface. Other works (Kanazawa et al. 2018; Zhang et al. 2023) used SMPL (Loper et al. 2015) and SMPL-X (Bogo et al. 2016) to represent 3D humans. These works recovered 3D mesh from single-view image by regressing the parameters of the parametric body. Limited by the representation ability of the parametric body, they can not recover complex clothes.

To reconstruct surface details, some studies applied vertex deformation based on parametric body. Zhu et al. (Zhu et al. 2019, 2021) employed a network to estimate vertex displacements of the parametric body hierarchically. Alldieck et al. (Alldieck et al. 2019) attempted to adjust vertex displacements in UV space for clothed human reconstruction. Liu et al. (Liu et al. 2024) proposed a stretch-refine strategy for human reconstruction using shift fields and a graph convolutional network. However, these methods suffer from poor generalizability on in-the-wild images and produce over-smoothed outputs.

Inspired by multi-view diffusion-based 3D generation (Wu et al. 2024; Long et al. 2024), some works (Kim et al. 2023; Li et al. 2025) directly perform a mesh optimization initialized by a SMPL-X mesh for 3D clothed human reconstruction. Due to the gap between naked SMPL-X meshes and clothed humans, this vanilla initialization increases inference burden and reduces reconstruction quality, especially for loose clothes. On the contrary, our progressive normal guidance employs different deformations to effectively

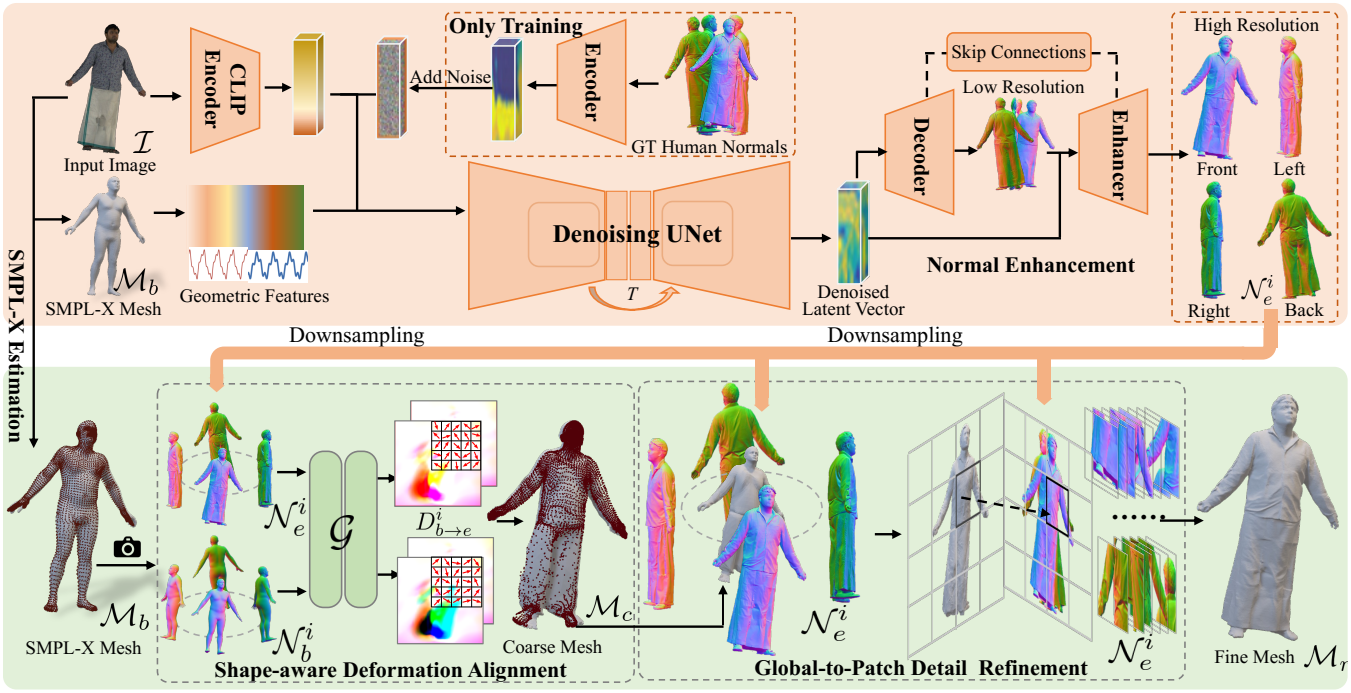

Figure 2: Approach overview. HumanPro takes a human image $\mathcal{I}$ and the corresponding estimated SMPL-X mesh $\mathcal{M}_b$ as input to reconstruct a high-fidelity clothed human mesh $\mathcal{M}_r$. HumanPro involves two main steps: (***Top***) Human Normal Estimation. The human normals $\{\mathcal{N}_e^i\}_{i=0}^3$ from four views are estimated from the image $\mathcal{I}$ and the SMPL-X mesh $\mathcal{M}_b$ using the proposed geometry-aware latent diffusion model with a normal enhancer. (***Bottom***) Human Mesh Reconstruction. $\mathcal{M}_r$ is deformed from the SMPL-X mesh $\mathcal{M}_b$ with the estimated human normals $\{\mathcal{N}_e^i\}_{i=0}^3$ via progressive normal guidance.

handle the gap between naked and clothed humans, accelerating convergence and improving reconstruction quality.

Methodology

Overview

We aim to reconstruct a complete human mesh $\mathcal{M}_r$ with fine-grained geometries from a single-view image $\mathcal{I}$. We first employ PyMAF-X (Zhang et al. 2023) to estimate the SMPL-X mesh $\mathcal{M}_b$ from the single-view image $\mathcal{I}$ in an off-the-shelf manner. Then, HumanPro deforms the SMPL-X mesh $\mathcal{M}_b$ into a complete and fine-grained clothed human $\mathcal{M}_r$ guided by normals. The overview of HumanPro is illustrated in Fig. 2. HumanPro contains two important parts: (1) ***Human Normal Estimation.*** We inject $\mathcal{M}_b$ and $\mathcal{I}$ as conditions into the diffusion model, which iteratively performs denoising to ultimately obtain the denoised results from four views (front, left, back, and right). These denoised results are then refined by a normal enhancer to produce high-quality normals $\{\mathcal{N}_e^i\}_{i=0}^3$. (2) ***Human Mesh Reconstruction.*** Guided by human normals $\{\mathcal{N}_e^i\}_{i=0}^3$, we reconstruct a final 3D clothed human $\mathcal{M}_r$ via progressively deforming the SMPL-X mesh $\mathcal{M}_b$. This progressive optimization consists of a shape-aware deformation alignment and a global-to-local detail refinement. They leverage the shape-level and pixel-level differences between the estimated human normals and the body normals rendered from $\mathcal{M}_b$, enabling large-scale deformations for human shape recovery and fine-scale deformations for clothing detail generation, respectively.

Human Normal Estimation

We propose a geometry-aware latent diffusion model with a normal enhancer, which takes the image $\mathcal{I}$ and the SMPL-X mesh $\mathcal{M}_b$ as conditions to estimate human normals from four views. Its key is to directly encode the high-frequency signals extracted from $\mathcal{M}_b$ to guide denoising, enabling cross-view consistency and normal estimation. Additionally, we further refine the initial denoised results using the normal enhancer to produce high-resolution normals.

Geometry-aware Latent Diffusion Model. We fine-tune a pre-trained latent diffusion model UNet (Rombach et al. 2022; Kim et al. 2023) as our denoising network, denoted as $\boldsymbol{\epsilon}_\theta$. To model the distribution of normals, following (Rombach et al. 2022), we first train a normal VAE $(\mathcal{E}_l, \mathcal{D}_l)$ to encode human normals into a latent space. Given the ground-truth human normal $\mathcal{N}_l^i \in \mathbb{R}^{512\times512\times4}$ with an alpha channel, the encoder $\mathcal{E}_l$ can map $\mathcal{N}_l^i$ into latent vectors $z^i \in \mathbb{R}^{128\times128\times4}$. The corresponding decoder $\mathcal{D}_l$ decodes z^i back to the normal space. Since the normals from four views are estimated simultaneously, the final latent vector $z \in \mathbb{R}^{128\times128\times16}$ is formed by concatenating the individual latent vector z^i from each view.

In the latent diffusion, according to a specified timestep t, the forward diffusion adds a Gaussian noise $\boldsymbol{\epsilon} \sim \mathbb{N}(\mathbf{0}, \mathbf{I})$

to the latent vectors z, and obtains z_t. Our denoising network $\boldsymbol{\epsilon_\theta}$ learns the injected noise during the reverse diffusion, guided by the input image $\mathcal{I}$ and the SMPL-X mesh $\mathcal{M}_b$. The SMPL-X mesh $\mathcal{M}_b$ can provide robust pose priors for diffusion, enhancing the completeness of human normals. The input image $\mathcal{I}$ serves as a correct identity, ensuring the appearance consistency between the estimated human normals and the input image. To incorporate the 3D prior of the SMPL-X mesh, we directly inject point-based geometric features extracted from $\mathcal{M}_b$ into the latent denoising process. Specifically, we treat all vertices $\mathbf{V}_b$ and normals $\mathbf{N}_b$ of the SMPL-X mesh $\mathcal{M}_b$ as oriented point clouds. We apply the position embedding operation γ to extract high-frequency normal signals of these points:

$$\gamma(p) = \{\sin(2^l \pi p), \cos(2^l \pi p) | l = 0, 1, ...L-1\}, \quad (1)$$

where $L = 32$ is the dim of the position embedding. Finally, we merge the high-frequency signals and vertices to obtain $\mathcal{H} = \gamma(\mathbf{N}_b) \cup \mathbf{V}_b$ as geometric features for denoising. Following (Ho, Jain, and Abbeel 2020), the loss function $\mathcal{L}_{ldm}$ of the latent diffusion model is :

$$\mathcal{L}_{ldm} = \mathbb{E}_{z,\mathcal{I},\mathcal{H},\boldsymbol{\epsilon}\sim\mathbb{N}(\mathbf{0},\mathbf{I}),t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(z_t, t, \mathcal{I}, \mathcal{H}\right)\|_2^2\right]. \quad (2)$$

Normal Enhancer. For efficient modeling of data distribution, we only retain low-resolution normals in the latent diffusion model. However, some discarded high-resolution details are crucial for capturing fine-grained human geometries. To address this, we develop a normal enhancer to estimate the high-resolution normal $\mathcal{N}_e^i$. Specifically, we design a reference-guided decoding network $\mathcal{D}_h$ that leverages the latent vector z_0 and intermediate outputs from the decoder $\mathcal{D}_l$ to predict the high-resolution normal:

$$\mathcal{N}_e^i = \mathcal{D}_h\left(\mathcal{D}_l(z_0^i), z_0^i\right), i = 0, 1, 2, 3, \quad (3)$$

where z_0^i is the denoised latent vector of the i-th view. To enhance contextual perception from the latent space, we incorporate skip connections across layers in $\mathcal{D}_h$ to fuse corresponding feature maps from $\mathcal{D}_l(z_0^i)$. During training, we supervise the predicted high-resolution normal $\mathcal{N}_e^i$ using an L1 loss and a VGG loss with respect to the ground-truth high-resolution normal $\mathcal{N}_h^i \in \mathbb{R}^{1024\times1024\times4}$:

$$\mathcal{L}_h = \lambda_{\text{L1}}||\mathcal{N}_e^i - \mathcal{N}_h^i||_1 + \lambda_{\text{VGG}}\mathcal{L}_{\text{VGG}}(\mathcal{N}_e^i, \mathcal{N}_h^i), \quad (4)$$

where $\mathcal{L}_{\text{VGG}}(\cdot,\cdot)$ denotes the perceptual loss between two normals, λ_{L1} and λ_{VGG} are the balanced weights.

In the inference stage, we start from the time T and give the Gaussian noise $z_T \sim \mathbb{N}(\mathbf{0}, \mathbf{I})$, then iteratively denoise from the previous step until z_0. We chunk z_0 into four latent vectors $\{z_0^i\}_{i=0}^3$. Finally, we take z_0^i and the intermediate result decoded by $\mathcal{D}_l$ into the normal enhancer $\mathcal{D}_h$, generating the high-resolution human normal $\mathcal{N}_e^i \in \mathbb{R}^{1024\times1024\times4}$.

Human Mesh Reconstruction

After obtaining the human normals $\{\mathcal{N}_e^i\}_{i=0}^3$, we apply mesh deformation to reconstruct clothed human mesh. A straightforward and common strategy is to employ the SMPL-X mesh as an initial mesh. However, a major limitation of this way lies in the significant discrepancy between the minimally naked SMPL-X mesh and the clothed human mesh. This discrepancy introduces challenges in deformation convergence, often leading to undesirable detail recovery or degraded robustness. To address this problem, we propose a progressive optimization strategy guided by human normals, which consists of a shape-aware deformation alignment and a global-to-local detail refinement.

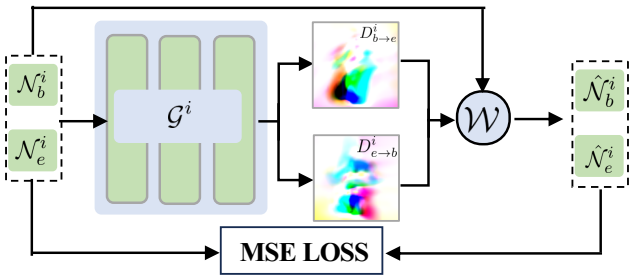

Figure 3: Self-supervised mechanism of the shape-aware deformation alignment. The concatenation of $\mathcal{N}_b^i$ and $\mathcal{N}_e^i$ is fed into the network $\mathcal{G}^i$ to output the warp maps $D_{b\to e}^i$ and $D_{e\to b}^i$. The image morphing $\mathcal{W}$ can transform $\mathcal{N}_b^i$, and $\mathcal{N}_e^i$ into the corresponding maps $\hat{\mathcal{N}}_b^i$ and $\hat{\mathcal{N}}_e^i$ according to the warp maps.

Shape-aware Deformation Alignment. The shape-aware deformation alignment aims to address the large deformations between the naked SMPL-X mesh and the clothed human, recovering the complete shape of a human. Inspired by the advanced work (Truong et al. 2021; Liu et al. 2024), we apply image morphing to infer warp maps from body and human normals for human shape recovery. The warp map is specified by offset vectors that define the correspondences of spatial locations in the target image to those in the source image. Therefore, the image morphing can learn the shape-level gap between body normals and human normals, and the warp map can effectively indicate large deformations from the SMPL-X mesh to the clothed mesh.

To this end, we design a lightweight network $\mathcal{G} = \{\mathcal{G}^i\}_{i=0}^3$ to learn warp maps from the rendered body normals $\mathcal{N}_b^i$ to the estimated human normals $\mathcal{N}_e^i$. Lack of ground-truth labels, we adopt training-per-instance to obtain warp maps and corresponding inverse maps in a self-supervised manner (see Fig. 3). This process is summarized as follows:

$$(D_{b\to e}^i, D_{e\to b}^i) = \mathcal{G}^i(\mathcal{N}_b^i, \mathcal{N}_e^i), \quad (5)$$

where $D_{b\to e}^i$ are the warp maps from the body normal to the clothed normal, and $D_{e\to b}^i$ are the inverse maps of $D_{b\to e}^i$.

To train our network $\mathcal{G}$, we exploit the generated warp maps and inverse maps to achieve the conversion between two groups of normals by image morphing. The image morphing transform body normals into human normals using the predicted warp maps and reverse the process using the predicted inverse maps:

$$\hat{\mathcal{N}}_e^i = \mathcal{W}(\mathcal{N}_b^i, D_{b\to e}^i), \hat{\mathcal{N}}_b^i = \mathcal{W}(\mathcal{N}_e^i, D_{e\to b}^i), \quad (6)$$

where a function $\mathcal{W}(\mathcal{N}, D)$ represent a non-linear warp to the image $\mathcal{N}$, and the warp offsets are specified by D. $\hat{\mathcal{N}}_e^i$

Methods	Publications	THuman2			3DHumans			2K2K		
		CD	PSD	NC	CD	PSD	NC	CD	PSD	NC
HiLo	CVPR'2024	2.0708	2.5151	0.1006	2.6325	2.6980	0.1133	2.7850	2.6435	0.1134
SiTH	CVPR'2024	0.9707	1.0463	0.0486	1.9561	1.8222	0.0909	2.4101	2.3038	0.0847
ECON	CVPR'2023	0.9667	0.9556	0.0515	1.7327	1.5624	0.0848	2.6037	2.2905	0.0803
Chupa	ICCV'2023	1.2290	1.2027	0.0704	2.4076	2.0052	0.1114	4.9089	4.2627	0.1674
VSNet	CVPR'2024	0.9641	0.9362	0.0528	1.7382	1.4859	0.0925	2.4290	2.3554	0.0867
PSHuman	CVPR'2025	0.8641	0.8332	0.0471	1.3312	1.3812	0.0621	2.3210	2.1524	0.0823
Ours	-	**0.8323**	**0.8234**	**0.0402**	**1.1223**	**1.2249**	**0.0518**	**2.2841**	**2.1332**	**0.0711**

Table 1: Quantitative comparison with baselines on three benchmarks. Smaller numbers indicate better performance.

are the warped normals corresponding to $\mathcal{N}_e^i$, respectively. Our loss functions $\mathcal{L}_C^i$ consist of two items:

$$\mathcal{L}_C^i = \lambda_e ||\hat{\mathcal{N}_e^i} - \mathcal{N}_e^i||_2^2 + \lambda_b ||\hat{\mathcal{N}_b^i} - \mathcal{N}_b^i||_2^2, \tag{7}$$

where $\lambda_e = 0.3$, and $\lambda_b = 0.3$ are the balanced weights. In practice, since this stage primarily recovers the coarse shape of the clothed human, we employ the downsampled normals $\{\mathcal{N}_e^i\}_{i=0}^3$ with a resolution of 128×128 for optimization.

Once the lightweight network $\mathcal{G}$ is trained, we query offset vectors for vertices $\mathbf{V}_b$ of the SMPL-X mesh $\mathcal{M}_b$ on the predicted warp map $D_{b\to e}^i$ by bilinear interpolation $\mathcal{B}$. The deformed vertices $\mathbf{V}_c^i$ are written as:

$$\mathbf{V}_c^i = s\mathcal{B}(D_{b\to e}^i, \pi^i(\mathbf{V}_b)) + \mathbf{V}_b, i = 0, 1, 2, 3, \tag{8}$$

where π^i is the weak perspective camera of the i-th view, s is the ratio of image resolution and mesh space. We aggregate the multi-view deformed vertices $\mathbf{V}_c^i$ into a unified vertex set $\mathbf{V}_c$. Finally, the coarse mesh $\mathcal{M}_c$ is composed of the fused vertex set $\mathbf{V}_c$ and the faces of $\mathcal{M}_b$.

Global-to-Patch Detail Refinement. The coarse mesh $\mathcal{M}_c$ generated by the shape-aware deformation alignment has the overall silhouette and pose of a clothed human but lacks fine details. These details can be measured in the pixel-level values of normals. Hence, we infer the fine-scale deformation of $\mathcal{M}_c$ to obtain fine details by minimizing the pixel-level difference with the estimated human normals. Initialized by $\mathcal{M}_c$, we design the global-to-patch detail refinement to recover fine-grained geometry from human normals via deforming and remeshing (Palfinger 2022). Our refinement first performs a global adjustment on the coarse mesh $\mathcal{M}_c$ to enforce overall geometric consistency, and then refines local surface patches for high-quality geometric details.

Let $\mathcal{M} = \{\mathbf{V}, \mathbf{F}\}$ be an optimizing mesh with vertices $\mathbf{V}$ and faces $\mathbf{F}$. We iteratively optimize the vertices $\mathbf{V}$ by minimizing the normals $\mathcal{N}_e^i$ with the normals rendered from a differentiable rasterizer $\mathcal{R}$. The normal loss $\mathcal{L}_n$ is:

$$\begin{aligned} &\mathcal{N}_r^i, \mathcal{S}_r^i = \mathcal{R}(V, F, \pi^i), i = \{0, 1, 2, 3\}, \\ &\mathcal{L}_n = \sum_{i=0}^{3} \|\mathcal{N}_e^i - \mathcal{N}_r^i\|_2^2 + \|\mathcal{S}_e^i - \mathcal{S}_r^i\|_2^2, \end{aligned} \tag{9}$$

where π^i denotes the weak perspective camera of the i-th view, $\mathcal{N}_r^i$ and $\mathcal{S}_r^i$ are the rendered clothed normals and silhouettes from four views, and $\mathcal{S}_e^i$ is the corresponding silhouettes of $\mathcal{N}_e^i$.

To avoid undesired topologies, we introduce the Laplacian term $\mathcal{L}_l = \|\mathbf{L}\mathbf{V}\|_2^2$ and normal consistency term $\mathcal{L}_{nc} = \|1 - n_i n_j\|_2^2$ as regularization for smoothing mesh. Among them, $\mathbf{L}$ is the Laplacian matrix of the deformed mesh. n_i and n_j are the normals of two neighboring faces. Based on the above loss functions, the optimized mesh can be summarized as follows:

$$\arg\min_{\mathbf{V},\mathbf{F}} \ \lambda_n \mathcal{L}_n + \lambda_l \mathcal{L}_l + \lambda_{nc} \mathcal{L}_{nc}, \tag{10}$$

where $\lambda_n = 2, \lambda_l = 40$, and $\lambda_{nc} = 0.1$ are the balanced weights. The iterative optimization Eq.(10) applies the optimizer proposed in (Palfinger 2022) to perform vertex displacement and remeshing on $\mathcal{M} = \{\mathbf{V}, \mathbf{F}\}$ for global adjustment, allowing a reconstruction of complex topologies. During the iterative optimization, we employ the downsampled normals $\{\mathcal{N}_e^i\}_{i=0}^3$ with a resolution of 512×512. The vertices $\mathbf{V}$ can be deformed according to the loss using gradient descent. We also increase the number of faces by interpolating triangles at fixed iterations.

To further enhance local surface fidelity, we introduce a patchify optimizing strategy following the global adjustment. It divides the human mesh and corresponding full-resolution normals into equal-sized patches and performs refinement on these patches individually. This patchify can recover more compact geometric details and avoids getting stuck in local minima. Specifically, we divide each human normal $\mathcal{N}_e^i$ into $4 \times 4 = 16$ equal-sized patches. Each patch $\mathcal{N}_e^{(i,j)}$ is treated as a local supervision window, where (i, j) denotes the j-th patch of the i-th view. We define a dedicated camera projection $\pi^{(i,j)}$ for each mesh patch, under which the mesh is projected onto the patch-specific image plane, producing rendered normals $\mathcal{N}_r^{(i,j)}$. We apply a valid-pixel threshold τ to filter out patches with insufficient valid pixels, reducing noise and computational burden.

Following Eq.(9), the per-patch normal consistency loss $\mathcal{L}_n^{(i,j)}$ is:

$$\mathcal{L}_n^{(i,j)} = \left[\left\|\mathcal{S}_e^{(i,j)}\right\|_0 > \tau\right] \cdot \left\|\mathcal{N}_e^{(i,j)} - \mathcal{N}_r^{(i,j)}\right\|_2^2, \tag{11}$$

where $||\cdot||_0$ computes the number of foreground pixels in the silhouette $\mathcal{S}_e^{(i,j)}$. Like the global adjustment, the vertices and faces of each patch are refined based on Eq.(10) via deforming and remeshing. Finally, fine deformations can be inferred by iteratively optimizing mesh, producing the complete clothed human mesh $\mathcal{M}_r$ with fine-grained details.

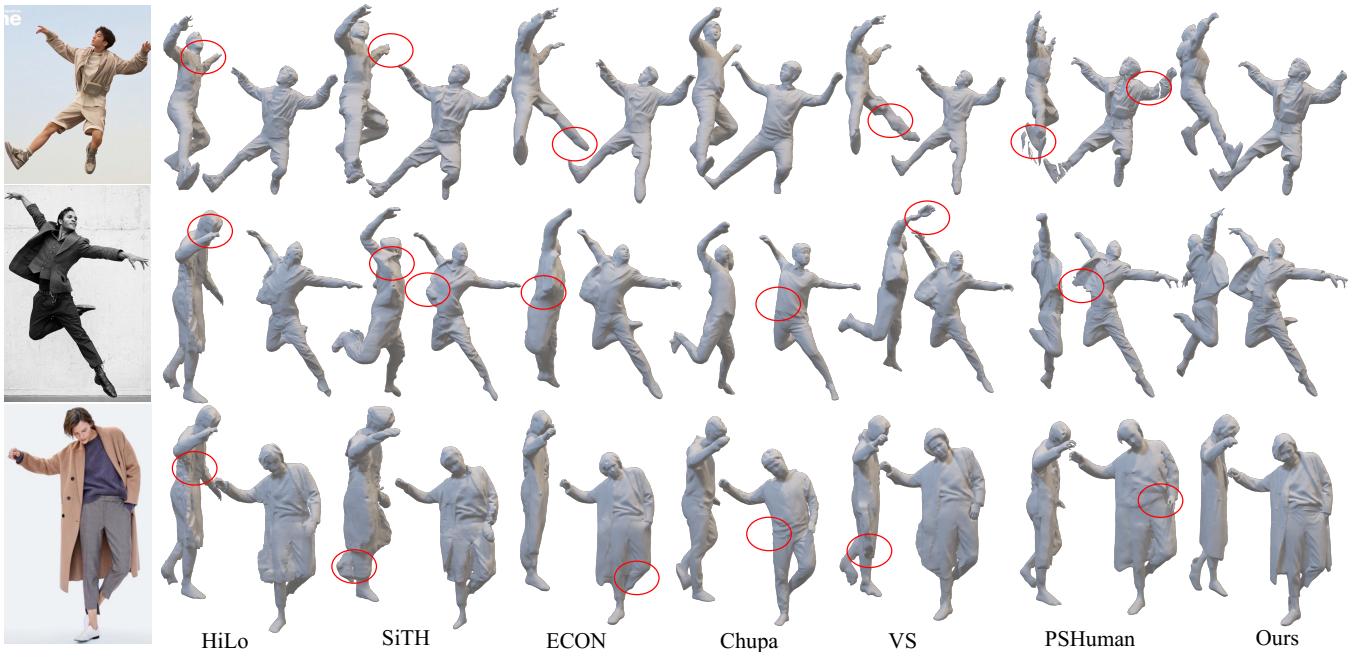

Figure 4: Qualitative comparison with baselines on in-the-wild images. These results are presented from front and side views. Red mark indicates reconstruction error. Please **zoom in** to see details.

Experiments

Experimental Setups

Datasets. We extensively conduct experiments on three benchmarks: THuman2 (including THuman2.0 and THuman2.1 (Yu et al. 2021)), 3DHumans (Jinka et al. 2023), and 2K2K (Han et al. 2023). We chose 2500 scans of THuman2 as training data, and the rest 26 scans as testing data. The training data is only used to train our human normal estimation. The randomly selected 250 raw meshes in 3DHumans and 2K2K are used for testing. We also collect in-the-wild images to evaluate the generalizability of the compared methods.

Baselines. We compare the proposed method with three types of methods, i.e., implicit field, explicit shape, and mesh optimization. 1) HiLo (Yang et al. 2024) and SiTH (Ho et al. 2024), which employ implicit functions to predict occupancy values by neural networks for surface reconstruction. 2) ECON (Xiu et al. 2023) and VSNet (Liu et al. 2024), which are explicit mesh-based reconstruction methods. 3) Chupa (Kim et al. 2023) and PSHuman (Li et al. 2025), which are optimization-based reconstruction methods. All baselines use original implementations provided by authors. These methods and ours are tested on the same datasets and environment.

Evaluation Metrics. To quantitatively evaluate our method, we report geometric accuracy as chamfer distance (CD) and point-to-surface distance (PSD), and local detail as normal consistency (NC). Following ICON (Xiu et al. 2022), all meshes are normalized 3D meshes to a unit space for evaluation.

Comparison

Tab. 1 reports the quantitative results on three public datasets. HumanPro achieves SOTA on all benchmarks and metrics. On 3DHumans with loose clothes, the CD and PSD of competing methods increase substantially, indicating that baselines are difficult to recover complete surface for complex clothes. On 2K2K with high-quality geometry, baselines also have poor evaluation on NC. Notably, our method shows the smallest reduction in quantitative results across datasets, demonstrating stronger robustness. The corresponding qualitative results can be found in the supplementary. These results highlight the performance of HumanPro in accurate and detailed reconstruction across various scenarios.

To further evaluate the generalizability of methods, we showcase reconstruction results of HumanPro and baselines on in-the-wild images. The input images exhibit complex poses and loose clothes. As illustrated in Fig. 4, their performance all experience varying degrees of degradation. The results reconstructed by PSHuman, HiLo, SiTH, and ECON suffer from thickness errors and missing parts. Chupa and VSNet produce some over-smooth surfaces. In contrast, HumanPro can effectively deal with these issues, delivering high-quality clothed human meshes. More visual results can be seen in the supplementary.

Ablation Study

Effectiveness of Geometry-aware Latent Diffusion with a Normal Enhancer. We validate the effectiveness of the high-frequency geometric features and the designed normal enhancer. We study three variants: (a) *Image Only*, takes only the input image $\mathcal{I}$ for diffusion model; (b) *Body Nml*,

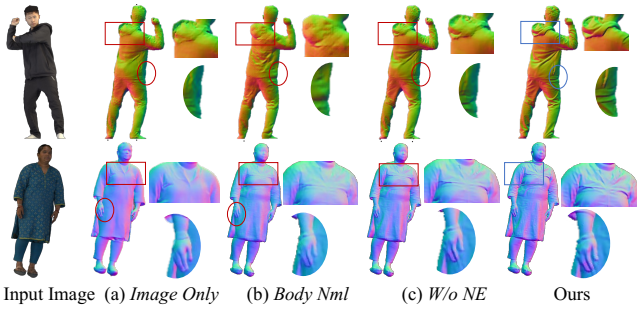

Figure 5: Qualitative ablation comparison of geometry-aware latent diffusion with a normal enhancer.

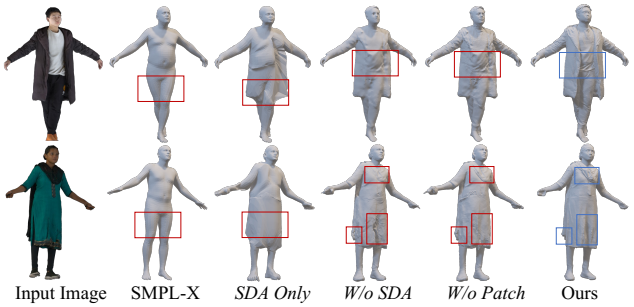

Figure 6: Qualitative ablation comparison of progressive mesh optimization. The results from *W/o SDA*, *W/o Patch*, and ours are executed with the same optimization steps.

takes the input image $\mathcal{I}$ and body normals rendered from the SMPL-X mesh $\mathcal{M}_b$ for diffusion model; (c) *W/o NE*, takes the input image $\mathcal{I}$ and high-frequency geometric features from the SMPL-X mesh $\mathcal{M}_b$, without the normal enhancer. Tab. 2 and Fig. 5 show these results from three variants and ours on THuman2.0 and 3DHumans. These studies validate that our geometric features provide more effective guidance than body normals, and the normal enhancer significantly improves surface details by refining the initial denoising results.

Effectiveness of Progressive Mesh Optimization. We evaluate our progressive mesh optimization. Tab. 2 and Fig. 6 show quantitative and qualitative results. We refer to the one using the shape-aware deformation alignment as *SDA Only*. Compared to the naked SMPL-X mesh, the results produced by *SDA Only* have rough shapes that cover the complete clothes. It shows our *SDA* can deal with large deformation, such as loose clothes. However, limited by deformation ability, its results lack rich surface details. We refer to our method without the shape-aware deformation alignment as *W/o SDA*. It directly uses the SMPL-X mesh as initialization for detail refinement. This refinement can tackle fine deformation to generate intricate details. Due to the lack of full initial shapes, *W/o SDA* easily obtains error surfaces with broken clothes. Additionally, we refer to our method without the patchify optimizing strategy as *W/o Patch*. It directly uses the global mesh optimization for detail refinement, resulting in suboptimal local detail reconstruction. These demonstrate that our progressive mesh optimization can achieve high-quality and efficient reconstruction in a coarse-fine manner.

Type	CD	PSD	NC
Image Only	1.2341	1.4581	0.0995
Body Nml	1.0241	1.1503	0.0695
W/o NE	0.9871	1.0547	0.0625
SDA Only	1.8053	1.5866	0.1061
W/o SDA	1.2451	1.5831	0.0736
W/o Patch	1.0112	1.1233	0.0616
Full	**0.8323**	**0.8234**	**0.0402**

Table 2: Quantitative ablation study. We report the quantitative results of six variants.

Figure 7: Failure cases. We illustrate our limitations on extremely occluded regions and hair.

Limitations

Although HumanPro can produce impressive results with various poses and clothes, supporting extremely occluded regions remains a tremendous challenge. Severe occlusions make it difficult to estimate accurate normals. While our method can recover a complete overall shape, it struggles to reconstruct detailed inter-layer geometries (left of Fig. 7). Moreover, for hair, since all vertices of the generated human mesh derive from the SMPL-X mesh and its expansion, our results exhibit a block shape, making it difficult to achieve satisfactory hair reconstruction (right of Fig. 7).

Conclusion

In this paper, we have proposed HumanPro, a single-view 3D clothed human reconstruction with progressive normal guidance. First, we proposed a geometry-aware latent diffusion model with a normal enhancer to estimate high-quality human normals. This latent diffusion model leverages the input image and the parametric body as conditions to guide the generation of human normals. Then, we proposed a progressive human mesh reconstruction approach from the estimated human normals and the parametric body. This approach consists of the shape-aware deformation alignment and the global-to-patch detail refinement, which can reconstruct a detailed human mesh in a coarse-fine manner. The proposed progressive mesh optimization can balance between global structure and local geometric details. Extensive experiments demonstrated that our method outperforms the SOTA methods, and can cope with challenging poses and clothes.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (No. 62372336) and the Key Research and Development Program for Technological Innovation Project of Hubei Province (No. 2025BAB020).

References

Alldieck, T.; Pons-Moll, G.; Theobalt, C.; and Magnor, M. 2019. Tex2Shape: Detailed Full Human Body Geometry from a Single Image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2293–2303.

Alldieck, T.; Zanfir, M.; and Sminchisescu, C. 2022. Photorealistic Monocular 3D reconstruction of Humans Wearing Clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1506–1515.

Bao, Z.; Fu, G.; Sun, J.; Zhou, J.; Yu, Z.; and Xiao, C. 2025. I2HDiffuser: Image Illumination Harmonization Meets the Diffusion Model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1627–1636.

Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; and Black, M. J. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proceedings of European Conference on Computer Vision*, 561–578.

Chan, K. Y.; Lin, G.; Zhao, H.; and Lin, W. 2022. Integratedpifu: Integrated pixel aligned implicit function for single-view human reconstruction. In *European conference on computer vision*, 328–344.

Corona, E.; Zanfir, M.; Alldieck, T.; Bazavan, E. G.; Zanfir, A.; and Sminchisescu, C. 2023. Structured 3d features for reconstructing controllable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16954–16964.

Feng, Q.; Liu, Y.; Lai, Y.-K.; Yang, J.; and Li, K. 2022. Fof: Learning fourier occupancy field for monocular real-time human reconstruction. *Advances in Neural Information Processing Systems*, 35: 7397–7409.

Gabeur, V.; Franco, J.-S.; Martin, X.; Schmid, C.; and Rogez, G. 2019. Moulding humans: Non-parametric 3D human shape estimation from single images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2232–2241.

Han, S.-H.; Park, M.-G.; Yoon, J. H.; Kang, J.-M.; Park, Y.-J.; and Jeon, H.-G. 2023. High-fidelity 3D human digitization from single 2K resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12869–12879.

He, T.; Collomosse, J.; Jin, H.; and Soatto, S. 2020. Geo-PIFu: Geometry and Pixel Aligned Implicit Functions for Singleview Human Reconstruction. In *Advances in Neural Information Processing Systems*, 1–12.

He, X.; Wu, Z.; Li, X.; Kang, D.; Zhang, C.; Ye, J.; Chen, L.; Gao, X.; Zhang, H.; and Zhuang, H. 2025. Magicman: Generative novel view synthesis of humans with 3D-aware diffusion and iterative refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3437–3445.

Ho, I.; Song, J.; Hilliges, O.; et al. 2024. SiTH: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 538–549.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Huang, Z.; Xu, Y.; Lassner, C.; Li, H.; and Tung, T. 2020. ARCH: Animatable Reconstruction of Clothed Humans. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 3093–3102.

Jinka, S. S.; Srivastava, A.; Pokhariya, C.; Sharma, A.; and Narayanan, P. 2023. Sharp: Shape-aware reconstruction of people in loose clothing. *International Journal of Computer Vision*, 131(4): 918–937.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end Recovery of Human Shape and Pose. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 7122–7131.

Kazhdan, M.; and Hoppe, H. 2013. Screened poisson surface reconstruction. *ACM Transactions on Graphics*, 32(3): 1–13.

Kim, B.; Kwon, P.; Lee, K.; Lee, M.; Han, S.; Kim, D.; and Joo, H. 2023. Chupa: Carving 3D clothed humans from skinned shape priors using 2d diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15965–15976.

Li, P.; Zheng, W.; Liu, Y.; Yu, T.; Li, Y.; Qi, X.; Chi, X.; Xia, S.; Cao, Y.-P.; Chi, W.; Luo, W.; and Guo, Y. 2025. PSHuman: Photorealistic Single-image 3D Human Reconstruction using Cross-scale Multiview Diffusion and Explicit Remeshing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16008–16018.

Li, Y.; Luo, F.; and Xiao, C. 2024. Diffusion-FOF: Single-View Clothed Human Reconstruction via Diffusion-Based Fourier Occupancy Field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9525–9534.

Liu, L.; Li, Y.; Gao, Y.; Gao, C.; Liu, Y.; and Chen, J. 2024. VS: Reconstructing Clothed 3D Human from Single Image via Vertex Shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10498–10507.

Long, X.; Guo, Y.-C.; Lin, C.; Liu, Y.; Dou, Z.; Liu, L.; Ma, Y.; Zhang, S.-H.; Habermann, M.; Theobalt, C.; et al. 2024. Wonder3d: Single image to 3D using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9970–9980.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Mode. *ACM Transactions on Graphics*, 34(6): 1–16.

Luo, C.; Luo, F.; Wang, Y.; Zhao, E.; and Xiao, C. 2024. DLCA-Recon: Dynamic Loose Clothing Avatar Reconstruction from Monocular Videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3963–3971.

Palfinger, W. 2022. Continuous remeshing for inverse rendering. *Computer Animation and Virtual Worlds*, 33(5): e2101.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 165–174.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Saito, S.; Huang, Z.; Natsume, R.; and et al. 2019. PIFu: Pixel-aligned Implicit Function for High-resolution Clothed Duman Digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2304–2314.

Shen, S.; Bao, Z.; Xu, W.; and Xiao, C. 2025. IllumiDiff: Indoor Illumination Estimation From a Single Image With Diffusion Model. *IEEE Transactions on Visualization and Computer Graphics*, 31(10): 7752–7768.

Song, D.-Y.; Lee, H.; Seo, J.; and Cho, D. 2023. DIFu: Depth-guided implicit function for clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8738–8747.

Truong, P.; Danelljan, M.; Yu, F.; and Van Gool, L. 2021. Warp consistency for unsupervised learning of dense correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10346–10356.

Varol, G.; Ceylan, D.; Russell, B.; Yang, J.; Yumer, E.; Laptev, I.; and Schmid, C. 2018. BodyNet: Volumetric inference of 3D human body shapes. In *Proceedings of European Conference on Computer Vision*, 20–36.

Wu, K.; Liu, F.; Cai, Z.; Yan, R.; Wang, H.; Hu, Y.; Duan, Y.; and Ma, K. 2024. Unique3d: High-quality and efficient 3d mesh generation from a single image. *Advances in Neural Information Processing Systems*, 37: 125116–125141.

Xiu, Y.; Yang, J.; Cao, X.; Tzionas, D.; and Black, M. J. 2023. ECON: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 512–523.

Xiu, Y.; Yang, J.; Tzionas, D.; and Black, M. J. 2022. ICON: Implicit Clothed Humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13296–13306.

Yang, X.; Luo, Y.; Xiu, Y.; Wang, W.; Xu, H.; and Fan, Z. 2023. D-if: Uncertainty-aware human digitization via implicit distribution field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9122–9132.

Yang, Y.; Liu, D.; Zhang, S.; Deng, Z.; Huang, Z.; and Tan, M. 2024. 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*, 10671–10681.

Yu, T.; Zheng, Z.; Guo, K.; Liu, P.; Dai, Q.; and Liu, Y. 2021. Function4D: Realtime Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 5746–5756.

Zhang, G.; Yao, N.; Zhang, S.; Zhao, H.; Pang, G.; Shu, J.; and Wang, H. 2025. Multigo: Towards multi-level geometry learning for monocular 3D textured human reconstruction. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 338–347.

Zhang, H.; Tian, Y.; Zhang, Y.; Li, M.; An, L.; Sun, Z.; and Liu, Y. 2023. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 12287–12303.

Zhang, M.; Feng, Q.; Su, Z.; Wen, C.; Xue, Z.; and Li, K. 2024a. Joint2Human: High-quality 3D Human Generation via Compact Spherical Embedding of 3D Joints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1429–1438.

Zhang, Z.; Sun, L.; Yang, Z.; Chen, L.; and Yang, Y. 2024b. Global-correlated 3D-decoupling transformer for clothed avatar reconstruction. *Advances in Neural Information Processing Systems*, 36.

Zhang, Z.; Yang, Z.; and Yang, Y. 2024. 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*, 9936–9947.

Zhao, E.; Sun, J.; Luo, F.; and Xiao, C. 2025. EE-Head: emotion estimation for precise facial expression in NeRF head avatars. *The Visual Computer*, 41(2): 6865–6878.

Zheng, H.; Bao, Z.; Fu, G.; Jiao, X.; and Xiao, C. 2025a. PHR-DIFF: Portrait Highlights Removal via Patch-aware Diffusion Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10555–10563.

Zheng, H.; Xu, W.; Wang, Z.; Lu, X.; and Xiao, C. 2025b. Facial Highlight Removal With Cross-Context Attention and Texture Enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2): 1519–1533.

Zheng, Z.; Yu, T.; Liu, Y.; and Dai, Q. 2021. PaMIR: Parametric modelconditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 3170–3184.

Zheng, Z.; Yu, T.; Wei, Y.; Dai, Q.; and Liu, Y. 2019. DeepHuman: 3D Human Reconstruction From a Single Image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7739–7749.

Zhu, H.; Zuo, X.; Wang, S.; Cao, X.; and Yang, R. 2019. Detailed Human Shape Estimation from a Single Image by Hierarchical Mesh Deformation. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 4491–4500.

Zhu, H.; Zuo, X.; Yang, H.; Wang, S.; Cao, X.; and Yang, R. 2021. Detailed avatar recovery from single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7363–7379.