3D Clothed Human Reconstruction from Sparse Multi-View Images

Jin Gyu Hong[1*] Seung Young Noh[1*] Hee Kyung Lee[2] Won Sik Cheong[2] Ju Yong Chang[1]
[1]Dept of ECE, Kwangwoon University, Seoul, South Korea
[2]Electronics and Telecommunications Research Institute, Daejeon, South Korea
{sony03330,kelvinnoh,jychang}@kw.ac.kr, {lhk95,wscheong}@etri.re.kr

Abstract

Clothed human reconstruction based on implicit functions has recently received considerable attention. In this study, we explore the most effective 2D feature fusion method from multi-view inputs experimentally and propose a method utilizing the 3D coarse volume predicted by the network to provide a better 3D prior. We fuse 2D features using an attention-based method to obtain detailed geometric predictions. In addition, we propose depth and color projection networks that predict the coarse depth volume and the coarse color volume from the input RGB images and depth maps, respectively. Coarse depth volume and coarse color volume are used as 3D priors to predict occupancy and texture, respectively. Further, we combine the fused 2D features and 3D features extracted from our 3D prior to predict occupancy and propose a technique to adjust the influence of 2D and 3D features using learnable weights. The effectiveness of our method is demonstrated through qualitative and quantitative comparisons with recent multi-view clothed human reconstruction models.

1. Introduction

The development of deep-learning networks has facilitated extensive research on clothed human reconstruction. The goal of clothed human reconstruction is to estimate a 3D human mesh that accurately represents the shape of a person dressed in clothes from monocular or multi-view image(s). The estimated 3D human mesh can be utilized in various fields such as clothing design and manufacturing, gaming, augmented reality (AR), and virtual reality (VR).

In this study, we focus on the sparse multi-view clothed human reconstruction problem with the aim of estimating a 3D human mesh from calibrated sparse multi-view images. Conventional multi-view clothed human reconstruction methods typically estimate a 3D human mesh via the following three steps: *(A) 2D feature extraction and fusion from input images, (B) 3D feature extraction from 3D prior, and (C) occupancy prediction from extracted 2D and 3D features.*

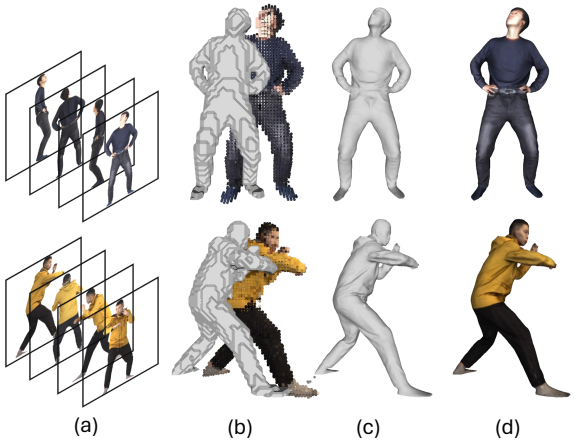

Figure 1. (a) Input multi-view images, (b) coarse depth and color volumes, (c) geometry prediction results, (d) color prediction results.

The process of fusing 2D features in Step *(A)* is essential in multi-view environments and several fusion methods have been proposed for multi-view feature fusion. In PIFu [37], the 2D features are fused using average pooling. ICON [47] uses the visibility provided by SMPL-X [36] to perform feature fusion. DeepMultiCap [57] proposed attention-aware fusion utilizing self-attention mechanisms, and SeSDF [6] suggested occlusion-aware fusion method based on the depth value derived from SMPL-X.

The 3D prior in Step *(B)* is widely used for clothed human reconstruction, regardless of the monocular or multi-view environments. In PIFu [37], the z-coordinates are concatenated to 2D features without a 3D prior; this lack of a 3D prior causes the method to suffer from hard poses. PaMIR [58] and DeepMultiCap [57] voxelize SMPL [29] and extract 3D features from that volume, whereas ICON [47] uses the signed distance value from SMPL-X as a 3D prior. Methods based on parametric body model priors [3, 17, 29, 36, 49] have demonstrated en-

* Equal contribution

hanced robustness to challenging poses and require the use of a fitted SMPL. However, SMPL does not capture details, such as hair or clothing. In DIFu [42], instead of using the SMPL as a 3D prior, a depth volume is generated based on the front/back depth maps, from which 3D features are extracted and used. In this method, the depth maps are unprojected into the 3D space through heuristically calculated offsets, often resulting in unstable depth volume and adversely affecting the reconstruction performance.

When predicting occupancy in Step *(C)* from the fused 2D features extracted in Step *(A)* and the computed 3D features in Step *(B)*, previous works [14, 16, 37, 38, 42, 47, 58] simply combined the features via a concatenation operation and then predicted occupancy using multi-layer perceptron (MLP). However, the degree to which each feature aids the network differs.

In this study, we conducted ablation experiments on the 2D feature fusion method in Step *(A)* in a consistent environment, and compared the quantitative and qualitative results of each method to determine the best fusion method. Further, we propose a depth projection network (DPN) and a color projection network (CPN) for predicting coarse depth volume and coarse color volume from input RGB images and depth maps, respectively. The coarse volumes include detailed information on hair and clothes, unlike the SMPL, and provide a stable human shape prior that differs from the unstable depth volume in the DIFu. We experimentally demonstrated that using the coarse volume as a 3D prior in Step *(B)*, 3D human mesh reconstruction was quantitatively and qualitatively improved compared with other existing 3D priors. Finally, when predicting occupancy in Step *(C)* from fused 2D and 3D features, we combined the weighted features by introducing learnable weights that can modulate the influence of the 2D and 3D features, unlike previous works, in which the features were simply concatenated. The experimental results of the proposed method are shown in Fig. 1.

The contributions of this paper can be summarized as follows:

- An ablation study is performed on existing 2D feature fusion methods to determine the performance differences and suggest the best method. In addition, an ablation study was conducted on the 3D prior and learnable weights to deduce the best combination.
- We propose DPN and CPN to predict coarse volumes that contain detailed information regarding hair and cloth; these methods are more stable than handcrafted methods [42], and the predicted coarse volumes are used as 3D priors in the proposed method.
- Extensive experiments using the Thuman2.0 [53] and BUFF [54] datasets demonstrate that our proposed method performs well in multi-view feature fusion and provides a stable 3D prior. Moreover, we demonstrate that our framework outperforms previously reported clothed human reconstruction methods.

2. Related Work

Implicit function-based representation. Conventional 3D modeling methods include polygonal meshes [20], voxel surfaces [43, 46], point-based representations [62], and etc. According to recent studies [8, 24, 31, 34, 51, 52], implicit function-based representations are widely used in geometric reconstruction, 3D modeling, and rendering because they can effectively represent complex structures. Compared with conventional methods, this method can represent high-dimensional data more concisely. Additionally, implicit representations require less memory and provide natural deformations and flexibility.

Single-view human reconstruction. A parametric 3D human body model [29, 36, 49] is a human modeling method based on pose and shape parameters, that can reproduce various human movements and body shapes. Numerous studies [18, 21, 22, 26, 32, 35, 45, 50, 55] focused on estimating the parametric human model, particularly the SMPL [29] or SMPL-X mesh [36], from single-view images. In particular, the SMPL model can accurately represent various body shapes and poses, is compatible with existing graphics pipelines, and has high computational efficiency. However, it is designed based on the naked human body; therefore presents limitations in expressing detailed features, such as clothing and hair.

Saito *et al.* [37] first introduced an implicit function in the field of clothed human reconstruction and proposed the PIFu model for reconstructing a human mesh using pixel-aligned features. Subsequently, various models [2, 5, 6, 9, 13, 14, 16, 23, 25, 38, 42, 47, 48, 58] employing implicit functions have been proposed for single-view image reconstruction. PaMIR [58] is based on an implicit function and uses a pretrained GraphCMR [22] to estimate the SMPL mesh as a 3D prior for mesh reconstruction. ICON [47] utilizes the signed distance from the query point to the nearest SMPL-X body surface, along with the SMPL-X features and normal. Alldieck *et al.* [2] employed PHORHUM, a pixel-aligned approach, to estimate 3D geometry and infer scene illumination and shading to produce photorealistic results. JIFF [5] uses a clothed human body and parametric face model [4] as a prior for recovering fine facial details. ECON [48] combines implicit representations with explicit body regularization. DIFu [42] generates a back-side image through a trained hallucinator, creates a 3D coarse volume from the front and back depth maps, and uses it as a 3D prior. Although these methods provide visually appealing results, they do not fully address the issues of self-occlusion, depth ambiguity, and the lack of backside information.

Multi-view human reconstruction. In 3D human mesh

reconstruction, more precise reconstruction can be achieved by effectively fusing multi-view images, thus utilizing information that is not available from a single view. Several studies have attempted to achieve accurate human reconstruction from multi-view images. Some methods [1, 10, 11, 15, 27, 28, 57] reconstruct human meshes from multi-view videos, whereas other methods [7, 39, 59, 60] achieve high-fidelity human rendering from multi-view input. In addition, some approaches [37, 58] involve the aggregation of features of each view using average pooling. DiffuStereo [40], a diffusion-based model, is a generative model in a multi-view stereo environment, which effectively handles high-resolution inputs. DoubleField [39] fuses multi-view features in a sparse-view environment for photorealistic human rendering and introduces a view-to-view transformer to learn view-dependent features from high-resolution inputs. SeSDF [6] proposes an occlusion-aware feature fusion strategy that self-calibrates to the SMPL-X model fitted from uncalibrated multi-view images and fuses features considering occlusion. DeepMultiCap [57] is a method for multi-person clothed human reconstruction from multi-view videos that fuses features from each view in an attention-based framework. However, DeepMultiCap can suffer from performance degradation in the real world owing to its dependence on well-fitted SMPL. Both SeSDF and DeepMultiCap rely on the parametric human body model that is unstable in challenging clothes.

Our work builds upon the attention-based 2D feature fusion approach, as in DeepMultiCap, and introduces a novel depth projection network for generating a 3D coarse volume that captures human poses and clothing details. The generated 3D coarse volume is utilized as a robust 3D prior in the final reconstruction process, which improves the accuracy and details of human reconstruction. Furthermore, our model significantly reduces the dependence on the SMPL model of previous methods by using the estimated SMPL model only for depth-map prediction. This approach provides the ability to utilize 3D coarse volumes as a 3D prior without directly using SMPL models to reconstruct complex clothing and difficult poses more accurately. This improves the qualitative and quantitative performances of the overall human mesh reconstruction process.

3. Proposed Method

3.1. Overview

We propose a method for reconstructing clothed human meshes from N multiple calibrated images. Fig. 2 shows the overall structure of the proposed method comprising a depth estimator [42], geometry reconstructor, and texture reconstructor.

The depth estimator $\mathcal{F}_{DE}$ is an off-the-shelf image-to-image translation module that generates multi-view depth maps $D_n \in \mathbb{R}^{1\times512\times512}$ $(n = 1, \ldots, N)$ from input multi-view images $I_n \in \mathbb{R}^{3\times512\times512}$ $(n = 1, \ldots, N)$ using estimated SMPL [29] $\mathcal{M}$ as prior, and is given as follows:

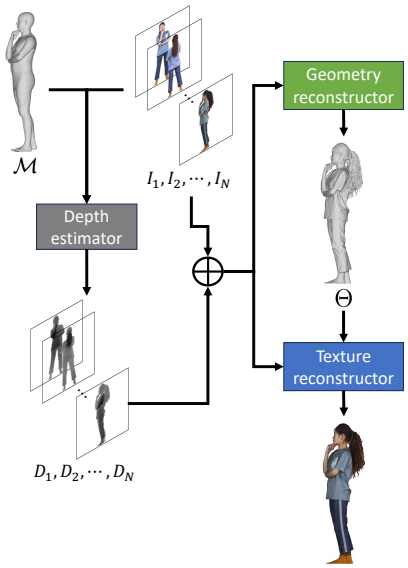

Figure 2. **Overall structure of the proposed method.** The depth estimator generates depth maps $\{D_n\}_{n=1}^N$ from input images $\{I_n\}_{n=1}^N$ and estimated SMPL $\mathcal{M}$. The geometry reconstructor predicts occupancy from $\{I_n\}_{n=1}^N$ and $\{D_n\}_{n=1}^N$, then extracts a 3D human mesh Θ using a marching cube algorithm. The texture reconstructor predicts texture from $\{I_n\}_{n=1}^N$, $\{D_n\}_{n=1}^N$, and Θ and finally outputs a textured 3D human mesh. $\oplus$ denotes concatenate operation.

$$\{D_n\}_{n=1}^N = \mathcal{F}_{DE}(\{I_n\}_{n=1}^N, \mathcal{M}). \tag{1}$$

Geometry reconstructor $\mathcal{F}_{GEO}$ predicts the occupancy via implicit function from $\{I_n\}_{n=1}^N$ and $\{D_n\}_{n=1}^N$, and then converts the predicted occupancy to a 3D human mesh Θ using the marching cubes algorithm [30]:

$$\Theta = \mathcal{F}_{GEO}(\{I_n\}_{n=1}^N, \{D_n\}_{n=1}^N). \tag{2}$$

Texture reconstructor $\mathcal{F}_{TEX}$ takes $\{I_n\}_{n=1}^N$, $\{D_n\}_{n=1}^N$, and the vertex coordinates $\mathcal{K} \in \mathbb{R}^{K\times3}$ of Θ computed from $\mathcal{F}_{GEO}$ and outputs the RGB value $\hat{C} \in \mathbb{R}^{K\times3}$ corresponding to each vertex $\mathcal{K}$ and the parameters $\Omega \in \mathbb{R}^{K\times1\times N}$ and $\lambda \in \mathbb{R}^{K\times1}$ for computing the final texture:

$$\{\hat{C}, \Omega, \lambda\} = \mathcal{F}_{TEX}(\{I_n\}_{n=1}^N, \{D_n\}_{n=1}^N, \mathcal{K}), \tag{3}$$

where K is the number of vertices in the 3D human mesh Θ. Geometry reconstructor and texture reconstructor are described in detail in Sections 3.2 and 3.3, respectively.

3.2. Geometry Reconstructor

Geometry reconstructor $\mathcal{F}_{GEO}$ predicts the 3D human mesh Θ from $\{I_n\}_{n=1}^N$ and $\{D_n\}_{n=1}^N$ as shown in Fig. 3. $\{I_n\}_{n=1}^N$ and $\{D_n\}_{n=1}^N$ are concatenated and then fed into

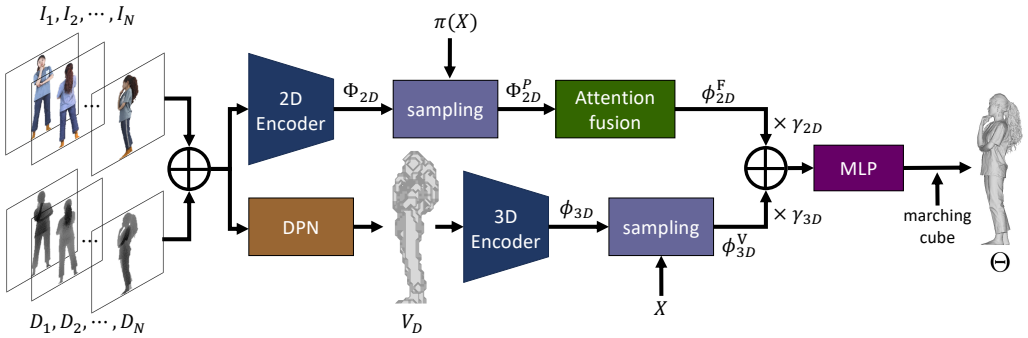

Figure 3. **Geometry Reconstructor.** $\{I_n\}_{n=1}^N$ and $\{D_n\}_{n=1}^N$ are concatenated and fed into the 2D encoder and DPN. The 2D feature Φ_{2D} extracted from the 2D encoder is sampled as pixel-aligned feature Φ^P_{2D} and fused into fused feature ϕ^F_{2D} through the attention-aware fusion module. The coarse depth volume V_D output from the DPN is used as a 3D prior to extract voxel-aligned feature ϕ^V_{3D}. Each feature is converted to a weighted feature by learnable weights γ_{2D} and γ_{3D}, respectively, and fed into the MLP.

the 2D encoder and DPN. DPN is a module that outputs coarse depth volume $V_D \in \mathbb{R}^{64\times64\times64}$ from $\{I_n\}_{n=1}^N$ and $\{D_n\}_{n=1}^N$ and is described in detail in Section 3.4.

The 2D encoder f_{HG} computes the 2D feature maps $\Phi_{2D} = \{\Phi^{(s)}_{2D} \in \mathbb{R}^{N\times256\times128\times128}\}_{s=1}^4$ from the concatenated input. This encoder is based on a stacked hourglass network [33] described in [37], where s is the stack index of the stacked hourglass network. Φ_{2D} is extracted as follows:

$$\Phi_{2D} = f_{HG}(\{I_n\}_{n=1}^N \oplus \{D_n\}_{n=1}^N). \tag{4}$$

The 3D encoder f_{VE} computes the 3D feature volumes $\phi_{3D} = \{\phi^{(r)}_{3D} \in \mathbb{R}^{32\times64\times64\times64}\}_{r=1}^3$ from the coarse depth volume V_D. The 3D encoder is based on the volume encoder as in [58], where r is the residual block [12] index of the volume encoder. ϕ_{3D} is extracted as follows:

$$\phi_{3D} = f_{VE}(V_D). \tag{5}$$

The sampling module samples voxel-aligned feature $\phi^V_{3D} \in \mathbb{R}^{32}$ and pixel-aligned feature $\Phi^P_{2D} \in \mathbb{R}^{N\times256}$ from 3D feature volumes ϕ_{3D} and 2D feature map Φ_{2D}, respectively, utilizing the 3D query point $X \in \mathbb{R}^3$ and its 2D projection $\pi(X) \in \mathbb{R}^2$ as follows:

$$\Phi^P_{2D} = \mathcal{B}(\Phi_{2D}, \pi(X)),\ \phi^V_{3D} = \mathcal{T}(\phi_{3D}, X), \tag{6}$$

where $\mathcal{B}(\cdot)$ and $\mathcal{T}(\cdot)$ denote bilinear and trilinear interpolations, respectively.

Φ^P_{2D} is transformed into Φ_q, Φ_k, and Φ_v by learnable weights W_q, W_k, and W_v, respectively. Note that Φ_q, Φ_k, and Φ_v are the query, key, and value features, respectively:

$$\Phi_q = \Phi^P_{2D}W_q,\ \Phi_k = \Phi^P_{2D}W_k,\ \Phi_v = \Phi^P_{2D}W_v, \tag{7}$$

where W_q, W_k, and W_v are learnable projection matrices for multi-head attention [57].

Subsequently, we compute the fused pixel-aligned feature $\phi^F_{2D} \in \mathbb{R}^{256}$ from Φ_q, Φ_k, and Φ_v using the self-attention encoder [44]. We use two self-attention encoder layers to obtain ϕ^F_{2D}. Within each encoder layer [57], Φ_q, Φ_k, and Φ_v pass through self-attention mechanism and point-wise feed-forward process as follows:

$$\begin{aligned}\Phi_{att} &= attention(\Phi_q, \Phi_k, \Phi_v)\\ &= softmax(\tfrac{\Phi_q(\Phi_k)^T}{\sqrt{d_k}}),\end{aligned} \tag{8}$$

$$\phi^F_{2D} = FF(\Phi_{att}), \tag{9}$$

where $d_k \in \mathbb{R}$ is a constant used to prevent the gradient vanishing problem, and $FF(\cdot)$ denotes the point-wise feed-forward process [44].

Instead of directly using the fused pixel-aligned feature ϕ^F_{2D} and the voxel-aligned feature ϕ^V_{3D} for occupancy prediction MLP f_{geo}, we feed the weighted feature computed using the learnable weights $\gamma_{2D} \in \mathbb{R}$ and $\gamma_{3D} \in \mathbb{R}$ into the occupancy prediction MLP f_{geo} as follows:

$$f_{geo}(\gamma_{2D}\phi^F_{2D} \oplus \gamma_{3D}\phi^V_{3D}) \longmapsto [0, 1]. \tag{10}$$

After the occupancy corresponding to the grid points is predicted by f_{geo}, it is converted into a 3D human mesh Θ using the marching cube algorithm.

3.3. Texture Reconstructor

Texture reconstructor $\mathcal{F}_{TEX}$ predicts a textured 3D human mesh from $\{I_n\}_{n=1}^N$, $\{D_n\}_{n=1}^N$, and a 3D human mesh Θ predicted by geometry reconstructor $\mathcal{F}_{GEO}$ as shown in Fig. 4. $\{I_n\}_{n=1}^N$ and $\{D_n\}_{n=1}^N$ are concatenated and fed into a 2D encoder and CPN. The CPN is the module outputs the coarse color volume $V_C \in \mathbb{R}^{64\times64\times64\times3}$ from $\{I_n\}_{n=1}^N$ and $\{D_n\}_{n=1}^N$ and is described in detail in Section 3.4.

The 2D encoder f_{CG} computes a 2D feature map $\bar{\Phi}_{2D} \in \mathbb{R}^{N\times256\times128\times128}$ from the concatenated input, where the 2D encoder is based on CycleGAN [61] as described in [58]. $\bar{\Phi}_{2D}$ is extracted as follows:

$$\bar{\Phi}_{2D} = f_{CG}(\{I_n\}_{n=1}^N \oplus \{D_n\}_{n=1}^N). \tag{11}$$

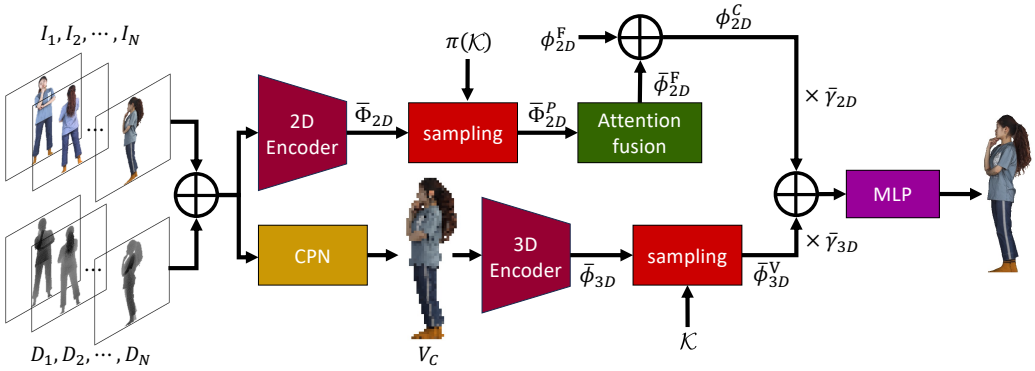

Figure 4. **Texture Reconstructor.** The texture reconstructor has almost the same structure as the geometry reconstructor with three differences. First, it uses CPN instead of DPN and thus uses V_C instead of V_D as the 3D prior. It also uses vertex $\mathcal{K}$ of Θ instead of query point X in the sampling module. Finally, it concatenates ϕ^F_{2D} and $\bar{\phi}^F_{2D}$ before being input to the MLP. Please note that the features generated from texture reconstructor are marked with a bar symbol to distinguish them from geometry reconstructor.

The 3D encoder f_{TVE} [58] computes the 3D feature volume $\bar{\phi}_{3D} \in \mathbb{R}^{N\times 32\times 64\times 64\times 64}$ from the coarse color volume V_C. $\bar{\phi}_{3D}$ is extracted as follows:

$$\bar{\phi}_{3D} = f_{TVE}(V_C). \tag{12}$$

The sampling process is the same as the sampling module described in Section 3.2, except that the vertex $\mathcal{K}$ of Θ is used as input instead of a 3D query point X as follows:

$$\bar{\Phi}^P_{2D} = \mathcal{B}(\bar{\Phi}_{2D}, \pi(\mathcal{K})),\ \bar{\phi}^V_{3D} = \mathcal{T}(\bar{\phi}_{3D}, \mathcal{K}). \tag{13}$$

Similar to the attention-aware fusion module mentioned in Section 3.2, we compute a fused pixel-aligned feature $\bar{\phi}^F_{2D}$ from a stacked pixel-aligned feature $\bar{\Phi}^P_{2D}$ by transforming it with learnable projection matrices, applying a self-attention mechanism, and point-wise feed-forward as follows:

$$\bar{\Phi}_q = \bar{\Phi}^P_{2D}\bar{W}_q,\ \bar{\Phi}_k = \bar{\Phi}^P_{2D}\bar{W}_k,\ \bar{\Phi}_v = \bar{\Phi}^P_{2D}\bar{W}_v, \tag{14}$$

$$\begin{aligned}\bar{\Phi}_{att} &= attention(\bar{\Phi}_q, \bar{\Phi}_k, \bar{\Phi}_v) \\ &= softmax(\tfrac{\bar{\Phi}_q(\bar{\Phi}_k)^T}{\sqrt{d_k}}),\end{aligned} \tag{15}$$

$$\bar{\phi}^F_{2D} = FF(\bar{\Phi}_{att}). \tag{16}$$

Instead of using $\bar{\phi}^F_{2D}$ alone for texture prediction, we obtain ϕ^F_{2D} from Section 3.2 and combine it with $\bar{\phi}^F_{2D}$ to get a concatenated feature [37] as follows:

$$\phi^C_{2D} = \phi^F_{2D} \oplus \bar{\phi}^F_{2D}. \tag{17}$$

The weighted features obtained through learnable weights $\bar{\gamma}_{2D}$ and $\bar{\gamma}_{3D}$ are fed into the texture prediction MLP f_{tex} as follows:

$$f_{tex}(\bar{\gamma}_{2D}\phi^C_{2D} \oplus \bar{\gamma}_{3D}\bar{\phi}^V_{3D}) \longmapsto [\hat{C}, \Omega, \lambda]. \tag{18}$$

Rather than directly using the RGB value $\hat{C}$ corresponding to each vertex $\mathcal{K}$ computed by f_{tex}, the final texture C_{fin} is calculated as a linear combination [42, 58] using the estimated parameters $\Omega = \{\omega_n \in \mathbb{R}^{K\times 1}\}^N_{n=1}$ and λ as follows:

$$C_{img} = \sum_{n=1}^{N} \omega_n \mathcal{B}(I_n, \pi(\mathcal{K})), \tag{19}$$

$$C_{fin} = \lambda C_{img} + (1-\lambda)\hat{C}, \tag{20}$$

where C_{img} is the result of sampling the RGB value corresponding to each vertex $\mathcal{K}$ from $\{I_n\}^N_{n=1}$ and weighted sum by Ω, and $\{\omega_n\}^N_{n=1}$ are the weights that determine the extent to which RGB values from each view are represented when calculating C_{img}. λ is the weight that determines the strength of contribution of C_{img} and $\hat{C}$.

3.4. Depth/Color Projection Network

From multi-view input images $\{I_n\}^N_{n=1}$ and estimated multi-view depth maps $\{D_n\}^N_{n=1}$, the DPN and CPN predict coarse depth volumes $V_D \in \mathbb{R}^{64\times 64\times 64}$ and coarse color volumes $V_C \in \mathbb{R}^{64\times 64\times 64\times 3}$ as shown in Fig. 5. The predicted depth and color volumes are used as 3D prior for the geometry reconstructor $\mathcal{F}_{GEO}$ and texture reconstructor $\mathcal{F}_{TEX}$, respectively.

The 2D encoder f_{RS} extracts a 2D feature map $\psi \in \mathbb{R}^{N\times 2048\times 16\times 16}$ from the concatenated input. ResNet50 [12] is used as the 2D encoder, and ψ is extracted as follows:

$$\psi = f_{RS}(\{I_n\}^N_{n=1} \oplus \{D_n\}^N_{n=1}). \tag{21}$$

Next, a deconvolution layer, f_{DC}, is used to obtain a feature map $\psi_{DC} \in \mathbb{R}^{N\times 256\times 64\times 64}$ with increased spatial resolution as follows:

$$\psi_{DC} = f_{DC}(\psi). \tag{22}$$

Then, we use the 'repeat' operation to extend the 2D feature map ψ_{DC} to the 3D space and transform it into an $\mathbb{R}^{256\times 64\times 64\times 64}$-shaped feature map. Afterwards, we concatenate the depth coordinate tensor $\tau \in \mathbb{R}^{1\times 64\times 64\times 64}$,

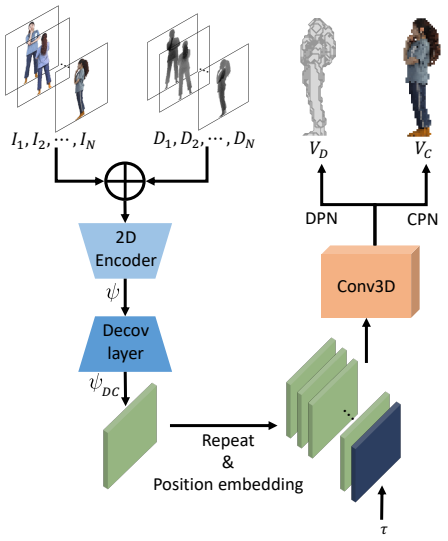

Figure 5. **Depth(Color) Projection Network.** The depth(color) projection network generates coarse depth(color) volume as a 3D prior by taking $\{I_n\}_{n=1}^{N}$ and $\{D_n\}_{n=1}^{N}$ as inputs. The generated 3D volume is utilized as a 3D prior in both geometry reconstructor and texture reconstructor.

which is divided into uniform steps from -1 to 1 to represent the depth information [37]. Then, the obtained feature map is processed through the 3D convolution operation f_{CV} to predict occupancy as follows:

$$f_{CV}(\psi_{DC} \oplus \tau) \longmapsto [0, 1]. \tag{23}$$

CPN follows the same process as DPN until $\psi_{DC} \oplus \tau$ is fed into f_{CV}. However, after it is fed into f_{CV}, the color map is predicted instead of the occupancy as follows:

$$f_{CV}(\psi_{DC} \oplus \tau) \longmapsto [0, 1]^3. \tag{24}$$

3.5. Loss Functions

We sample M 3D query points to train the geometry reconstructor $\mathcal{F}_{GEO}$. For each query point, the mean squared error between the predicted occupancy $\hat{\mathcal{O}}$ by f_{geo} and the ground-truth (GT) occupancy $\mathcal{O}^*$ are used as the loss function:

$$\mathcal{L}_{geo} = \frac{1}{M}\sum_{i=1}^{M}(\hat{\mathcal{O}}_i - \mathcal{O}_i^*)^2. \tag{25}$$

The texture reconstructor has two outputs: intermediate texture $\hat{C}$ predicted by f_{tex} and final texture C_{fin} calculated by a linear combination. We apply the L1 loss to both outputs [58] as follows:

$$\mathcal{L}_{color} = \frac{1}{K}\sum_{i=1}^{K}(\|\hat{C}_i - C_i^*\|_1 + \|C_{fin_i} - C_i^*\|_1), \tag{26}$$

where C_i^* is the GT texture for the i^{th} vertex of Θ. As shown in Eq. (19) and Eq. (20), the larger the value of λ, the more we can get the RGB values of $\{I_n\}_{n=1}^{N}$. Therefore, we use a loss [42] to avoid making the value of λ small by using a loss as follows:

$$\mathcal{L}_{\lambda} = -\mathbb{E}[\log(\lambda)]. \tag{27}$$

The final loss for the texture reconstructor is defined as follows:

$$\mathcal{L}_{tex} = \mathcal{L}_{color} + \mathcal{L}_{\lambda}. \tag{28}$$

Our proposed DPN predicts a probability $\hat{d}$ between 0 and 1 within a $64 \times 64 \times 64$ grid from the input, and produces a coarse depth volume V_D indicating occupancy. The GT 3D volume map has the same size and shape, and the network is trained to minimize the difference from the GT occupancy d^* using binary cross entropy loss as follows:

$$\mathcal{L}_{dpn} = \frac{1}{N}\sum_{i=1}^{N}[d_i^* \log(\hat{d}_i) + (1 - d_i^*)\log(1 - \hat{d}_i)]. \tag{29}$$

CPN predicts continuous color values $\hat{c}$ between 0 and 1 within the $64 \times 64 \times 64$ grid from the input, and produces a coarse color volume V_C indicating texture. The network uses mean squared error to minimizes the difference between the predicted color values $\hat{c}$ and GT color values c^* by using the mean squared error as follows:

$$\mathcal{L}_{cpn} = \frac{1}{N}\sum_{i=1}^{N}\|\hat{c}_i - c_i^*\|_2^2. \tag{30}$$

4. Experimental Results

4.1. Implementation Details

To train all methods, the THuman2.0 [53] dataset was used, with the training images subjected to color jittering augmentation. We sampled 8,000 points from the mesh of each human subject during the training process. For all methods requiring SMPL [29] parameters, we used the parameters predicted by the pre-trained GCMR [22]. When training our method, the DPN, CPN, geometry reconstructor, and texture reconstructor were trained separately. Adam [19] was used as the optimizer and the learning rate was set to 1e-4. We utilized four RTX 4090 GPUs for training, with approximately 24 h required for the DPN and CPN, and approximately 36 h required for the geometry and texture reconstructors.

4.2. Datasets

The THuman2.0 dataset provides 526 high-quality 3D human scans along with GT SMPL parameters. We used 495 human scans for training and 31 human scans for evaluation. All the training images were generated through a

	THuman2.0					BUFF				
Models	P2S ↓	Chamfer ↓	Normal ↓	MSE ↓	LPIPS ↓	P2S ↓	Chamfer ↓	Normal ↓	MSE ↓	LPIPS ↓
PIFu [37]	2.644	3.016	0.122	0.089	0.133	4.186	4.739	0.196	0.099	0.190
PaMIR [58]	1.544	1.568	0.063	0.055	0.080	1.181	1.209	0.052	0.013	0.076
DeepMultiCap [57]	1.311	1.313	0.052	0.049	0.065	0.875	0.885	0.039	0.011	0.061
Ours	**0.921**	**0.923**	**0.039**	**0.040**	**0.056**	**0.839**	**0.842**	**0.035**	**0.010**	**0.058**

Table 1. **Quantitative result for THuman2.0 and BUFF datasets.**

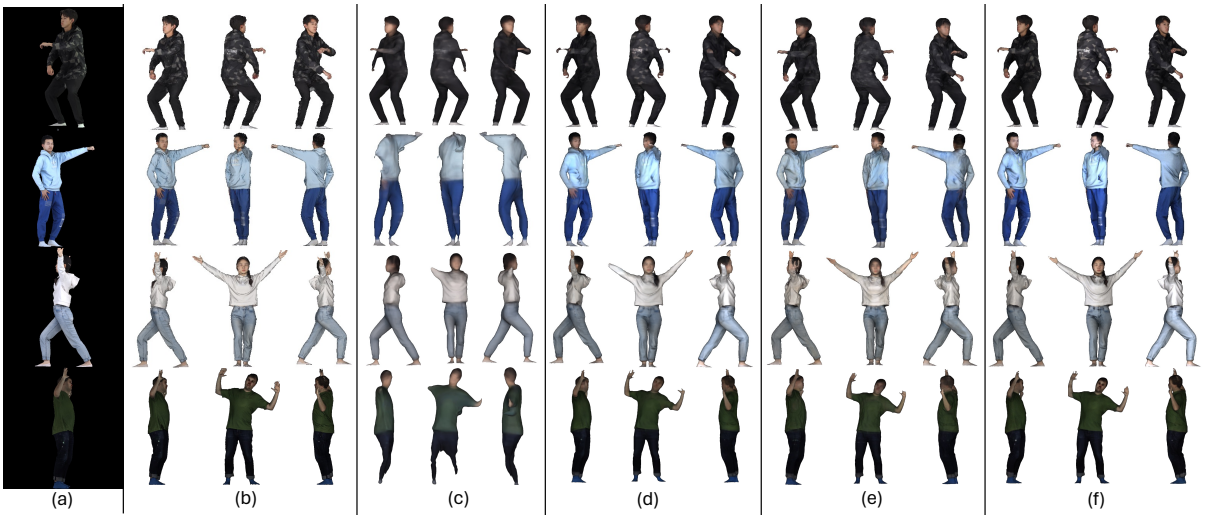

Figure 6. **Qualitative result for THuman2.0 and BUFF datasets.** (a) 1 of 4views, (b) GT mesh, (c) PIFu, (d) PaMIR, (e) DeepMultiCap, (f) Ours. The bottom row is the results from the BUFF dataset.

rendering process from a 3D human mesh at every degree using OpenGL. To test the generalization ability of the proposed method, we employed the BUFF [54] dataset. BUFF provides a sequence of human meshes for five subjects, with two clothing styles and three motions. We used all 143 meshes for the evaluation, and the evaluation images were similarly generated using OpenGL.

4.3. Evaluation Metrics

The P2S distance evaluates the precision of mesh reconstruction by averaging the shortest distance between points on the predicted mesh and the actual mesh surface. The chamfer distance evaluates the geometric similarity between the predicted mesh and the actual mesh by selecting points from both meshes and averaging their distances. Normal difference quantifies the discrepancy between the four normal maps of the predicted mesh and the GT normal map, calculated after simulating a virtual camera rotation at 0°, 90°, 180°, and 270°, and subsequently reprojecting into 2D.

Texture inference performance was evaluated using by two metrics, MSE and LPIPS [56], after the mesh was reprojected to 2D. MSE measures the accuracy as the root mean square of the pixel difference between the predicted and GT textures. LPIPS evaluates the similarity between image patches using a trained VGG [41] network.

4.4. Comparison with Existing Methods

We compared our proposed model with previous multi-view methods, namely, PIFu [37], PaMIR [58], and DeepMultiCap [57]. For a fair comparison, all the methods were processed in the same environment, and trained on THuman2.0 dataset. We render 4-view images as the input. As reported in Table 1, the proposed method significantly outperforms other existing methods by the large margins on THuman2.0. Among existing methods, DeepMultiCap also fuses 2D features by attention-aware method, while our method consistently outperforms DeepMultiCap. Even in the BUFF dataset, the performance gap with other models is relatively small, but the proposed method still achieves the best performance. In addition, as shown in Fig. 6, our model produces more plausible results for both geometry and texture. In particular, the detailed geometry parts (e.g., hand and wrist) and face textures are generated plausibly. We guess that this is due to the effect of the DPN and CPN that create the coarse volumes.

4.5. Ablation Experiments

2D feature fusion method. In this experiment, we explore the effect of 2D feature fusion methods on 3D clothed human reconstruction using PIFu [37] as a baseline. We evaluate the performance of different 2D feature fusion methods while preserving the z-coordinate, the 3D feature in PIFu. We compare existing fusion methods: (1) average-

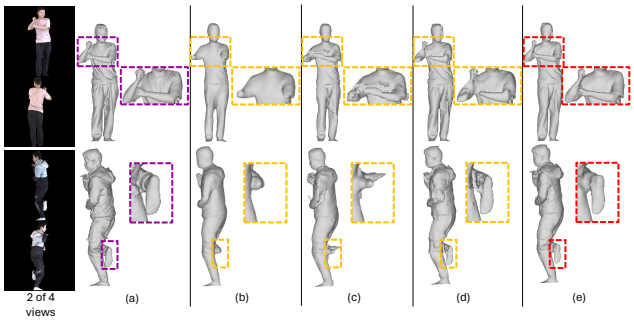

Figure 7. **Qualitative comparison of 2D feature fusion methods.** (a) GT mesh, (b) average, (c) SMPLX-vis, (d) occlusion, (e) attention (ours). Best viewed in PDF with zoom.

2D fusion	3D prior	P2S	Chamfer	Normal
average [37]	Z-coord [37]	2.645	3.020	0.122
SMPLX-vis [47]		2.817	2.957	0.126
occlusion [6]		1.378	1.340	0.053
attention [57]		**1.038**	**1.043**	**0.043**

Table 2. **Ablation results for the 2D feature fusion method.**

pooling [37], (2) SMPL-X visibility [47], (3) occlusion-aware fusion method [6], and (4) attention-aware fusion method [57]. Average-pooling simply averages the features of multiple views, which means that the features of important views, including key features, are not fully reflected. In Fig. 7(b), we observe that cropping occurs at the limbs, such as arms and legs. In the case of the feature selection method that utilizes SMPL-X visibility to determine features, differences from the actual mesh occur due to the estimated SMPL-X, which causes noise on the surface as shown in Fig. 7(c). The occlusion-aware feature fusion method uses the occlusion of the SMPL mesh obtained through ray tracing to control the influence of the features, and shows relatively favorable results, but has limitations when it is difficult to accurately estimate the SMPL mesh. On the other hand, as shown in Table 2 and Fig.7(e), the attention-based 2D feature fusion method can effectively fuse multiple features and perform well in non-SMPL environments.

3D prior. After fixing the best performing attention-aware fusion method, we compare our coarse depth volume with other 3D priors, namely, (1) z-cooordinate [37], (2) voxelized SMPL volume [58], and (3) depth volume by heuristic depth projection [42]. It can be seen in the Fig. 8 that the existing 3D priors lose the details of the clothes or body, but our prior recovers the details well and appropriately. Furthermore, our method demonstrates quantitative improvement over existing methods, as shown in Table 3.

Learnable parameters. We conducted an ablation experiment for our proposed learnable parameters $\gamma_{2D} \in \mathbb{R}$ and $\gamma_{3D} \in \mathbb{R}$, which adjust the influence of 2D and 3D features to optimize learning and thereby improve 3D human reconstruction. The results of this experiment can be found in Table 4, which shows the effectiveness of our approach.

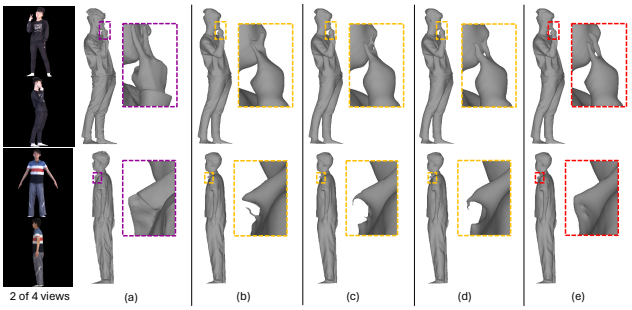

Figure 8. **Qualitative comparison of 3D prior methods.** (a) GT mesh, (b) Z-coordinate, (c) voxelized SMPL, (d) depth volume by depth projection, (e) coarse depth volume (ours).

2D fusion	3D prior	P2S	Chamfer	Normal
attention	Z-coord [37]	1.038	1.043	0.043
	VSM [58]	0.955	0.960	0.040
	DP [42]	0.951	0.951	0.040
	CDV(ours)	**0.939**	**0.944**	**0.039**

Table 3. **Ablation results for the 3D prior method.** Please note that Z-coord, VSM, DP, and CDV denote z-coordinate, voxelized SMPL, depth volume by depth projection, and coarse depth volume, respectively.

Moodels	P2S	Chamfer	Normal
ours w/o γ	0.939	0.944	**0.039**
ours	**0.921**	**0.922**	**0.039**

Table 4. **Ablation results for the learnable weights.**

5. Conclusion

In this study, we developed a method for reconstructing high quality 3D clothed human from calibrated sparse multi-view images. Herein, the 2D features are fused using through an attention-based fusion module, and the 3D features are extracted using the coarse depth volume output from the DPN as a 3D prior. These features are transformed into weighted features using learnable weights, and the occupancy is predicted using features that reflect their importance. We proposed a texture reconstructor that uses the coarse color volume predicted by the CPN as a 3D prior. Our experiments demonstrated that the 3D coarse volume significantly improves mesh reconstruction compared with other 3D priors, and highlighted the effectiveness of attention-based 2D fusion module. The proposed method outperforms existing methods in terms of both qualitative and quantitative results.

Acknowledgement

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2018-0-00207, Immersive Media Research Laboratory)

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 3

[2] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1506–1515, 2022. 2

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH*, pages 408–416. 2005. 1

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *ACM SIGGRAPH*, 1999. 2

[5] Yukang Cao, Guanying Chen, Kai Han, Wenqi Yang, and Kwan-Yee K Wong. Jiff: Jointly-aligned implicit face function for high quality single view clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2729–2739, 2022. 2

[6] Yukang Cao, Kai Han, and Kwan-Yee K Wong. Sesdf: Self-evolved signed distance field for implicit 3d clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4647–4657, 2023. 1, 2, 3, 8

[7] Jianchuan Chen, Wentao Yi, Liqian Ma, Xu Jia, and Huchuan Lu. Gm-nerf: Learning generalizable model-based neural radiance fields from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20648–20658, 2023. 3

[8] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 2

[9] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6970–6981, 2020. 2

[10] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM TOG*, 34(4):1–13, 2015. 3

[11] Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. In *ACM SIGGRAPH*, pages 1–10. 2008. 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4, 5

[13] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *Advances in Neural Information Processing Systems*, 33:9276–9287, 2020. 2

[14] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11046–11056, 2021. 2

[15] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep volumetric video from very sparse multi-view performance capture. In *Proceedings of the European Conference on Computer Vision*, pages 336–354, 2018. 3

[16] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 2

[17] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 1

[18] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 2

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 6

[20] Leif P Kobbelt, Jens Vorsatz, Ulf Labsik, and Hans-Peter Seidel. A shrink wrapping approach to remeshing polygonal surfaces. In *Computer Graphics Forum*, pages 119–130. Wiley Online Library, 1999. 2

[21] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 2

[22] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019. 2, 6

[23] Jungeun Lee, Sanghun Kim, Hansol Lee, Tserendorj Adiya, and Hwasup Lim. Pidiffu: Pixel-aligned diffusion model for high-fidelity clothed human reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5172–5181, 2024. 2

[24] Tianyang Li, Xin Wen, Yu-Shen Liu, Hua Su, and Zhizhong Han. Learning deep implicit functions for 3d shapes with dynamic code clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12840–12850, 2022. 2

[25] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. High-fidelity clothed avatar reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8662–8672, 2023. 2

[26] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1954–1963, 2021. 2

[27] Yebin Liu, Qionghai Dai, and Wenli Xu. A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE transactions on Visualization and Computer Graphics*, 16(3):407–418, 2009. 3

[28] Yebin Liu, Juergen Gall, Carsten Stoll, Qionghai Dai, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2720–2735, 2013. 3

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 34(6):248:1–248:16, 2015. 1, 2, 3, 6

[30] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 3

[31] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2

[32] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *Proceedings of the European Conference on Computer Vision*, pages 752–768. Springer, 2020. 2

[33] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 483–499. Springer, 2016. 4

[34] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2

[35] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018. 2

[36] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10967–10977, 2019. 1, 2

[37] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[38] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 2

[39] Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yan-Pei Cao, Tao Yu, and Yebin Liu. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15872–15882, 2022. 3

[40] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *Proceedings of the European Conference on Computer Vision*, pages 702–720. Springer, 2022. 3

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 7

[42] Dae-Young Song, HeeKyung Lee, Jeongil Seo, and Donghyeon Cho. Difu: Depth-guided implicit function for clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8738–8747, 2023. 2, 3, 5, 6, 8

[43] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision*, pages 20–36, 2018. 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 4

[45] Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13033–13042, 2021. 2

[46] Guy Windreich, Nahum Kiryati, and Gabriele Lohmann. Voxel-based surface area estimation: from theory to practice. *Pattern Recognition*, 36(11):2531–2541, 2003. 2

[47] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13286–13296. IEEE, 2022. 1, 2, 8

[48] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 512–523, 2023. 2

[49] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6183–6192, 2020. 1, 2

[50] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7760–7770, 2019. 2

[51] Xianghui Yang, Guosheng Lin, Zhenghao Chen, and Luping Zhou. Neural vector fields: Implicit representation by explicit learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16727–16738, 2023. 2

[52] Jianglong Ye, Yuntao Chen, Naiyan Wang, and Xiaolong Wang. Gifs: Neural implicit function for general shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12829–12839, 2022. 2

[53] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5746–5756, 2021. 2, 6

[54] Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4191–4200, 2017. 2, 7

[55] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11446–11456, 2021. 2

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 7

[57] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6239–6249, 2021. 1, 3, 4, 7, 8

[58] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 3170–3184, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[59] Taotao Zhou, Kai He, Di Wu, Teng Xu, Qixuan Zhang, Kuixiang Shao, Wenzheng Chen, Lan Xu, and Jingyi Yu. Relightable neural human assets from multi-view gradient illuminations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4315–4327, 2023. 3

[60] Tiansong Zhou, Jing Huang, Tao Yu, Ruizhi Shao, and Kun Li. Hdhuman: High-quality human novel-view rendering from sparse views. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 3

[61] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2223–2232, 2017. 4

[62] Matthias Zwicker and Mark Pauly. Point-based computer graphics. In *ACM SIGGRAPH*, pages 799–800. 2004. 2