Complete 3D Human Reconstruction from a Single Incomplete Image

Junying Wang[1] Jae Shin Yoon[2] Tuanfeng Y. Wang[2]
Krishna Kumar Singh[2] Ulrich Neumann[1]
[1]University of Southern California [2]Adobe Research
{junyingw,uneumann}@usc.edu {jaeyoon,yangtwan,krishsin}@adobe.com

Abstract

This paper presents a method to reconstruct a complete human geometry and texture from an image of a person with only partial body observed, e.g., a torso. The core challenge arises from the occlusion: there exists no pixel to reconstruct where many existing single-view human reconstruction methods are not designed to handle such invisible parts, leading to missing data in 3D. To address this challenge, we introduce a novel coarse-to-fine human reconstruction framework. For coarse reconstruction, explicit volumetric features are learned to generate a complete human geometry with 3D convolutional neural networks conditioned by a 3D body model and the style features from visible parts. An implicit network combines the learned 3D features with the high-quality surface normals enhanced from multiviews to produce fine local details, e.g., high-frequency wrinkles. Finally, we perform progressive texture inpainting to reconstruct a complete appearance of the person in a view-consistent way, which is not possible without the reconstruction of a complete geometry. In experiments, we demonstrate that our method can reconstruct high-quality 3D humans, which is robust to occlusion.

1. Introduction

How many portrait photos in your albums have the whole body captured? Usually, the answer is not many. Taking a photo of the whole body is often limited by a number of factors of occlusion such as camera angles, objects, other people, and self. While existing single-view human reconstruction methods [3, 43] have shown promising results, they often fail to handle such incomplete images, leading to significant artifacts with distortion and missing data in 3D for invisible body parts. In this paper, we introduce a method to reconstruct a complete 3D human model from a single image of a person with occlusions as shown in Figure 1. The complete 3D model can be the foundation for a wide range of applications such as film production, video games, virtual teleportation, and 3D avatar printing from a group-shot photo. 3D human reconstruction from an image [2, 16]

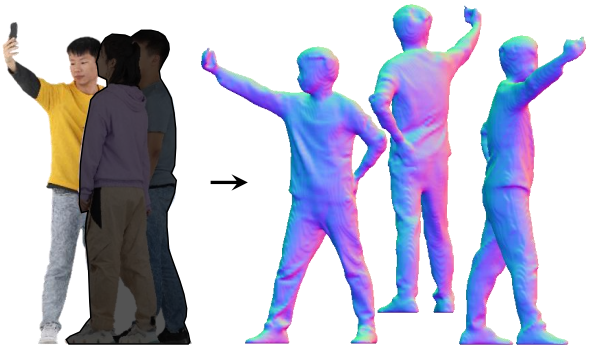

Figure 1. The complete reconstruction results using our method from the image of a person with occlusion by other people.

has been studied for two decades. The recent progress in this topic indicates the neural network based implicit approach [3, 44] is a promising way for accurate detail reconstruction. Such an approach often formulated the 3D human reconstruction problem as a classification task: an implicit network is designed to learn image features at each pixel, *e.g.*, pixel-aligned features [18, 43, 44], which enable continual classification of the position in 3D along the camera ray. While the implicit approaches have shown a strong performance to produce the geometry with high-quality local details, the learning of such an implicit model is often characterized as 1) reconstructive: it estimates 3D only for the pixels that are captured from a camera, *i.e.*, no 3D reconstruction is possible for missing pixels of invisible parts; and 2) globally incoherent: the ordinal relationship (front-back relationship in 3D) of the reconstructed 3D points is often not globally coherent, *e.g.,* while the reconstruction of the face surface is locally plausible, its combination with other parts such as torso looks highly distorted. These properties fundamentally limit the implicit network to reconstruct the complete and coherent 3D human model from the image with a partial body.

In this paper, we overcome these fundamental limitations

of the implicit network by modeling generative and globally coherent 3D volumetric features. To this end, we use a 3D convolutional neural network that can explicitly capture the global ordinal relation of a human body in the canonical 3D volume space. It generates volumetric features by encoding an incomplete image and a 3D body model, *i.e.*, SMPL [30, 37], where the 3D body model provides the unified guidance of the body pose in the coherent 3D space. These volumetric features are jointly learned with a 3D discriminator in a way that generates a coarse yet complete 3D geometry.

The complete 3D geometry enables the coherent rendering of its shape over different viewpoints, which makes it possible to enhance surface normals and inpaint textures in a multiview-consistent way. Specifically, for surface normal enhancement, a neural network takes as input a coarse rendering of the surface normal and style features; and outputs the fine surface normal with plausible high-frequency details. We design a novel normal fusion network that can combine the fine surface normals from multiviews with the learned volumetric features to upgrade the quality of local geometry details. For texture inpainting, a neural network conditioned on the fine surface normal and an incomplete image generates the complete textures. The inpainted textures are progressively combined from multiviews through the 3D geometry.

Unlike previous methods [43, 43, 44, 52] which have utilized the surface normals from limited views (*e.g.*, front and back), our multiview normal fusion approach can produce more coherent and refined reconstruction results by incorporating fine-grained surface normals from many views.

Our experiments demonstrate that our method can robustly reconstruct a complete 3D human model with plausible details from the image of a partial body, outperforming previous methods while still obtaining comparable results in the full-body case.

The technical contributions of this work include (1) a new design of generative and coherent volumetric features which make an implicit network possible to reconstruct a complete 3D human from an incomplete image; (2) a novel multiview normal fusion approach that upgrades the quality of local geometry details in a view-coherent way; and (3) an effective texture inpainting pipeline using the reconstructed 3D geometry.

2. Related Work

Monocular human reconstruction with explicit shape models. One of the primary challenges in human modeling is reconstructing accurate and high-fidelity explicit 3D surfaces. However, due to the limited availability of 3D human explicit geometry data, achieving high-quality human reconstruction with various styles remains a long-term problem. One effective approach to address this challenge is to use explicit shape models: there is a wide range of explicit 3D shape representations including voxels [21, 49], point clouds [10, 27, 40, 41, 48, 54] or parametric meshes [2, 13, 23, 30, 50, 61]. Voxel-based methods are often limited by low resolution and difficulty in predicting shape details. Point clouds, on the other hand, have advantages in achieving topological modeling, but require tedious point estimation for obtaining fine surface details. In human modeling, 3D parametric shape models play a crucial role, particularly in single-view 3D human reconstruction. Human body templates can overcome occlusion problems and avoid fundamental depth ambiguity. These approaches [23–25, 38, 39, 57, 58] can estimate SMPL [30] shapes and coefficients from a given image. However, the given 3D parametric model only provides an occlusion-free full-body geometry, lacking garment shape and style information. This results in less detailed surface reconstruction.

Pixel-aligned implicit function. The field of image-based human modeling [3, 16, 18, 19, 43, 44, 52] has enabled the generation of a wide range of human models. Utilizing implicit functions, the reconstructed mesh is independent of the volume resolution. Implicit shape functions such as [28, 33, 36, 53] can represent 3D surfaces in a continuous SDF or occupancy field, which requires dense sampling around the mesh for detailed surface reconstruction. These single-view human body reconstruction methods [18, 43], take advantage of the 2D pixel-aligned features to encode the occupancy values of each sampling point, and can reconstruct a clothed human body with rich surface detail. Despite advancements in pixel-aligned implicit functions, feature ambiguity and lack of global shape robustness still pose challenges. Additionally, when given input images are largely occluded, these functions are unable to handle full-body reconstruction. While some recent works [3, 44, 52] have attempted to adapt to higher resolution inputs or complex poses, none of them have been able to achieve partial image human body reconstruction due to the inherent limitations of pixel-aligned local features.

Existing methods often lack global consistency and heavily rely on local image features, resulting in unnatural body shapes or missing parts in occluded areas. To address this, recent works [15, 60] combine explicit 3D models, such as SMPL or voxel features, with pixel-aligned implicit functions to regularize global shape and ensure consistency. However, generating local details in occluded parts remains challenging.

2D and 3D generative model for occlusion. Recent advances in Generative Adversarial Networks (GANs) [20] and Diffusion Models [17, 17, 42, 47] have enabled high-fidelity image synthesis. Previous 2D human generative models [1, 6, 12, 26, 32, 45] have demonstrated impressive results in generating synthetic human images. However, these meth-

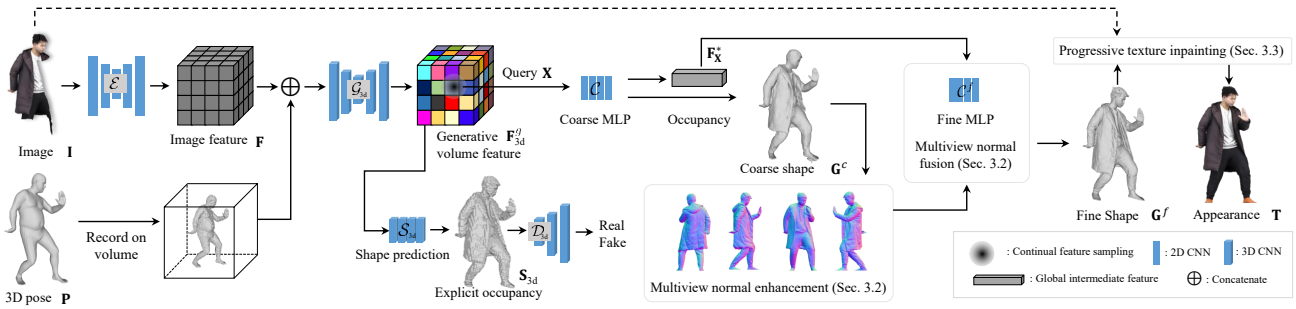

Figure 2. The overview of our approach. Given an image $\mathbf{I}$ of a person with occlusion and a guiding 3D body pose $\mathbf{P}$, we reconstruct a complete 3D human model $\mathbf{G}^f$ in a coarse-to-fine manner: we first build the volume of image features $\mathbf{F}$ by extracting the 2D image features and copying them in a depth direction. This image feature volume is concatenated with the 3D body pose $\mathbf{P}$ recorded on the volume. Our 3D CNN $\mathcal{G}_{3d}$ generates complete and coherent volumetric features whose generative power is enabled by jointly learning with 3D discriminator $\mathcal{D}_{3d}$ with explicit shape prediction $\mathbf{S}_{3d}$. The coarse MLP $\mathcal{C}$ produces the coarse yet complete occupancy of the continually sampled 3D points and their intermediate global features $\mathbf{F}^*$ where we represent the 3D surface by using 0.5 level-set occupancy field. The fine MLP $\mathbf{C}^f$ combines $\mathbf{F}^*$ and surface normals enhanced from multiviews to output fine-grained occupancy. We also complete the appearance by performing view-progressive texture inpainting.

ods are based on 2D reasoning and do not guarantee 3D consistency. In addition, current 2D human-specific inpainting models struggle with large hole completion as it is a challenging task to complete the structure in 2D. Consequently, the 3D geometry obtained from these methods often has low fidelity for the occluded regions due to erroneous inpainting results.

Recently, a 3D human generative model (gDAN) [5] has been developed to directly generate various 3D avatars in 3D space with unconditional style. Motivated by gDAN and other 3D shape generative models [7, 35, 46], we propose a coarse-to-fine 3D generative model that reconstructs human bodies from single-view incomplete images. And our generative style conditioned on the incomplete image.

3. Method

We design a novel coarse-to-fine 3D generative framework to achieve a complete 3D human body reconstruction from a single incomplete image. Figure 2 illustrates an overview of our framework. The input to our system is a single image of a person with a partial body, and we assume the unclothed 3D body mesh, *i.e.*, SMPL [30, 37] model, aligned with the image is given. We develop generative volumentic features using a 3D convolutional neural network by learning to reconstruct a coarse yet complete 3D human geometry with a 3D discriminator (Section 3.1). We further improve the high-frequency details of the coarse geometry by generating fine-detailed surface normals from multiviews and combining them through an implicit fusion network (Section 3.2). Finally, we perform view-progressive 2D appearance inpainting to obtain fully textured and coherent 3D human avatar (Section 3.3).

3.1. Learning Generative Volumetric Features

We cast the single-view 3D reconstruction problem as a binary feature classification of a 3D point:

$$\mathbf{F} = \mathcal{E}(\mathbf{I}), \quad \mathcal{C}(\mathbf{F}_{\mathbf{x}_p}; \mathbf{X}_p) \rightarrow [0, 1], \tag{1}$$

where $\mathbf{I} \in \mathbb{I}^{w \times h \times 3}$ is the image of a person with partial body, $\mathcal{E}$ is the feature extraction function often enabled by an encoder-decoder network, $\mathbf{F} \in \mathbb{R}^{w \times h \times c}$ is the 2D map of image features, $\mathcal{C}$ is an implicit classifier which classifies a continually sampled 3D point $\mathbf{X} \in \mathbb{R}^3$ into 0 (inside) and 1 (outside), so that the 3D surface can be represented as a 0.5 level-set of continuous occupancy field [31]. $\mathbf{x} \in \mathbb{R}^3$ is the 2D projection of $\mathbf{X}$, *i.e.*, $\mathbf{\Pi X} = \mathbf{x}$ where $\mathbf{\Pi}$ is the projection matrix, $p \in P$ is the index of the points set on the visible body parts. For the pixels lying on invisible body parts $\mathbf{x}_q$ where $q \in Q$ is the index of the invisible points set, C always classifies the features as outside the surface, *i.e.*, $\mathcal{C}(\mathbf{F}_{\mathbf{x}_q}; \mathbf{X}_q) = 1$, due to the missing data in the image: there exists no pixel information (*e.g.*, black patches) to encode onto the image features.

One can augment this incomplete image features by propagating the features from the visible to invisible parts with the joint learning of a 2D shape discriminator for generative adversarial training:

$$\mathcal{G}(\mathbf{F}) = \mathbf{F}^g, \quad \mathcal{S}(\mathbf{F}^g) = \mathbf{S}, \quad \mathcal{D}(\mathbf{S}) \rightarrow [0, 1], \tag{2}$$

where $\mathcal{G}$ is the generative function that generates the complete features $\mathbf{F}^g$, $\mathcal{S}$ is the function that predicts 2D binary shape mask $\mathbf{S} \in [0, 1]^{w \times h}$ (0 is background, 1 is foreground), $\mathcal{D}$ is the 2D discriminator that distinguishes the real and fake of a complete human shape. By taking advantage of a generative framework, the augmented image features allows $\mathcal{C}$ to

classify the 3D points on the invisible body parts in a way that construct a complete human, *i.e.*, $\mathcal{C}(\mathbf{F}^g_{\mathbf{X}_q}; \mathbf{X}_q) \rightarrow [0, 1]$.

While now the image features are complete, they are holding a significant pose ambiguity: any plausible body poses for invisible parts that harmonize with visible ones can be possible. We disambiguate it by further conditioning pose information:

$$\mathcal{G}(\mathbf{F}; \mathbf{P}) = \mathbf{F}^g, \quad \mathcal{S}(\mathbf{F}^g) = \mathbf{S}, \quad \mathcal{D}(\mathbf{S}; \mathbf{P}) \rightarrow [0, 1],$$

where $\mathbf{P} \in \mathbb{R}^{w \times h \times m}$ is the map of a guiding 2D body pose, *e.g.*, keypoints [4] and densepose [14]. Conditioning $\mathbf{P}$ enables the features to be aware of the global body poses, leading to shape generation without pose ambiguity.

Still, however, since the augmented features $\mathbf{F}^g$ are modeled totally from 2D space, it is not possible to capture the global ordinal relationship of a human body in 3D, *e.g.*, while the generated 3D surface of a leg looks plausible, its combination with visible torso is highly distorted. To capture such a global relationship, we propose to upgrade the entire featuring modeling pipeline from 2D to 3D:

$$\begin{aligned} \mathcal{G}_{3\text{d}}(\mathbf{F}; \mathbf{P}_{3\text{d}}) &= \mathbf{F}^g_{3\text{d}}, \qquad \mathcal{C}(\mathbf{F}^g_{3\text{d},\mathbf{X}}; \mathbf{X}) \rightarrow [0, 1], \\ \mathcal{S}_{3\text{d}}(\mathbf{F}^g_{3\text{d}}) &= \mathbf{S}_{3\text{d}}, \qquad \mathcal{D}_{3\text{d}}(\mathbf{S}_{3\text{d}}; \mathbf{P}_{3\text{d}}) \rightarrow [0, 1], \end{aligned} \tag{3}$$

where $\mathbf{S}_{3\text{d}}$, $\mathbf{P}_{3\text{d}}$, and $\mathbf{F}^g_{3\text{d}}$ are defined in the canoncial volume space. The generation of the volumetric features $\mathbf{F}^g_{3\text{d}}$ allows C to reconstruct the globally coherent and complete 3D human geometry. $\mathbf{F}^g_{3\text{d}}$ is learned by minimizing the following objectives:

$$L_{\text{feat}} = \mathcal{L}_{\text{c}} + \lambda_{\text{g}}\mathcal{L}_{\text{g}} + \lambda_{\text{cGAN}}\mathcal{L}_{\text{cGAN}}, \tag{4}$$

where λ balances the contribution of each loss. $\mathcal{L}_{\text{c}}$ makes a direct supervision on the implicit classifier:

$$\mathcal{L}_{\text{c}} = \sum_i \|\mathcal{C}(\mathbf{F}^g_{3\text{d},\mathbf{X}}; \mathbf{X}) - \mathcal{C}_{\text{gt}}(\mathbf{X})\|^2, \tag{5}$$

where $\mathcal{C}_{\text{gt}} : \mathbb{R}^3 \rightarrow \{0, 1\}$ outputs ground-truth label of the 3D occupancy. $\mathcal{L}_{\text{g}}$ supervise the 3D shape prediction by comparing with ground truth volume, $\mathcal{L}_g = \sum \|\mathbf{S}_{3\text{d}} - \mathbf{S}_{3\text{d,gt}}\|$. $\mathcal{L}_{\text{cGAN}}$ is the conditional adversarial loss [20] where we use $\{\mathbf{S}_{3\text{d}}, \mathbf{P}_{3\text{d}}\}$ for fake $\{\mathbf{S}_{3\text{d,gt}}, \mathbf{P}_{3d}\}$ for real inputs.

Implementation details. We enable the features extraction function with a 2D convolutional neural network (*e.g.*, U-net [34]) which takes as an input image $\mathbf{I}$ and produces pixel-aligned features $\mathbf{F}$. We use a 3D convolutional neural network (*e.g.*, 3D U-net [9]) to design $\mathcal{G}_{3\text{d}}$ that generates 3D volumetric features $\mathbf{F}^g_{3\text{d}}$ from a 3D body pose $\mathbf{P}$ and $\mathbf{F}$. In practice, to build the input volumes for $\mathcal{G}_{3\text{d}}$, we discretize the vertices of the posed SMPL body model and record them on a canonical volume ($128 \times 128 \times 128$); $\mathbf{F}$ is copied over the three-dimensional direction; and the two volumes for $\mathbf{P}$ and $\mathbf{F}$ are concatenated.

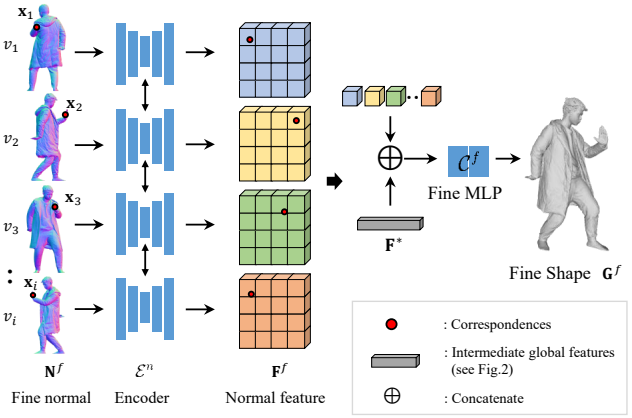

Figure 3. The overview of our multiview surface normal fusion.

The volumetric features are decoded in two ways: explicit and implicit. For explicit decoding $\mathcal{S}_{3\text{d}}$, 3D convolutional networks reconstruct a complete occupancy $\mathbf{S}_{3\text{d}}$ at each voxel grid, whose geometric distribution are classified by a 3D discriminator $\mathcal{D}_{3\text{d}}$ [51]. For implicit decoding $\mathcal{C}$, we utilize multilayer perceptron (MLP) to classify the learned volumetric features of a 3D query point $\mathbf{X}$ (which is the resultant of dynamic sampling around the ground truth mesh), where we perform trilinear interpolation of the volumeric features that are neighboring the query point to construct the continuous features representation. Inspired by existing multi-level MLP processing [5], we further design $\mathcal{C}$ in a way that produces not only occupancy but also its intermediate feature representation as shown in Figure 2:

$$\mathcal{C}(\mathbf{F}^g_{3\text{d},\mathbf{X}}; \mathbf{X}) \rightarrow \{\ [0, 1], \ \ \mathbf{F}^*_{\mathbf{X}}\ \} \tag{6}$$

where $\mathbf{F}^*_{\mathbf{X}} \in \mathbb{R}^{256}$ is the intermediate feature that captures the structure and visibility of the 3D point in the context of the global body pose. We show the detailed network structure in the supplementary material.

3.2. Multiview Surface Normal Fusion

We improve the quality of local geometric details of the coarse reconstruction from Section 3.1 by combining fine-detailed surface normals:

$$\mathbf{F}^n = \mathcal{E}^n(\mathbf{N}^f), \quad \mathcal{C}^f(\mathbf{F}^n_{\mathbf{X}}; \mathbf{F}^*_{\mathbf{X}}, \mathbf{X}) \rightarrow [0, 1], \tag{7}$$

where $\mathbf{N}^f$ is the surface normal map with high-frequency details, $\mathcal{E}^n$ is a surface normal encoder that produces pixel-aligned normal features, $\mathcal{C}^f$ is the fine classifier that classifies the in/out occupancy status of the 3D point $\mathbf{X}$, and $\mathbf{F}^*$ is the intermediate features of the coarse classifier, *i.e.*, $\mathcal{C}$ (see implementation details in Section 3.1).

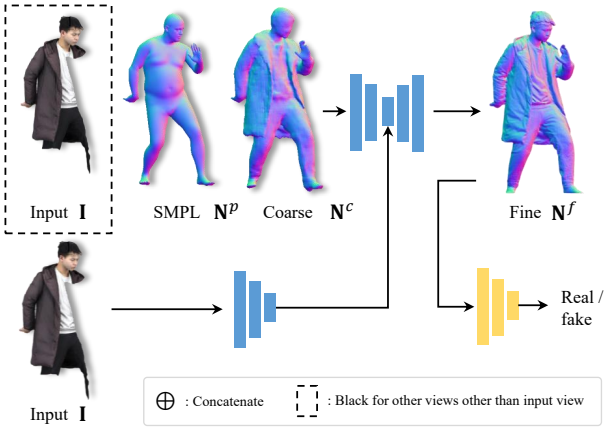

Figure 4. The details of our surface normal enhancement network.

To obtain $\mathbf{N}^f$, existing methods (*e.g.*, [44]) have often utilized a human surface normal detection from an image. However, for the single input image with occlusion, $\mathbf{N}^f$ is missing two elements, 1) body parts: there exists no pixel to detect, and 2) viewpoints: only single-view input is available, so thus, the surface normal from other views is unknown. Those missing data prevent $\mathcal{C}^f$ from performing fine-grained occupancy reconstruction for the invisible parts. For these reasons, we reformulate the surface normal detection problem as generation:

$$\mathcal{R}(\mathbf{G}^c; v_i) = \mathbf{N}^c_{v_i}, \quad \mathbf{N}^f_{v_i} = \mathcal{G}^n(\mathbf{N}^c_{v_i}; \mathbf{I}), \tag{8}$$

where $\mathcal{R}$ is the function that renders the surface normal $\mathbf{N}^c_{v_i}$ from the coarse geometry $\mathbf{G}^c \in \mathbb{R}^{n\times 3}$ (obtained from $\mathcal{C}$ in Section 3.1) and a specific viewpoint v_i, $\mathcal{G}$ is the generation function that generates high-frequency normal details from $\mathbf{N}^c_{v_i}$. The input partial image $\mathbf{I}$ is used to guide the appearance style of the person in the latent space.

Importantly, our coarse geometry $\mathbf{G}^c$ is complete, and therefore, rendering the coarse surface normal from any view is always possible. This allows us to combine the features of fine surface normals from multiviews:

$$\mathcal{C}^f(\{\mathbf{F}^n_{v_1,\mathbf{x}_1}, .., \mathbf{F}^n_{v_i,\mathbf{x}_i}\}; \mathbf{F}^*_{\mathbf{X}}) \rightarrow [0, 1]; \tag{9}$$

where $\mathbf{F}^f$ is the outcome of the feature extraction (Eq. 7) i is the number of views and we use $i = 4$ in practice (front, back, right, and left).

We enable $\mathcal{E}^n$ and $\mathcal{C}^f$ using multiview fusion networks and $\mathcal{G}^n$ using normal enhancement networks whose details and training objectives are in below.

Multiview Surface Normal Fusion Network Figure 3 shows the overall framework for our multiview normal fusion pipeline. An encoder-decoder network $\mathcal{E}^n$ extracts the pixel-aligned features from the fine surface normal $\mathbf{N}^f$ at each view. We enable the surface normal fusion function $\mathcal{C}^f$ using multilayer perceptron (MLP). For each dynamically sampled 3D point $\mathbf{X}$, it takes as input surface normal features from multiviews and global intermediate features $\mathbf{F}^*_{\mathbf{X}}$, and outputs fine-grained occupancy where $\mathbf{F}^*_{\mathbf{X}}$ is from coarse MLP as shown in Figure 3, which captures image features and viewpoints in the context of global geometry. We reconstruct the fine geometry $\mathbf{G}^f$ by applying 0.5 level-set marching cube algorithm. $\mathcal{E}^n$ and $\mathcal{C}^f$ are trained by minimizing the following loss:

$$\mathcal{L}_{\text{fusion}} = \sum_i \|\mathcal{C}^f(\{\mathbf{F}^n_{v_1,\mathbf{x}_1}, .., \mathbf{F}^n_{v_i,\mathbf{x}_i}\}; \mathbf{F}^*_{\mathbf{X}}) - \mathcal{C}_{\text{gt}}(\mathbf{X})\|^2.$$

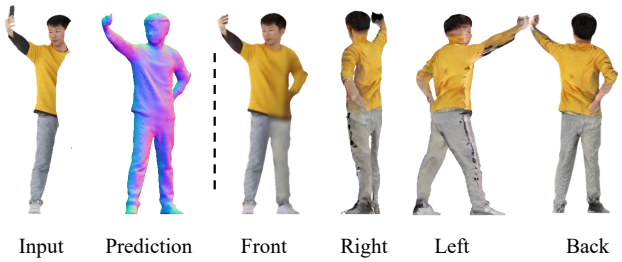

Figure 5. Results of our view-progressive texture inpainting pipeline (right side) which are effectively combined through the complete geometry predicted from our method. We use an existing human inpainting model [55] with minor modification.

Surface Normal Enhancement Network Figure 4 describes the overall framework for our surface normal enhancement network. This enables $\mathcal{G}^n$. In practice, it takes as input a coarse surface normal $\mathbf{N}^c$, the surface normal of a 3D body model $\mathbf{N}^p$, and the input image $\mathbf{I}$. $\mathbf{N}^p$ guides the global human pose, and an encoder encodes $\mathbf{I}$ to extract style features from latent space. Only for the input view, we concatenate $\mathbf{I}$ (otherwise, black image) with other surface normal maps $\{\mathbf{N}^c, \mathbf{N}^p\}$ to allow the network G to preserve the local patterns from visible texture. $\mathcal{G}^n$ is trained by minimizing the following objectives:

$$L_{\text{enhance}} = \mathcal{L}_1 + \lambda_{\text{vgg}}\mathcal{L}_{\text{vgg}} + \lambda_{\text{Adv}}\mathcal{L}_{\text{Adv}}, \tag{10}$$

where λ controls the weight of each loss. $\mathcal{L}_1$ measure the difference between the prediction $\mathbf{N}^f$ and ground truth $\mathbf{N}^f_{\text{gt}}$: $\mathcal{L}_1 = \|\mathbf{N}^f - \mathbf{N}^f_{\text{gt}}\|$ where we render $\mathbf{N}^f_{\text{gt}}$ from the ground truth geometry. $\mathcal{L}_{\text{vgg}}$ is designed to penalize the difference of $\mathbf{N}^f_{\text{gt}}$ and $\mathbf{N}^f$ from their VGG features space [22] to capture both high-frequency details and semantic validity. λ_{Adv} is the unconditional adversarial loss [11] to evaluate the plausibility of the surface normal where we use $\mathbf{N}^f_{\text{gt}}$ as real and $\mathbf{N}^f$ as fake, and we apply a patch discriminator [20].

3.3. View-Progressive Texture Inpainting

Given a complete geometry and partial input image, we

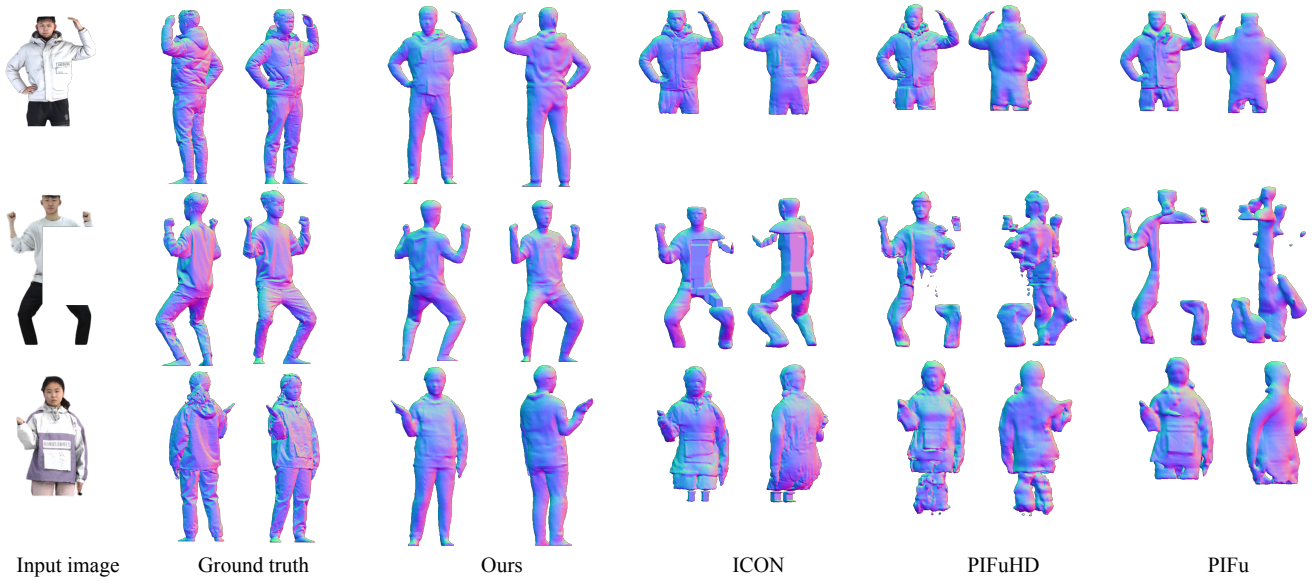

Figure 6. **Comparisons to the single view human body reconstruction models.** SOTA methods such as PIFu, PIFuHD, and ICON currently face challenges in reconstructing full body models from incomplete images, especially under challenging conditions such as large occlusion or half-body images. We evaluate our model with other frameworks on the unseen objects from Thuman2.0 [56] dataset. The first column displays partial body images randomly cropped from the original ones. In the second column, we show the ground truth geometry. The third column shows the output of our reconstruction, which ensures that the whole body is completed from the occluded image.

Full Body Image Reconstruction Evaluation				
Method	SMPL	Chamfer ↓	P2S ↓	Normal ↑
PIFu	✗	3.268	3.320	10.407
PIFuHD	✗	2.890	2.631	11.567
ICON	✓	0.965	0.848	**12.596**
Ours	✓	**0.798**	**0.808**	12.441

Table 1. **Comparison with SOTA on Human Modeling**.

generate the full texture of human by synthesizing the image of a complete human from many viewpoints in a progressive way: we iterate the surface rendering, texture inpainting, and 3D warping to other views. By starting from the input view, for each view, we render the fine surface normal using the reconstructed 3D geometry from our method (in Section 3.1-3.2). A human inpainting network generates a complete human image by taking as input a partial image and the surface normal (as shape guidance). We warp the generated texture to other views that are close to the current one through the 3D geometry by combining the textures in 3D and projecting them to other views. This allows us to render a partial body image from other views in a geometrically plausible way. We iterate these three steps to obtain a full texture in 3D as shown in Figure 5. For the inpainting model, we adopt an existing human inpainting network [55] with minor modifications. Additional details and results are available in the supplementary materials.

4. Experiment

We validate the performance of our coarse-to-fine framework quantitatively and qualitatively for the task of a complete 3D human reconstruction from a single image of a partial body.

Training details. During the training process, we train our model on our partial body images rendered from a human body dataset [56], and use Adam optimization with an initial learning rate lr = 0.0005. For the coarse model, we set the parameters in Eq. 4 with $\lambda_g = 1$ and $\lambda_{cGAN} = 0.01$, and in Eq. 10 with $\lambda_{\text{vgg}} = 1$ and $\lambda_{Adv} = 0.01$. We show the detailed network structure and parameters in the supplementary material.

Datasets. We use Thuman2.0 data [56] for both training and testing, which includes high-resolution photogrammetry scans as well as fitted SMPL mesh. We use 400 subjects for training and 20 subjects for evaluation. We create input images by performing weak perspective rendering of the 3D scan from 180 multiple viewpoints. We synthesize partial body images by masking the original images with random holes parameterized by occlusion ratio. We also use MultiThuman dataset [59] for testing as cross-dataset validation. This dataset includes the case with natural occlusion by objects and people and provides 3D surface ground truth and fitted SMPL for each person. For in-the-wild testing, we use

Figure 7. **In-the-wild testing results.** We present our reconstruction results, which demonstrate the accuracy of our approach in preserving local details while generating a complete model from an incomplete image. Our approach can generate highly-realistic human models from both incomplete and full-body images, even under challenging conditions such as large occlusions or half-body images in real-world settings.

an internet photo where we obtain the 3D body model by applying existing fitting method [57, 58].

Baseline. We compare our method with recent single-view human body reconstruction works: PIFu [43], PIFuHD [44] and ICON [52] where all their methods are based on implicit models. ICON uses parametric 3D models (SMPL) during training and inference. For the fair comparison, we retrained the baseline methods using the same dataset we used under the same experimental setting. We use ground truth SMPL during the comparison for ours and ICON. One effective approach for obtaining a complete human model is to perform 2D inpainting followed by 3D reconstruction. However, when dealing with larger holes in the image, 2D inpainting methods often struggle to produce realistic human structures, leading to artifacts such as distortion that can affect the final reconstruction results. In the supplementary materials, we provide a comparison of 2D inpainting-to-3D reconstruction results to further demonstrate this issue.

Metrics. We measure the reconstruction quality through three metrics: Chamfer, P2S, and surface normal errors. For Chamfer, we measure the bi-directional point-to-surface distances between the reconstruction and ground truth. For P2S, we measure the closest distance from the ground truth to the reconstruction with uniform sampling. For surface normal errors, we measure the distance between the rendered surface normal and ground truth from four views (one input and three synthetic views) in a PNSR space. For Chamfer and P2S, the lower score means better, while for normal error, the opposite.

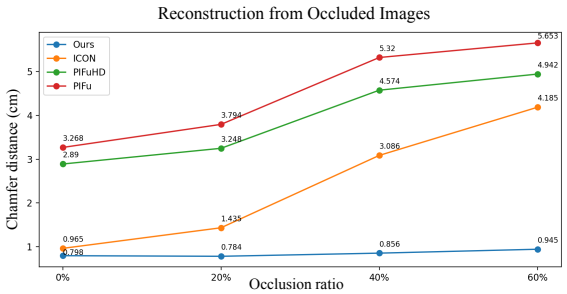

Figure 8. **A cumulative occlusion-to-reconstruction test.** This figure shows different models and their performance to reconstruct the human body from the occluded image. For a given input whole-body image, we first test the whole-body reconstruction with our method and other baselines, which is 0% occlusion. We then generate holes within the human body bounding box, and the generated holes will cover 20% 40% and 60% of the human body area. We then test the occluded images with different baseline models. When the occlusion area increases, our model is able to have a more robust reconstruction capability than other models.

Results. We summarize the quantitative comparison in Table 1 for the full-body testing cases and Figure 8 for the

Method	Chamfer↓	P2S↓	Normal↑
Ours - coarse MLP - fine MLP	1.978	1.720	6.320
Ours - fine MLP	0.818	0.926	10.704
Ours w/o GT SMPL	1.224	1.062	12.106
Ours	**0.798**	**0.808**	**12.441**

Table 2. **Ablation study results.** We show the average values of Chamfer distance, P2S, and normal PSNR over testing subjects.

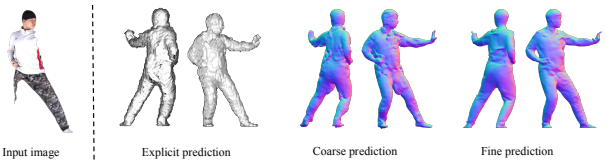

Figure 9. **Ablation study on explicit, coarse and fine model.**

partial body. The qualitative results are shown in Figure 6. For the cross-dataset validation on MultiHuman-Dataset [59], please refer to the supplementary document. Please also refer to the supplementary material for our texture inpainting results.

From Table 1, we can see that our method shows a strong performance even for the full-body testing case: our volumetric features allow our implicit model to reconstruct a globally coherent 3D human reconstruction, leading to the best quality in terms of Chamfer and P2S. While our method shows the second best under Normal metrics, it is still comparable to the best one (ICON). It implies that our globally coherent volumetric features slightly sacrifice the local details.

For the occlusion scenario shown in Figure 6, we can see that our method achieved high-quality 3D human body reconstruction from a partial body image, while other methods are struggling to handle the occlusions. Based on the graph in Figure 8, the performance gap between our method and others is largely magnified as the occlusion ratio increases.

We evaluate the performance of our model in real-world scenarios using the DeepFashion dataset [29]. Fig. 7 presents the quantitative results of our in-the-wild testing where our model is capable of effectively handling occlusions.

Ablation Study We conduct an ablation study on our coarse-to-fine human reconstruction framework to analyze the effect of each module. We study the following model combinations: (1) **Ours - corase MLP - fine MLP**: only an explicit model is only trained with a 3D convolutional neural network whose prediction result is explicit volumetric occupancy with $128 \times 128 \times 128$ voxel resolution (due to the limit of GPU resources). We use occupancy volumes as supervision. (2) **Ours - fine MLP**: we combine the explicit volume representation with coarse MLP. (3) **Ours**: this is our final model that combines explicit volume with both coarse and fine MLP with multiview surface normal enhancement as shown in Figure 2. (4) **Ours w/o GT SMPL**: to verify the effect of the accuracy of global 3D pose prior, we replace the ground truth 3D SMPL model to the fitted 3D SMPL from existing single-view prediction methods [57, 58].

Table 2 shows the summary of the performance of our ablation study. The explicit model ensures the general contour of the reconstructed mesh, but the quality of its local details is highly limited by the voxel resolution, bringing out significant discretized artifacts as shown in Figure 9. Combining a MLP with explicit volumes, *i.e.*, *Ours - fine MLP*, somewhat addresses this discretization issue by ensuring continual point sampling, but the limited resolution nature of volume features still prevents the coarse MLP from producing high-frequency details. The comparison of our approach and other ablation baselines demonstrates that combining the multiview fine surface normals is highly effective to upgrade the high-frequency details of the local 3D model surface as shown in Figure 9. Finally, *Ours w/o GT SMPL* implies that inaccurate 3D model fitting propagates its errors to our 3D reconstruction results.

Application Our method can also enable complete 3D reconstruction of people in a group-shot image. Please refer to the supplementary materials for more details and examples.

5. Conclusion

We present a method to reconstruct a complete human 3D model from a single image of a person with a partial body. To address the core occlusion problem, we introduce a new design of a coarse-to-fine human reconstruction framework. We learn generative and globally coherent volumetric features to reconstruct a coarse yet complete 3D human geometry using 3D generative adversarial networks. An implicit fusion network upgrades the quality of local geometry by combining the learned volumetric features and fine-grained multiview surface normals enhanced from coarse geometry. The evaluation on diverse subjects with various testing setups demonstrates that our framework performs well on the scenes with occlusion, showing a significant improvement over existing methods. We also show that the complete and high-quality geometry from our method makes it possible to reconstruct fully textured 3D human appearance by applying an existing inpainting model in a view-progressive way.

Limitation The requirement of an accurate 3D body model for our method is the main limitation. While it is possible to predict a 3D body model from a partial body image [8], the 3D pose prediction errors affect the global structure of our 3D reconstruction results. Our models sometimes face domain gap problems when tested on the image of a person with highly fashion styles, particularly loose clothing and complex hairstyles.

References

[1] Badour Albahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Transactions on Graphics (TOG)*, 40(6):1–11, 2021. 2

[2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2293–2303, 2019. 1, 2

[3] 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. 1, 2

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 4

[5] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. gdna: Towards generative detailed neural avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20427–20437, 2022. 3, 4

[6] Xu Chen, Jie Song, and Otmar Hilliges. Unpaired pose guided human image generation. In *Conference on Computer Vision and Pattern Recognition (CVPR 2019)*. Computer Vision Foundation (CVF), 2019. 2

[7] 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. 3

[8] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1475–1484, 2022. 8

[9] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, 2016. 4

[10] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 2

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 5

[12] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. Stylepeople: A generative model of fullbody human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5151–5160, 2021. 2

[13] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018. 2

[14] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 4

[15] 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

[16] 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. 1, 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[18] Yang Hong, Juyong Zhang, Boyi Jiang, Yudong Guo, Ligang Liu, and Hujun Bao. Stereopifu: Depth aware clothed human digitization via stereo vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 535–545, 2021. 1, 2

[19] 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

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2, 4, 5

[21] Aaron S Jackson, Chris Manafas, and Georgios Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[23] 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 conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2

[24] 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

[25] 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

[26] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 853–862, 2017. 2

[27] Chen-Hsuan Lin, Chen Kong, and Simon Lucey. Learning efficient point cloud generation for dense 3d object reconstruction. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[28] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2019–2028, 2020. 2

[29] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 8

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

[31] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 3

[32] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in neural information processing systems*, 30, 2017. 2

[33] 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

[34] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 4

[35] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 3

[36] 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

[37] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA 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 10975–10985, 2019. 2, 3

[38] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 803–812, 2019. 2

[39] 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 conference on computer vision and pattern recognition*, pages 459–468, 2018. 2

[40] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2

[41] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and BjÃ¶rn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[43] 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, 7

[44] 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. 1, 2, 5, 7

[45] Kripasindhu Sarkar, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Humangan: A generative model of human images. In *2021 International Conference on 3D Vision (3DV)*, pages 258–267. IEEE, 2021. 2

[46] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 3

[47] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[48] Neslisah Torosdagli, Denise K Liberton, Payal Verma, Murat Sincan, Janice S Lee, and Ulas Bagci. Deep geodesic learning for segmentation and anatomical landmarking. *IEEE transactions on medical imaging*, 38(4):919–931, 2018. 2

[49] 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 (ECCV)*, pages 20–36, 2018. 2

[50] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 2

[51] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016. 4

[52] 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 13296–13306, 2022. 2, 7

[53] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in neural information processing systems*, 32, 2019. 2

[54] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4541–4550, 2019. 2

[55] Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-guided human animation from a single image in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15039–15048, 2021. 5, 6

[56] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, June 2021. 6

[57] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *arXiv preprint arXiv:2207.06400*, 2022. 2, 7, 8

[58] 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, 7, 8

[59] 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 *IEEE Conference on Computer Vision (ICCV 2021)*, 2021. 6, 8

[60] 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. 2

[61] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2