Complete 3D Human Reconstruction from a Single Incomplete Image

Junying Wang\textsuperscript{1} \quad Jae Shin Yoon\textsuperscript{2} \quad Tuanfeng Y. Wang\textsuperscript{2} \quad Krishna Kumar Singh\textsuperscript{2} \quad Ulrich Neumann\textsuperscript{1}

\textsuperscript{1}University of Southern California \quad \textsuperscript{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 \cite{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 \cite{2, 16} has been studied for two decades. The recent progress in this topic indicates the neural network based implicit approach \cite{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 \cite{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, \textit{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, \textit{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-
Multiview normal enhancement (Sec. 3.2) 

Record on volume

3D pose

Shape prediction

Explicit occupancy

3D body pose

Coarse MLP

Multiview normal enhancement (Sec. 3.2)

Image feature

Continual feature sampling

Global intermediate feature

2D CNN

3D CNN

Coarse shape

Progressive texture inpainting (Sec. 3.3)

Fine Shape

G_f

Appearance T

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

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 volumetric 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:

\[ F = \mathcal{E}(I), \quad \mathcal{C}(F_{x_q}; X_p) \to [0, 1], \quad (1) \]

where \( I \in \mathbb{R}^{w \times h \times 3} \) is the image of a person with partial body, \( \mathcal{E} \) is the feature extraction function often enabled by an encoder-decoder network, \( F \in \mathbb{R}^{w \times h \times c} \) is the 2D map of image features, \( \mathcal{C} \) is an implicit classifier which classifies a continually sampled 3D point \( X \in \mathbb{R}^3 \) 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]. \( x \in \mathbb{R}^3 \) is the 2D projection of \( X \), i.e., \( \Pi X = x \) where \( \Pi \) 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 \( x_q \) where \( q \in Q \) is the index of the invisible points set, \( C \) always classifies the features as outside the surface, i.e., \( \mathcal{C}(F_{x_q}; X_p) = 1 \), 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:

\[ G(F) = F^g, \quad S(F^g) = S, \quad D(S) \to [0, 1], \quad (2) \]

where \( G \) is the generative function that generates the complete features \( F^g \), \( S \) is the function that predicts 2D binary shape mask \( S \in [0, 1]^{w \times h} \) (0 is background, 1 is foreground), \( D \) 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 \( C \) to
classify the 3D points on the invisible body parts in a way that construct a complete human, i.e., $C(F_{3d}^g; X) \rightarrow [0, 1]$.

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:

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

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

Still, however, since the augmented features $F^g$ 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 volumetric features on a canonical volume (see Fig.2), to build the input volumes for $G$ which takes as an input image $I$ and produces pixel-aligned features $F$. We use a 3D convolutional neural network (e.g., 3D U-net [9]) to design $G_{3d}$ that generates 3D volumetric features $F_{3d}^g$ from a 3D body pose $P$ and $F$. In practice, to build the input volumes for $G_{3d}$, we discretize the vertices of the posed SMPL body model and record them on a canonical volume ($128 \times 128 \times 128$); $F$ is copied over the three-dimensional direction; and the two volumes for $P$ and $F$ are concatenated.

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

$$C(F_{3d}^g; X) \rightarrow \{ [0, 1], F_X \}$$

where $F_X \in \mathbb{R}^{256}$ 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:

$$F^n = E^n(N^f), \quad C^f(F^n; F_X; X) \rightarrow [0, 1],$$

where $N^f$ is the surface normal map with high-frequency details, $E^n$ is a surface normal encoder that produces pixel-aligned normal features, $C^f$ is the fine classifier that classifies the in/out occupancy status of the 3D point $X$, and $F^*$ is the intermediate features of the coarse classifier, i.e., $C$ (see implementation details in Section 3.1).
To obtain $N_f$, existing methods (e.g., [44]) have often utilized a human surface normal detection from an image. However, for the single input image with occlusion, $N_f$ 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 $C_f$ from performing fine-grained occupancy reconstruction for the invisible parts. For these reasons, we reformulate the surface normal detection problem as generation:

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

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

Importantly, our coarse geometry $G^c$ 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:

$$C^f([F^a_{v_i, x_1}, \ldots, F^a_{v_i, x_4}]; F^c_X) \rightarrow [0, 1];$$

where $F^f$ 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 $E^n$ and $C^f$ using multiview fusion networks and $G^n$ 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 fusion pipeline. An encoder-decoder network $E^n$ extracts the pixel-aligned features from the fine surface normal $N^f$ at each view. We enable the surface normal fusion function $C^f$ using multilayer perceptron (MLP). For each dynamically sampled 3D point $X$, it takes as input surface normal features from multiviews and global intermediate features $F^c_X$, and outputs fine-grained occupancy where $F^h_X$ 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 $G^f$ by applying 0.5 level-set marching cube algorithm. $E^n$ and $C^f$ are trained by minimizing the following loss:

$$L_{fusion} = \sum_i \|C^f([F^a_{v_i, x_1}, \ldots, F^a_{v_i, x_4}]; F^c_X) - C_{gt}(X)\|^2.$$

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

$$L_{enhance} = L_1 + \lambda_{vgg}L_{vgg} + \lambda_{Adv}L_{Adv},$$

where $\lambda$ controls the weight of each loss. $L_1$ measure the difference between the prediction $N^f$ and ground truth $N^f_{gt}$:

$$L_1 = \|N^f - N^f_{gt}\|$$

where we render $N^f_{gt}$ from the ground truth geometry. $L_{vgg}$ is designed to penalize the difference of $N^f_{gt}$ and $N^f$ from their VGG features space to capture both high-frequency details and semantic validity. $\lambda_{Adv}$ is the unconditional adversarial loss [11] to evaluate the plausibility of the surface normal where we use $N^f_{gt}$ as real and $N^f$ as fake, and we apply a patch discriminator [20].

**3.3. View-Progressive Texture Inpainting**

Given a complete geometry and partial input image, we
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 $\lambda_{vgg} = 1$ 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 Multi-Thuman 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.

**Results.** We summarize the quantitative comparison in Table 1 for the full-body testing cases and Figure 8 for the...
We show the average values of Chamfer distance, P2S, and normal PSNR over testing subjects.

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Chamfer↓</th>
<th>P2S↓</th>
<th>Normal↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours - coarse MLP - fine MLP</td>
<td>1.978</td>
<td>1.720</td>
<td>6.320</td>
</tr>
<tr>
<td>Ours - fine MLP</td>
<td>0.818</td>
<td>0.926</td>
<td>10.704</td>
</tr>
<tr>
<td>Ours w/o GT SMPL</td>
<td>1.224</td>
<td>1.062</td>
<td>12.106</td>
</tr>
<tr>
<td>Ours</td>
<td>0.798</td>
<td>0.808</td>
<td>12.441</td>
</tr>
</tbody>
</table>

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 model 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 - coarse 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


