# Modeling Detailed Human Geometry with Adaptive Local Refinement

Bang Du Kunyao Chen Haochen Zhang
Fei Yin Baichuan Wu Truong Nguyen
University of California San Diego
{b7du,kuc017,haz035,fyin,bwu,tqn001}@ucsd.edu

## Abstract

*Estimating clothed human body shapes from monocular images has been a difficult problem due to occlusions, varying poses, and diverse clothing styles. Current methods involve directly regressing for either 3D positions of primitives or values in a volumetric space, but they struggle to balance generalization and accuracy, leading to suboptimal results. In this paper, we introduce a novel two-step framework that efficiently combines 2D and 3D representations to achieve both accurate surface detail inference and strong generalization capabilities: addressing challenging poses by occlusions and varying clothing styles. Our approach first uses an image-to-image translation framework to estimate a rough shape, which serves as an initial approximation of the human body. This step effectively captures global structure and coarse details, while being computationally efficient. Next, we employ a dedicated refinement module to enhance the surface details for a high-fidelity result. It utilizes an attention-based strategy that allows the 3D refinement module to focus on regions of interest, such as areas with complex clothing or occlusions. This strategy effectively improves the overall quality of the inferred shape by generating high-density patches of points in challenging regions. Our experiments show that, with the attention-based strategy, the proposed method outperforms state-of-the-art methods in terms of both qualitative and quantitative measures, demonstrating its effectiveness in handling diverse clothing styles and poses.*

## 1. Introduction

Estimating human body shapes from monocular images has become an increasingly important research topic in computer vision and graphics, with a wide range of applications such as virtual reality, gaming, fashion industry, and human-computer interaction. However, inferring clothed human body shapes poses significant challenges due to factors like occlusions, varying poses, and the diversity of clothing styles. It is crucial to design a simple acquisition system of high-quality human models with rich surface details. Traditional 3D-scan [27, 38] and multi-view reconstruction methods [14, 16, 23] rely heavily on hardware to capture dense inputs of a target scene, limiting its application in daily scenarios.

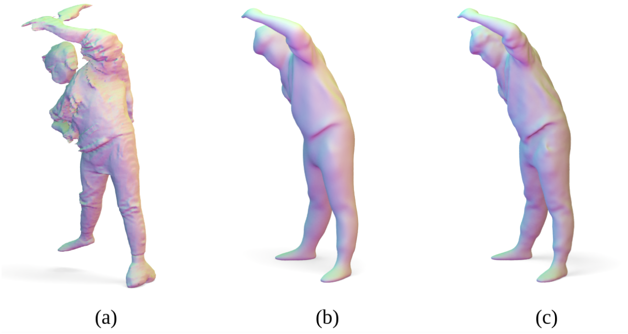


Figure 1. Illustration of reconstructed models under an uncommon pose from (a) a state-of-the-art implicit function-based method. (b) our coarse prediction. (c) our refined prediction.

Recent advancement in deep learning techniques makes single-view reconstruction possible. Some initial works [18, 32] adopt point cloud representation and regress on point positions with multilayer perceptron (MLP). Although follow-up works [33, 44–46] improve network structures by leveraging advanced 3D convolution and graph convolution (GCN) to learn with the correlation between neighboring points in a more elegant way, these approaches usually work with simple CAD models with fewer vertices. It's memory inefficient and computationally expensive to directly apply them to complex human models.

Another line of research [10, 47] seeks to establish shape priors through the utilization of volumetric representations, an approach that, while insightful, imposes significant computational demands. Recently, training neural networks as implicit functions gain popularity in 3D reconstruction and view synthesis. As demonstrated by works such as [36, 37, 40], this strategy achieves impressive results and, in the meantime, reduces memory complexity and

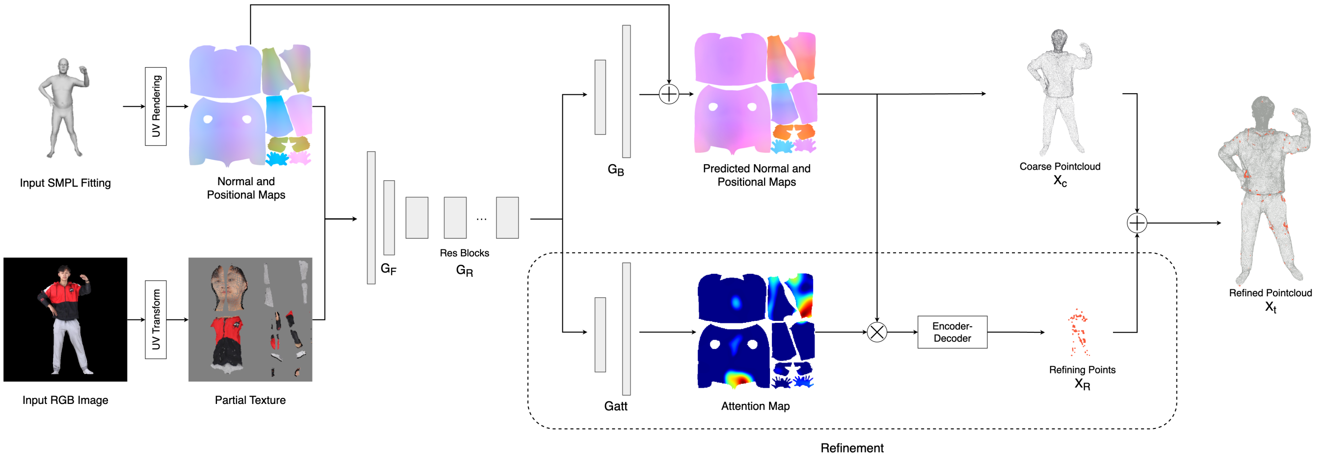


Figure 2. **Pipeline Overview**: The proposed framework utilizes unclothed SMPL-derived positional and normal maps in UV space, and an RGB image-derived partial UV texture as inputs. A generator uses these inputs to predict the coarse pointcloud $X_c$. To enhance the local expressiveness, we adopt an attention head $G_{att}$ to estimate regions of interest in the UV space as an attention map. The attention map is then multiplied with the predicted UV maps and decoded to local refinement points using the encoder-decoder structure. These points are then articulated with coarse points to generate a refined clothed human pointcloud. Notably, the generation of positional maps does not rely on SMPL template priors, allowing for the expression of arbitrary topology in the resulting pointclouds.

increases spatial resolution compared to classic volumetric-based frameworks. However, most of these methods' performance is subject to the variation of data. Although some works [42, 43] are proposed to mitigate the generalization issue by conditioning on the input pixels, the results are still suboptimal (See Figure 1). Recent works [48, 49] further improve the approach by integrating normal predictions into the reconstruction of clothed bodies. However, these methods require separate estimations of images from several viewpoints for the best performance.

Instead of utilizing features in projected space, there are many approaches that train in a special unwrapped space, i.e. UV parameterization space. The advantage is twofold. Firstly, models in the UV space partially preserve neighborhood information of vertices. Moreover, the model is arranged in a regularized form such that it is compatible with CNN frameworks. [15, 21] use a collection of UV patches and global embedding features to learn the shape prior, but the results are typically over-smoothed and lack details. For human models, Tex2Shape [4] transforms the shape estimation problem into an image-to-image translation problem, taking advantage of the power of 2D deep neural networks. Although simple and robust, the result is not comparable to the aforementioned implicit-function-based methods. Moreover, it relies on parametric templates for surface displacement, causing the resulting shape restricted by the template's characteristics, such as topology and resolutions. Recent work [35] achieves very promising results. It is based on the UV parameterization of a minimally-clothed parametric model [34] and focuses on training the residuals i.e. fine-grained garments and surface details. However, since the method is designed to reconstruct the cloth variance for a particular person, its ability to generalize across diverse human inputs remains a subject of further investigation. In addition, it adopts a structure that treats each surface element equally, ignoring the high variety of local surface structures. For human models with both wrinkles from clothes and smooth exteriors from the skin, elevating the inferring resolution everywhere increases the number of parameters exponentially, which not only leads to an increase of network parameters but also makes the representation less efficient.

In this paper, we introduce a novel two-step framework that generates fine-grained human geometry using a mixture of UV space training and point cloud-based regression. We first feed the RGB partial texture, the positional map, and the normal map of the corresponding SMPL model into a Tex2Shape-like image-to-image translation network. Through the employment of normalized coordinates and a residual learning strategy, we anticipate robust results across a broad spectrum of poses and 2D appearances in this initial stage. Subsequently, we improve the details with a simple-yet-effective refinement module for a high-expressive outcome. We surprisingly find that applying an explicit attention-based strategy in the framework to discover regions that need higher resolutions can significantly facilitate the overall training progress. Our design balances generalization capability and geometric flexibility to achieve adaptive local resolutions. Moreover, we notice that Chamfer discrepancy fails in serving as the loss of the refinement module owing to its widely-discussed global "blindness" issue [1, 17, 39]. Therefore, we apply sliced Wasserstein distance [8], which circumvents the aforementioned drawbacks and maintains lower computational cost compared with the Earth Mover's distance. We demonstrate the effectiveness of the proposed method through extensive experiments and show that our framework outperforms state-of-the-art methods in both qualitative and quantitative respects.

## 2. Related Works

**Learning-based Human Reconstruction**: A line of methods [5, 34] utilizes parametric models to explicitly model the human body shape. To regress these model parameters, bodies are fitted to the 2D poses [7]. The maturity of detecting accurate human pose on RGB images automates the fitting process [11, 19]. However, such approaches usually produce models "minimally clothed", since the topological changes on the surface are difficult to model. [2, 3] alleviate the problem through learning vertex displacement on top of the model. However, it fails to generalize to surfaces with arbitrary topology. Neural implicit surface representation [40, 41] leverages its powerful expressiveness to support various topologies [6], where [40] shows that the representation can also be learned from incomplete data. Recent works [20, 42, 43] successfully regress an implicit function to recover clothed human shapes using a single RGB image. [52] proposes a voxel-based method, which is hard to generate high-resolution shape due to memory restrictions. [4, 31] address the problem by employing UV parameterization to transform shape modeling into 2D image-to-image translation.

Another line of works make efforts to combine the benefits of the simplicity of the parametric model and the expressive capabilities of neural implicit surface representation. Works such as [25] and [24] reconstruct 3D human shapes in a canonical space by warping query points from the canonical to parametric model's posed space and projecting them onto the 2D image space. [53] conditions the implicit function on the SMPL template for robustness to pose variation and reconstruct local details from the image pixels. [48] regresses shapes from inferred normals and SDF features. [49] utilizes templates and predicted front-back normals to explicitly generate detailed surfaces. The occlusions are inpainted by implicit function networks.

**Point Set Similarities**: Methods to quantify the similarities between point sets are widely used in point cloud-related tasks. Chamfer distance is one of the most widely-used metrics. Many works [12, 18, 50] adopt it as the evaluation metric or as the loss to be optimized. However, Chamfer distance only considers the nearest neighbor of the point, making it highly dependent on the initialization. Therefore, Wasserstein distance, as the solution to the optimal transport problem, is adopted by recent studies [1, 18]. However, calculating the Wasserstein distance for high dimensional distributions is non-trivial. Even with 3D point clouds, it is expensive to compute. [9] designs sliced Wasserstein distance to reduce the dimension favoring the computation. [30] utilizes it for auto-encoders and [17, 39] introduces it to point cloud learning tasks.

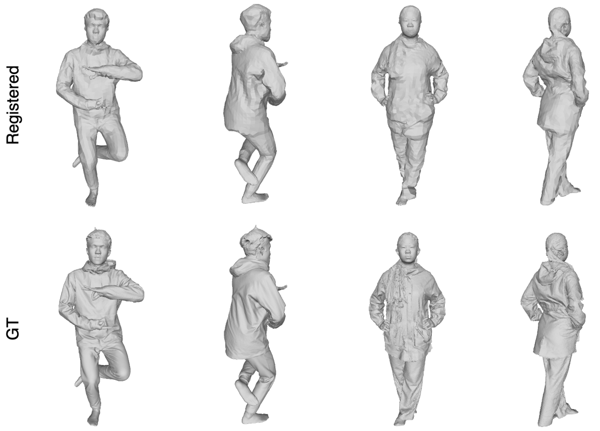
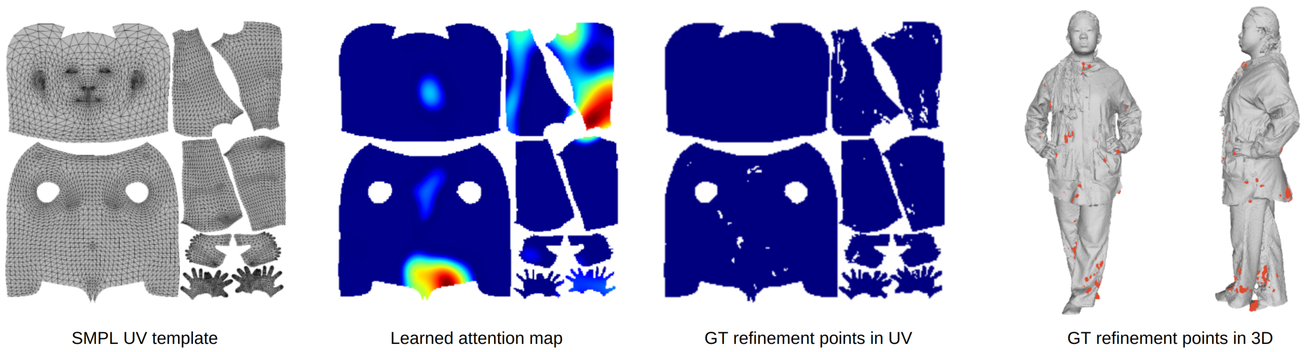


Figure 3. SMPL-registered meshes and Ground-truth meshes in two view angles

## 3. Proposed Method

To better leverage the power of the 2D neural network in 3D model reconstruction, we propose a framework, as illustrated in Figure 2, that employs both 2D UV-based and 3D point-based representations. Specifically, we adopt the coarse-to-fine strategy. To create a coarse model, we first parameterize both the color texture and an unclothed parametric template and render them into a uniform UV space. It is worth mentioning that this pre-processing step establishes pixel-aligned correspondences, thus enabling the network to generalize across various models. The aligned UV maps are then fed into an efficient 2D image-to-image translation module, which predicts the offsets of points and normals on the template UV map. For the purpose of clarity, all referenced 3D points in the following sections inherently include accompanying normals. Moreover, unlike previous work [4] we compute the loss directly in the 3D space, facilitating network training without necessitating strict data registration.

To enhance local expressiveness, we apply a 3D point-based refinement module, which predicts point sets corresponding to regions on the ground truth model that exhibit the most significant discrepancies from the coarse prediction. An attention map decoded from the UV image features is used to predict regions of interest. We generate additional point sets through a series of encoder-decoder layers with



SMPL UV template

Learned attention map

GT refinement points in UV

GT refinement points in 3D

Figure 4. Visualization of the learned attention map and ground-truth refinement points in UV and 3D space. The attention values are normalized to $[0, 1]$, with the dark blue for $0$ and the dark red for $1$. GT refinement points are illustrated as white pixels on the UV map.

the input of the predicted UV map multiplied by the attention map. These points are then added to the coarse prediction to yield the final result with adaptive local resolution. To ensure optimal performance, we split the training of the coarse prediction and the refinement module into two stages and subsequently train both modules in an end-to-end manner. In the following sections, we describe each component in detail.

### 3.1. Preliminaries

**UV Parameterization**: Unlike images, 3D models are not typically represented in a well-structured and ordered format, which disrupts the spatial locality assumptions, making the direct application of standard convolutional neural network (CNN) architectures challenging. Utilizing graph-based convolution on 3D data, however, introduces higher complexity. To overcome this limitation, we leverage the UV parameterization technique, which unwraps the surface of the 3D model onto a 2D plane with less projection distortion. After parameterization, each vertex in 3D is mapped to a position $(u, v)$ in 2D. The designated normal and positional maps are generated through barycentric interpolation on the related vertices' value, which can be formulated as $S(u, v) = \sum_{i=1}^{3} \omega_i P_i$, where $S(u, v)$ denotes the value corresponding to the pixel $(u, v)$, $P_i$ represents the vertex positions or normals of a vertex, and $\omega_i$ are the barycentric coordinates satisfying $\sum_{i=1}^{3} \omega_i = 1$.

**Unclothed Parametric Template**: We apply SMPL as the unclothed parametric template to serve as a base model and to unify the UV space. In SMPL, body shapes are driven by low-dimensional vectors, with $\beta \in \mathbb{R}^{10}$ representing shape parameters as weights for vertex offsets and $\theta \in \mathbb{R}^{72}$ denoting poses of 24 joints (including the pelvis as the root and 23 additional body joints). Each joint's pose is defined as the axis-angle rotation $R \in SO(3)$ relative to its parent in the kinematic tree. The SMPL model is accompanied by a pre-defined UV map that remains invariant to both $\beta$ and $\theta$. To ensure a consistent representation across all models, we employ a state-of-the-art algorithm [48] to estimate the unclothed SMPL fitting given an input RGB image. This fitting is then applied to generate both positional and normal maps.

### 3.2. Coarse Model Reconstruction

In this section, we intend to translate the parameterized images into XYZ position domain. We apply an approach similar to Tex2Shape [4], with alterations tailored to our optimization objective. In Tex2Shape, a pure 2D supervision and a PatchGAN discriminator [26] are used in loss calculation. However, generating 2D supervision in UV space requires registering the SMPL model to ground truth scan data, which is a challenging problem by itself. To demonstrate this point, we perform such registration using a state-of-the-art algorithm for human-body deformation [13] and report the registered meshes in Figure 3. As observed, the results are far from accurate, either over-smoothed or too noisy on resultant surfaces, implying that 2D supervision could limit the performance upper-bound and may not be optimal in practice.

Contrary to Tex2Shape, our framework adopts generative networks exclusively and innovatively supervises the output image in 3D by leveraging pixels sampled via the known UV mask. This design is amenable to registration-free data, thus enabling the network to fully harness the information offered by the raw models. Additionally, we parameterize the minimally-clothed SMPL model and concatenate it with the RGB texture map as the input together. It affords a reference to the result, thereby resolving the orientation and posing ambiguity. Hence, the network's focus is predominantly directed towards the reconstruction of the surface, as opposed to the learning of global absolute positions.

Our generator $G$ is built upon the architecture proposed by [28]. It consists of a convolutional downsampling front-end $G_F$, a set of residual blocks $G_R$, and an upsampling back-end $G_B$. The objective function is simply $\min \mathcal{L}_c(G(s), Y)$, where $G(s)$ is sampled into 3D using the SMPL UV mask and $Y$ is the ground truth pointcloud. We choose bidirectional Chamfer distance as the loss $\mathcal{L}_c$, which is defined as

$$\begin{aligned} d(S_1, S_2) = & \frac{1}{|S_1|} \sum_{x \in S_1} (\min_{y \in S_2} \|x - y\|) \\ & + \frac{1}{|S_2|} \sum_{y \in S_2} (\min_{x \in S_1} \|x - y\|) \end{aligned} \tag{1}$$

for two point sets $S_1$ and $S_2$.

### 3.3. Adaptive Refinement

After training a coarse prediction with the image-to-image translation network, in this section, we introduce our refinement module which adaptively enhances the local expressiveness. We define a resize-convolutional attention-head $G_{att}$ that decodes the regions of interest from global features $\mathbf{c}$, where $\mathbf{c} = G_R(G_F(\mathbf{s}))$. Replacing the transpose convolutional upsample layers in $G_B$ with the resize-convolution significantly reduces the checkerboard artifacts in the attention generation. Please refer to the supplementary for more details.

We explicitly apply attention to the inferred maps to obtain an image with information from the interested area only. We use a 5-layer convolutional encoder to learn the local features and an 8-layer MLP-based decoder to predict the refinement point set $X_r$. As each element in $X_r$ is an independent point, they support arbitrary topology and can be freely organized to refine surfaces regardless of size and location. The introduction of the attention-based strategy allows the following encoder-decoder network to achieve the "adaptive local resolution". We elaborate on this point in Section 4.4. We adopt a simple-yet-effective way to partition regions with finer details from the ground-truth model. Given the coarse prediction $X_c$ and the high-resolution model $Y$, we measure the local error of each point $p \in Y$ as the average of Chamfer distances to its $M$ neighboring points. From there, we sort the top $K$ points with the largest errors as the point set to be refined, denoted as $Y_r$.

The refinement points are trained with sliced Wasserstein distance (SW) [8]. Sliced Wasserstein distance is a point similarity measure with lower computational cost than EMD. Sliced-$p$-Wasserstein distance between distribution $\mu$ and $\nu$ is defined as

$$SW(\mu, \nu) = \left( \int_{\mathbb{S}^{d-1}} W_p^p(\mathcal{R}I_\mu(\cdot, \theta), \mathcal{R}I_\nu(\cdot, \theta)) d\theta \right)^{\frac{1}{p}} \tag{2}$$

where $I_\mu$ and $I_\nu$ are the probability density functions of measurements $\mu$ and $\nu$. This strategy facilitates the learning of effective features for the refinement module, overcoming the limitations of traditional Chamfer distance, which often fails to adequately guide the learning process. In practice, $SW(\mu, \nu)$ is usually computed through Monte-Carlo approximation:

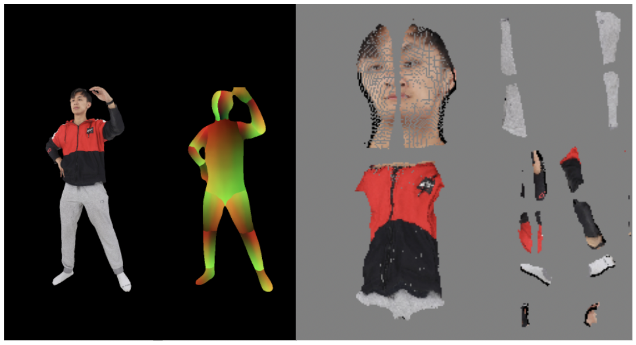


Figure 5. An example of the process to create an SMPL partial UV map. The input RGB image is first transformed into an IUV image through DensePose, which contains UV coordinates per part. Then we utilize a preset mapping to map the IUV to the partial UV on the right.

$$SW(\mu, \nu) \approx \left( \frac{1}{L} \sum_{l=1}^{L} W_p^p(\mathcal{R}I_\mu(\cdot, \theta_l), \mathcal{R}I_\nu(\cdot, \theta_l)) \right)^{\frac{1}{p}} \tag{3}$$

where $\theta_l$ is uniformly sampled on hypersphere $\mathbb{S}^{d-1}$. Therefore, the loss of the refinement module is formulated as

$$\mathcal{L}_r = SW(X_r, Y_r; L) \tag{4}$$

where $L$ is the size of the slice portfolio of Monte-Carlo estimation. Within the proposed framework, we adopt a modified version of Sliced Wasserstein distance, which further improves its performance. We present the content in the supplementary material.

## 4. Experiments

### 4.1. Experimental Settings

We utilize THuman2.0 dataset [51] for training and evaluation. It contains 526 high-fidelity human scans in various poses with 8K resolution textures as well as ground-truth SMPL fittings. The positional map input is created by rendering minimally-clothed SMPL fitting into the UV space. During the inference, SMPL's pose parameters $\theta$ and shape parameters $\beta$ are estimated with loop optimization following [48]. The unwrapped partial texture is created through DensePose [22]. DensePose predicts UV coordinates of 24 body parts separately based on the SMPL body model. We manually design an arrangement and synthesize the 24 patches into a whole UV map, illustrated in Figure 5.

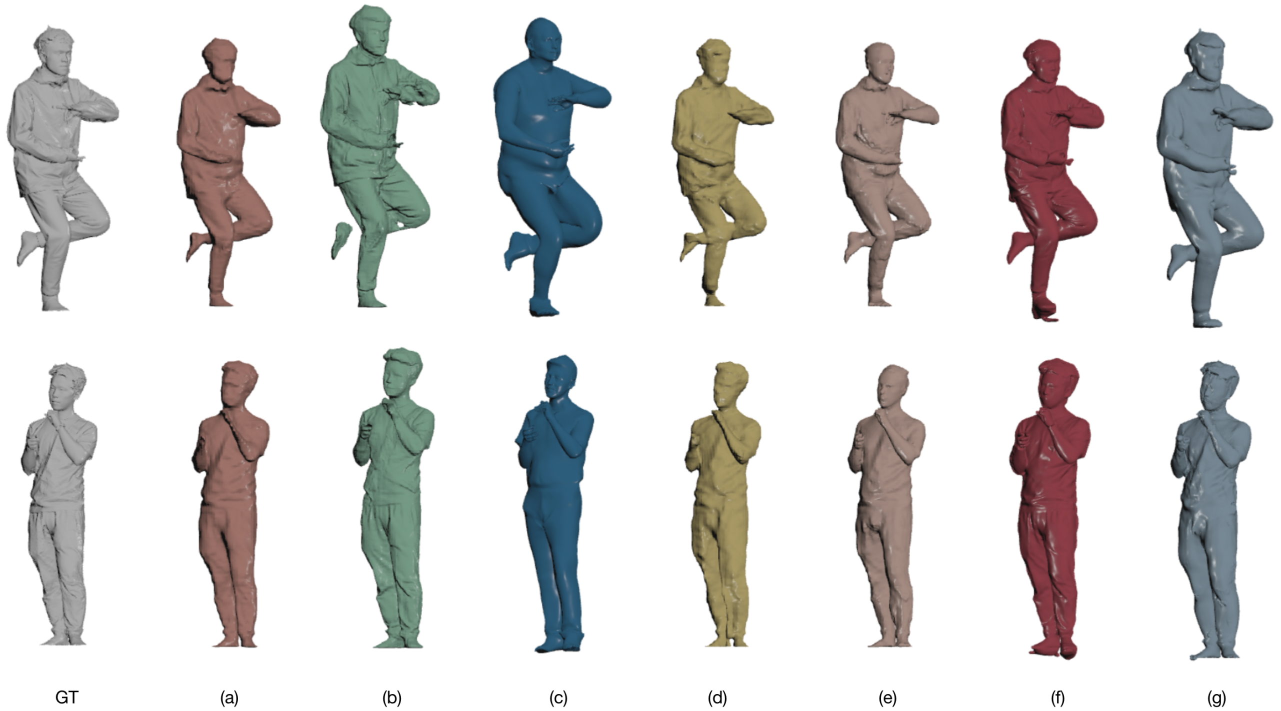


Figure 6. (a) PIFu, (b) PIFuHD, (c) Tex2Shape, (d) PaMIR, (e) ICON, (f) ECON, and (g) Ours. Our model presents the class with the finest class of details on visible regions.

To evaluate the effectiveness of the proposed two-step framework, we conduct a thorough comparison, both quantitatively and qualitatively, between the results of our method and several state-of-the-art methods, including PIFu [42], PIFuHD [43], Tex2Shape [4], PaMIR [53], ICON [48], and ECON [49]. To maintain a standard of comparison, all models are normalized to a uniform height of $1.8m$. The same set of SMPL parameters is shared across all pose-aware methods. As Tex2Shape and PIFuHD do not have training details revealed, the pre-trained models provided by the authors are used in the evaluation. More implementation details are described in the supplementary material.

## 4.2. Metrics

**Chamfer and Point-to-Surface distance**: To identify significant geometric errors, such as occlusions or mispositioned limbs, we use the widely accepted Chamfer distance (CD) and Point-to-Surface distance (PSD) for comparing reconstructed meshes to the ground truth. The PSD measures from the generated results to the GT.

**Normal Consistency**: To evaluate the accuracy of local detail reconstructions, we include the Normal Consistency (NC), a scale-invariant metric that complements CD and PSD. NC is calculated by the cosine distance between normals of the nearest faces in the reconstructed models and the ground truth.

## 4.3. Evaluation

**Qualitative Results**: Given that our method outputs in the form of 3D point clouds, to compare with other works, we employ classic Poisson Surface Reconstruction (PSR) [29] to generate the equivalent mesh results. Figure 6 provides a visual comparison of our method with the aforementioned state-of-the-art methods from the input viewangle. Tex2Shape's outcome is bounded by the topology of the SMPL model, making it unrealistic to the input. Our model creates the same class of visual appearance compared with much heavier methods, such as PIFuHD and ECON.

Nevertheless, in Figure 7, we compare the results from the input view angles with those from rotated view angles to illustrate the difference in reconstruction quality between visible and occluded regions. It can be observed that many methods, although with fine surface details, fail to preserve reasonable human body shape for occluded regions, even though some of them are conditioned with parametric models. Our method outperforms competing methods: consistently delivers superior body shape reconstruction accuracy in such challenging circumstances while achieving surface detail recovery commensurate with SOTA methods. It can be attributed to our innovative use of UV space learning, which ensures the preservation of accurate body shape from SMPL during the learning, regardless of visibility con-

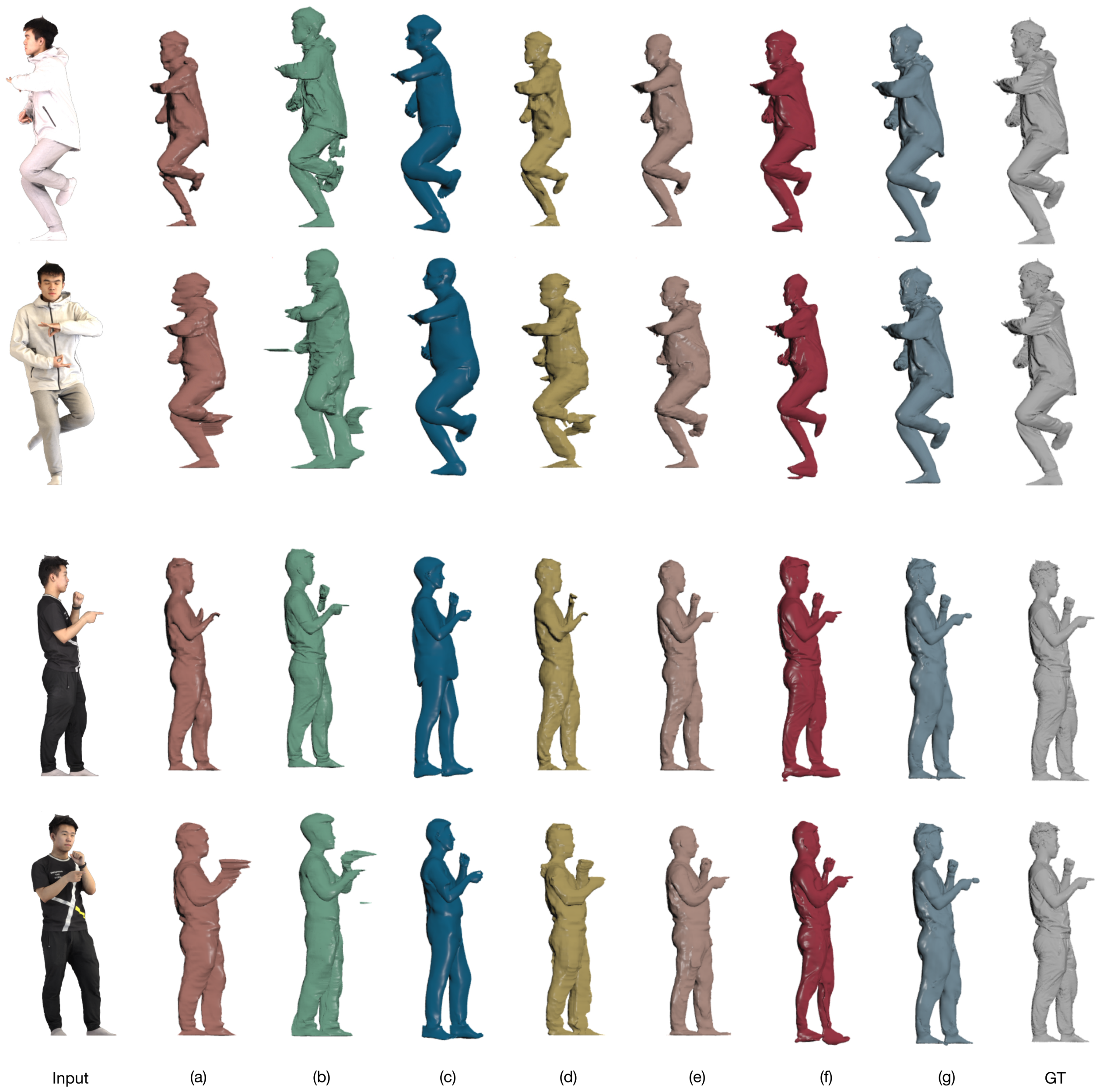


Figure 7. Reconstructed models when the visualized areas are visible (upper) and occluded (lower). (a) PIFu, (b) PIFuHD, (c) Tex2Shape, (d) PaMIR, (e) ICON, (f) ECON, and (g) Ours. Our method preserves the most of shape in occluded regions compared to pose-agnostic methods, such as PIFu and PIFuHD. Models from ECON also exhibit significant shape distortion on side views as well. Moreover, our results present the most detailed surfaces in comparison with all other competing approaches.

straints.

To highlight the improvement brought about by the refinement module, we offer a visual comparison in Figure 8. The difference between the surfaces reconstructed with and without the refinement module is clearly evident. With the addition of refinement points, our method successfully reproduces intricate details such as the deep wrinkles of trousers, which are loosely learned by the coarse prediction due to the efficiency concern. For additional comparisons, we refer readers to the supplementary material where we have included extended samples and viewpoints.

**Quantitative Evaluation**: The precise numerical results are tabulated in Table 1. Note that the raw point cloud

Table 1. Performance comparison with state-of-the-art methods. The best results are bolded.

| | CD (cm) | PSD (cm) | NC |
|---|---|---|---|
| PIFu [42] | 1.5443 | 1.3746 | 0.1095 |
| PIFuHD [43] | 1.4928 | 0.9586 | 0.1063 |
| Tex2Shape [4] | 5.0125 | 4.9463 | 0.2200 |
| PaMIR [53] | 1.5355 | 1.0187 | 0.1079 |
| ICON [48] | 1.1151 | 1.1884 | 0.0885 |
| ECON [49] | 1.2494 | 1.2931 | 0.0642 |
| Ours, w/o refine | 0.8597 | 0.8654 | 0.0876 |
| **Ours, with refine** | **0.8169** | **0.8412** | **0.0621** |

output is used instead of the mesh after PSR. Our method without refinement outperforms others in CD and PSD, as the metrics focus more on the general shape correctness. It is coherent with the visualization of models in rotated views. The introduction of our refinement module contributes high-frequency details to the output, resulting in much lower NC than competing methods and achieving the SOTA performance. Thanks to the high-level shape preservation, our NC is even lower than those from PIFuHD and ECON, although the local details are not visually superior in well-reconstructed regions.

### 4.4. Attention v.s. Global Input

In Section 3.3, we highlight the significance of our attention strategy to the refinement module. To quantify its efficacy, we execute an ablation experiment, wherein the attention map is eliminated, leaving only the predicted positional map as the input. Figure 9 exhibits the comparison of average losses per epoch with and without the attention map. The results clearly indicate that in the absence of the attention head $G_{att}$ and with exclusive reliance on the coarse UV positional map, the learning process fails to produce meaningful features that are conducive to the prediction of 3D refining points. This observation underscores the pivotal role the attention strategy serves in directing the locales of 3D point regression.

To further illustrate the significance of the attention module, we visualize the correlation between our learned attention map and the refinement points in Figure 4. By projecting the 3D refinement points onto the SMPL template, we are able to delineate their approximate locations on the UV map. We notice that regions of interest, such as the bottom of the torso and the two legs, corresponded with highlighted areas on the attention map. This correlation not only validates our attention-based strategy but also underscores its semantic significance.

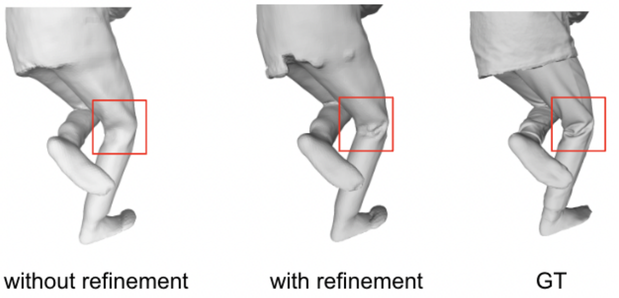


Figure 8. Surface reconstruction difference with and without refinement module. The wrinkles induced by the bending leg are more obvious with the refinement points.

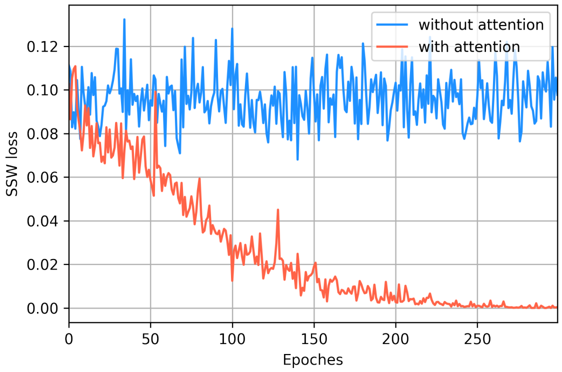


Figure 9. Per-epoch SW curves with and without attention head.

## 5. Conclusion

In summary, we present a novel framework that integrates 2D UV space training and 3D point cloud-based regression to generate fine-grained human geometry. We propose an attention-based strategy that pinpoints areas necessitating enhanced detail. The evaluations demonstrate that the proposed framework achieves state-of-the-art performance in single-view clothed human reconstruction tasks.

Despite the promising results, improvements are envisaged: On seen areas, surfaces generated by the PSR may not fully exploit information from the raw output and display less detail than those produced by implicit function-based methods. Moreover, similar to most methods that take a parametric model as input, our method relies on SMPL fitting for pose accuracy, which is a problem not fully solved. While we adopt the optimization pipeline from [48, 49] for modeling, there are still instances of reported failures.

**Acknowledgements.** This research was supported by the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D program in part (P0019797).

## References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 3

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

[3] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 3

[4] 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. 2, 3, 4, 6, 8

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

[6] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision*, pages 311–329. Springer, 2020. 3

[7] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*. Springer International Publishing, 2016. 3

[8] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1): 22–45, 2015. 3, 5

[9] Nicolas Bonnotte. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Paris 11, 2013. 3

[10] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Generative and discriminative voxel modeling with convolutional neural networks. *arXiv preprint arXiv:1608.04236*, 2016. 1

[11] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3

[12] Kunyao Chen, Fei Yin, Baichuan Wu, Bang Du, and Truong Nguyen. Mesh completion with virtual scans. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3303–3307. IEEE, 2021. 3

[13] Kunyao Chen, Fei Yin, Bang Du, Baichuan Wu, and Truong Q Nguyen. Efficient registration for human surfaces via isometric regularization on embedded deformation. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 4

[14] 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 Transactions on Graphics (ToG)*, 34(4):1–13, 2015. 1

[15] Theo Deprelle, Thibault Groueix, Matthew Fisher, Vladimir Kim, Bryan Russell, and Mathieu Aubry. Learning elementary structures for 3d shape generation and matching. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[16] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4): 1–13, 2016. 1

[17] Bang Du, Kunyao Chen, Haochen Zhang, and Truong Nguyen. Select-sliced wasserstein distance for point cloud learning. In *International Conference on 3D Vision (3DV)*, 2024. 3

[18] 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. 1, 3

[19] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 3

[20] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 3

[21] 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

[22] 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. 5

[23] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019. 1

[24] 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. 3

[25] 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. 3

[26] 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. 4

[27] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011. 1

[28] 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. 4

[29] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 6

[30] Soheil Kolouri, Phillip E Pope, Charles E Martin, and Gustavo K Rohde. Sliced wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018. 3

[31] Zorah Lahner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 667–684, 2018. 3

[32] 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*, 2018. 1

[33] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[34] 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

[35] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J Black. Scale: Modeling clothed humans with a surface codec of articulated local elements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16082–16093, 2021. 2

[36] 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. 1

[37] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[38] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. 1

[39] Trung Nguyen, Quang-Hieu Pham, Tam Le, Tung Pham, Nhat Ho, and Binh-Son Hua. Point-set distances for learning representations of 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10478–10487, 2021. 3

[40] 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. 1, 3

[41] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, pages 523–540. Springer, 2020. 3

[42] 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. 2, 3, 6, 8

[43] 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, 3, 6, 8

[44] 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. 1

[45] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019.

[46] Chao Wen, Yinda Zhang, Zhuwen Li, and Yanwei Fu. Pixel2mesh++: Multi-view 3d mesh generation via deformation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1042–1051, 2019. 1

[47] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 1

[48] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 2, 3, 4, 5, 6, 8

[49] 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 (CVPR)*, 2023. 2, 3, 6, 8

[50] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 206–215, 2018. 3

[51] 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)*, 2021. 5

[52] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7739–7749, 2019. 3

[53] 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. 3, 6, 8