Coarse-to-fine 3D Clothed Human Reconstruction using Peeled Semantic Segmentation Context*

Snehith Goud Routhu
IIIT Hyderabad
snehith.goud@research.iiit.ac.in

Sai Sagar Jinka
IIIT Hyderabad
jinka.sagar@research.iiit.ac.in

Avinash Sharma
IIIT Hyderabad
asharma@iiit.ac.in

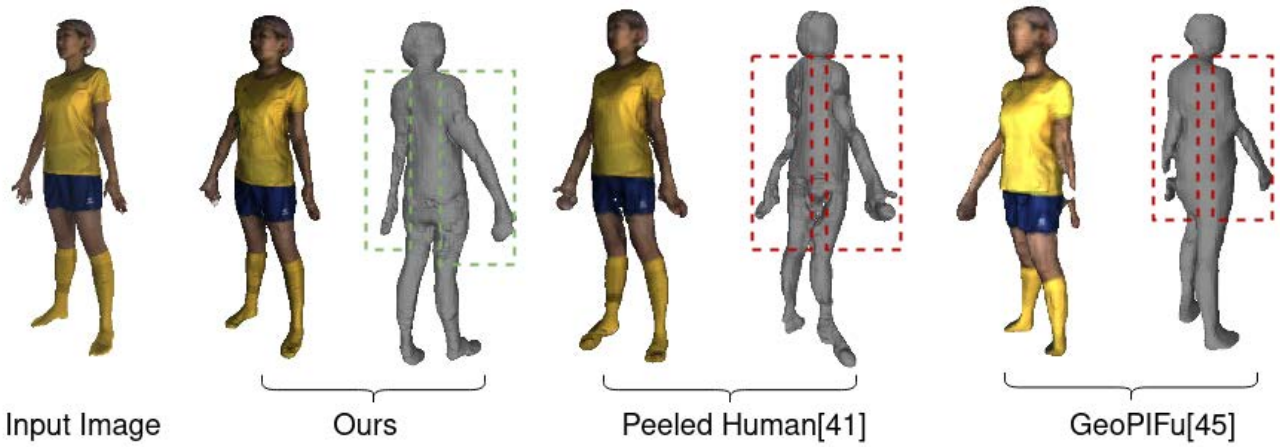

Figure 1: Our coarse-to-fine refinement approach using semantic segmentation context yields superior reconstruction of 3D human body from monocular input image as compared to other SOTA methods.

ABSTRACT

3D reconstruction of human body model from a monocular image is an under-constrained and challenging yet desired research problem in computer vision. Recently proposed multi-layered shape representation called PeeledHuman attempted a sparse non-parametric 2D representation that can handle severe self-occlusion. However, the key limitation of their PeeledHuman model is that the predicted depth maps of self-occluded parts are sparse and noisy, and hence after back-projection lead to distorted body parts and sometimes with discontinuity between them. In this work, proposed to introduce Peeled Segmentation map representation in a coarse-to-fine refinement framework which consist of a cascade three networks namely, PeelGAN, PSegGAN and RefGAN. At first, we use original PeeledHuman as baseline model to predict initial coarse estimation of peeled depth maps from input RGB image. These peeled maps are subsequently fed as input along with monocular RGB image to our novel PSegGAN which predict Peeled Segmentation maps in a generative fashion. Finally, we feed these peeled segmentation maps as additional context along with monocular input image to our RefGAN which predicts the refined peeled RBG and Depth maps. This also provides an additional output of 3D semantic segmentation of the reconstructed shape. We perform thorough empirical evaluation over four publicly available datasets to demonstrate superiority of our model.

CCS CONCEPTS

• **Computing methodologies** → **Reconstruction**.

KEYWORDS

Semantic Segmentation,3D Reconstruction

ACM Reference Format:
Snehith Goud Routhu, Sai Sagar Jinka, and Avinash Sharma. 2021. Coarse-to-fine 3D Clothed Human Reconstruction using Peeled Semantic Segmentation Context. In *Proceedings of 12th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'21),* Chetan Arora, Parag Chaudhuri, and Subhransu Maji (Eds.). ACM, New York, NY, USA, Article 133, 9 pages. https://doi.org/https://doi.org/10.1145/3490035.3490293

1 INTRODUCTION

3D reconstruction of human body model is a very challenging yet desired research problem with applications in VR/AR domain, entertainment & gaming as well as marketing domains (e.g., virtual try-on). Metrically accurate and precise reconstructions of humans is now possible with calibrated multi-view systems [21, 39] that uses RGB or structured light cameras. Nevertheless, these techniques have remained largely inaccessible to the general community due to its reliance on professional capture systems with strict environmental constraints like large number of cameras, controlled illuminations etc., that are overly expensive.

*Produces the permission block, and copyright information

ICVGIP'21, December 2021, Jodhpur, India

ACM ISBN 978-1-4503-7596-2.
https://doi.org/https://doi.org/10.1145/3490035.3490293

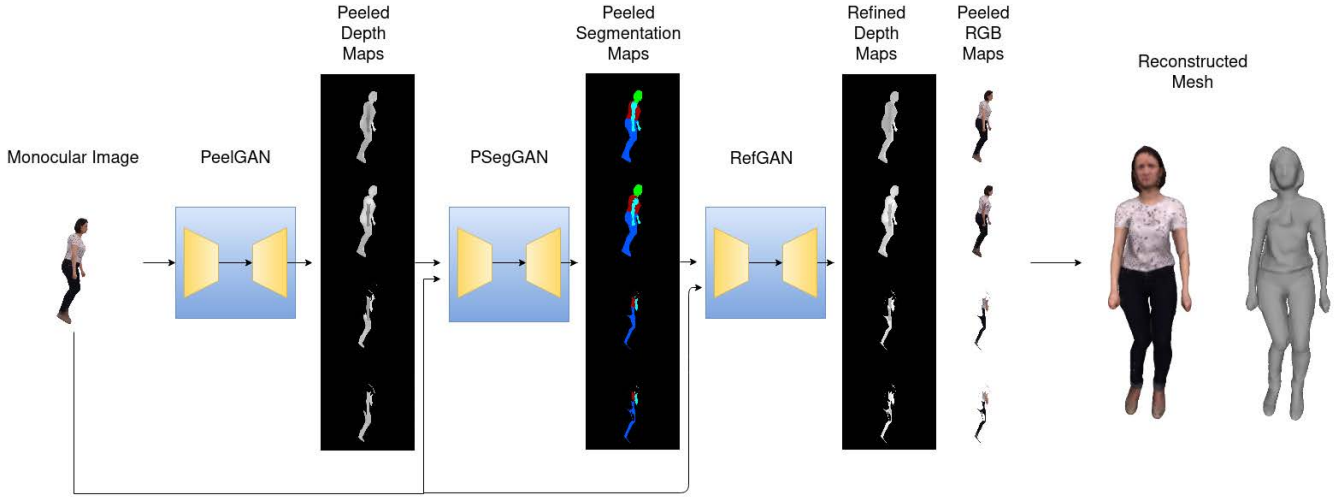

Figure 2: Coarse-to-fine refinement framework for 3D reconstruction of human body using peeled semantic segmentation.

On the other hand, reconstruction from a single view is an inherently under-constrained problem. However, recent advancement in deep learning domain [1, 14, 22, 30, 37, 45] have shown great promise in acquiring 3D body reconstructions from a monocular image. The first class of parametric body model based monocular shape reconstruction techniques infer the shape and pose parameter of a statistical body model, like SMPL [11, 12, 28]. These methods do well in recovering reasonably accurate body pose, but the reconstructed geometry is constrained to be within the model space, which can not capture the geometrical details including surface geometrical details and loose clothing. Other recent works [5, 6, 43, 45], recover static body shape, and clothing as displacements on top of the SMPL body model [28] (model-based). Semantic segmentation labels have been used to refine parametric body estimation [47]. Nevertheless, under this configuration only tight garments can be reconstructed. Thus, garments like skirts and robes, which have a different topology with body shape, are beyond their representation range.

The second class of monocular 3D reconstruction techniques [14, 37, 41, 42, 51] uses non-parametric representations to infer detailed geometry beyond basic body shape and pose. These non-parametric representations are either based on voxels or implicit function and can recover arbitrary shapes. BodyNet [14] leverages intermediate tasks (e.g. segmentation, 2D/3D skeletons) and estimates low-resolution body voxel grids. A similar work to BodyNet is VRN [2], but it directly estimates occupancy voxels without solving any intermediate task. Voxel-based methods [15, 37] often produce errors at the limbs of the body and require fitting a model post-hoc [15]. However, voxel representation are typically of lower resolution due to memory limitations and thus fails to recover high frequency surface geometry details. Whereas, deep implicit function learning techniques train deep neural networks to estimate dense, continuous occupancy fields from which meshes may be reconstructed e.g. via Marching Cubes [27]. These methods either extracts either a global image feature [8, 18, 25], or pixel-aligned features [41, 42] to drive this estimation. However, these methods fails to take into account the fine-grained local shape patterns and do not seek to enforce global consistency to encourage physically plausible shapes and poses in the reconstructed mesh. This can lead to unnatural body shapes or poses, and loss of high-frequency surface details within the reconstructed mesh. Recently proposed model GeoPiFu [44] attempt to combine both volumetric and implicit function learning by providing U-Net based volumetric features as prior for implicit function prior. However, their method is computationally intensive and perform dense inference at test time (similar to [42]).

An alternative set of non-parametric approaches attempt to model 3D objects/scenes as sparse layered representation. [46] attempted to predict 3D human body reconstruction by predicting the frontal and backside depth maps from monocular image. However, they failed to model the scenarios of self-occlusions by body parts, clothing etc. [40] attempted to address the self-occlusion problem by predicting multiple peeled depth maps. They estimate a fixed number of ray intersection points with the human body surface in the canonical view volume for every pixel in an image, yielding a multi-layered shape representation called **PeeledHuman**. Their representation encodes a 3D shape as a set of layered depth maps called as *Peeled Depth maps* and the surface texture/color maps called as *Peeled RGB maps*. Subsequently, they learn to predict these peeled RGB and Depth Maps from monocular input image using their *PeelGAN* model. Their proposed shape representation allows to recover multiple 3D points that project to the same pixel in the 2D image plane thereby overcoming the problem of self-occlusion. Recently, [19] proposes to introduce parametric body prior to variant of PeelGAN. However, they require good estimate of SMPL prior fitted to input image.

In this paper, we address the problem of detailed full-body shape reconstruction from a single image. We adopt the PeeledHuman [40] representation as this encoding is sparse and robust to self-occlusions. However, the key limitation of their PeelGAN model is that predicted depth maps of self-occluded parts are sparse and noisy, and

hence after back-projection lead to distorted body parts and sometimes discontinuity between them. We hypothesized that introducing coherence in semantic labels would alleviate such issues. Additionally, the semantic labels also helps in improving the generalisation of our model as shown in Figure 1 where we compare our method with SOTA. It is important to note that, we propose and use peeled semantic representation that enables capturing 3D shape context as compare to traditionally used segmentation map for monocular rgb image. To summarize, we propose to introduce *Peeled Segmentation map* representation in a coarse-to-fine refinement framework which consist of a cascade of three networks namely, PeelGAN, PSegGAN and RefGAN, as shown in Figure 2. At first, we use original PeelGAN [40] as baseline model to predict initial coarse estimation of peeled depth maps from input RGB image. These peeled maps are subsequently fed as input along with monocular RGB image to our novel PSegGAN which predict *Peeled Segmentation maps* in a generative fashion. These peeled segmentation maps are spatially aligned to peeled depth maps and thus we get semantic segmentation label corresponding to each pixel of the peeled depth maps. Finally, we fed these peeled segmentation maps as additional context along with monocular input image to our RefGAN to predict the refined peeled RBG and Depth maps. We observe that providing the semantic segmentation context improves the reconstructions both quantitatively and qualitatively especially in the case of self-occlusions. This also provide an additional output as 3D semantic segmentation of the reconstructed shape. We perform a thorough empirical evaluation over four publicly available datasets to demonstrate the superiority of our model over [40]. Additionally, we intend to release the semantic segmentation label data and proposed model code for the community to use.

2 LITERATURE SURVEY

2.1 Parametric 3D Reconstruction

Many existing methods have attempted estimating a naked body shape (statistical models [11, 28]) from image [1, 7, 10, 20, 24, 26, 30, 34, 49]. Most of the current model based approaches optimize the pose and shape of SMPL[13, 28] to match image features, which are extracted with bottom-up predictors. The most popular image features are 2D joints [7], or 2D joints and silhouettes [4, 5, 17]. HMR proposes to regress SMPL parameters when minimizing reprojection loss with the known 2D joints. For better representation ability, a displacement vector is added for each vertex. [4, 5, 45] adopt this strategy to reconstruct clothed body with skin-tight garment. Alldieck et al. [43] estimate detailed normal and vector displacement on the UV map, which leads to finer-scale details. Zhu et al. [54] model fine-scale details by adding free-form 3D deformation on top of parametric model. Instead of using a single surface to represent both garment and body, [6] separates SMPL mesh to represent upper garment and pant independently, leading to more flexible control. However, binding garment vertices to the body model strictly restricts the topology of support garment categories, and it is hard to represent more loose garment types, such as skirts. [35] also uses separate body and garment templates to register clothed body motion sequences. Pixel2Mesh [34] generates water-tight meshes by progressively deforming a sphere template.

2.2 Non-parametric 3D Reconstruction

Some non-parametric methods based on voxel or implicit function have been proposed to address the complex topology of garments. Volumetric regression [3, 14, 50] directly infers a voxel representation of clothed bodies with a deep network. Due to the large memory cost for high resolution, high-frequency details are often missed. [51] infer clothed body volume representation with an initial aligned SMPL body, and combine image features to enhance reconstruction details. Natsume et al. [37] propose a reconstruction method based on a multi-view framework using synthesizing new silhouettes from a single image. TSDFs [9] can represent the human surface implicitly, which is common in depth-fusion approaches [31, 32]. Occupancy Networks [25], DeepSDF [18] and LIF [8] proposed to use global representations of a single-view image input to learn deep implicit surface functions for mesh reconstruction. However, the global representation-based implicit function does not have dedicated query point encodings and thus lacks modeling power for articulated parts and fine-scale surface details.This motivates later works of PIFu [42] and DISN [36]. They utilize pixel-aligned 2D local features to encode each query point when estimating its occupancy value. The alignment is based on (weak) perspective projection from query points to the image plane, followed by bi-linear image feature interpolation. However, PIFu still suffers from the feature ambiguity problem and lacks global shape robustness. Another two PIFu variations are PIFuHD [41] and ARCH [52]. PIFuHD leverages higher resolution input then PIFu through patch-based feature extraction to accommodate GPU memory constraints. ARCH combines parametric human meshes (e.g. SMPL [28]) with implicit surface functions in order to assign skinning weights for the reconstructed mesh and enable animations. Both these methods require more input / annotations (e.g. 2×higher resolution color images, SMPL registrations) than PIFu. Geo-PIFu [44] is volumetric-regression based approach that incur high computational and memory costs. Unsupervised estimation of implicit functions has been addressed in [29, 38] . Peeled maps [40] proposed a sparse representation by estimating only surface intersections by posing the problem as an extension of ray tracing.

3 METHOD

3.1 PeeledHuman Representation

Peeled representation is a sparse, non-parametric encoding of 3D shape proposed in [40] where 3D human body is modeled as a set of *Peeled Depth & RGB maps*. Under the assumption that human body is a non-convex object placed in a virtual scene, a set of rays originating from the camera center are traced through each pixel to the 3D world. The first set of ray-intersections with the 3D body are recorded as depth map d_1 and RGB map r_1, capturing visible surface details nearest to the camera. Subsequently, the rays are extended beyond the first bounce to hit the next intersecting surface. The corresponding depth and RGB values of the next layer are represented by d_i and r_i, respectively. Additionally, [40] claimed that 4 intersections of each ray i.e., 4 *Peeled Depth & RGB maps* are sufficient to faithfully reconstruct a human body, which can handle self-occlusions caused by the most frequent body poses. Subsequently, a point-cloud can be reconstructed from these maps using classical camera back-projection methods. Please refer to [40]

for visualization and other details. This sparse 2D representation is more efficient to process than volumetric and implicit functions as it only stores ray-surface intersection in a 2D multi-layered layout.

3.2 Overview & Notations

Our objective is to reconstruct a 3D Human with clothing from a given monocular RGB Image I_{inp}. We employ a coarse-to-fine strategy with a cascade of multiple GANs. Initially, pre-trained **PeelGAN** [40] is used to predict the coarse peeled depth maps D_{peel} from monocular image I_{inp}. Subsequently, these predicted D_{peel} along with input RGB image are fed to our proposed **PSegGAN** which outputs *Peeled Segmentation maps* $\hat{S}$ in a generative fashion using the supervision of ground-truth *Peeled Segmentation maps* S_{gt}. Later, we feed these predicted segmentation maps $\hat{S}$ as a semantic context along with the monocular RGB image I_{inp} to the proposed **RefGAN** and predict the final *Peeled Depth maps* $\hat{D}$ *and RGB maps* $\hat{R}$. We propose **RefGAN** as a conditional generative adversarial network to refine the coarse peeled maps. We train the network in supervised manner using the ground-truth *Peeled Depth* (D_{gt}) *and RGB* (R_{gt}) *maps.*

In the case of self-occlusion, the peeled semantic segmentation maps provides refined information about the boundary of occluded parts. Even in case of poses without self-occlusion the semantic information provides the spatial homogeneity context for pixels belonging to the same part. Finally, we can back-project the predicted $\hat{D}$ and $\hat{R}$ into a 3D point-cloud using the known camera intrinsic parameters, which are dependent only on the image resolution. We further extract the mesh from the resulting point-cloud using the Poisson surface reconstruction [23] Note that the camera intrinsic parameters is dependent on input resolution and is fixed across the datasets.

3.3 Peeled Segmentation Network (PSegGAN)

In order to generate the *Peeled Segmentation maps* ($\hat{S}$) from *Peeled Depth maps* D_{peel} and RGB image I_{inp} in a generative fashion, we propose PSegGAN, a conditional GAN. The input to our PSegGAN is D_{peel} & I_{inp} and it generates *Peeled Segmentation maps* $\hat{S}$ as output as shown in Figure 3. We use cross-entropy loss over the predicted segmentation maps $\hat{S}$ and ground-truth segmentation maps S_{gt} as it is a classification task along with the GAN loss. The final loss function is shown below:

$$loss_{total} = w_g * loss_{GAN} + w_s * loss_{seg} \tag{1}$$

where w_g and w_s are the weighting factors for $loss_{GAN}$ and $loss_{seg}$.

$$loss_{GAN} = E_{I_{inp},S_{gt}}[logD(I_{inp}, S_{gt})] + E_{I_{inp},\hat{S}}[log(1 - D(I_{inp}, \hat{S}))] \tag{2}$$

and

$$loss_{seg} = \sum_{i=1}^{n}\sum_{k=1}^{c} -t_k^i * log(\hat{y}^i) \tag{3}$$

where n is the total number of pixels in the image,c is the total number of part labels of human, t_k^i is 1 when $S_{gt}^i = c$ other wise it is 0 and $\hat{y}$ is the last layer of PSegGan i.e $argmax(\hat{y}) = \hat{S}$

3.4 Refined Peeled Map Prediction (RefGAN)

We predict the refined *Peeled Depth and RGB maps* using a generative network, named **RefGAN**. The RefGAN takes a set of peeled segmentation maps $\hat{S}$ and monocular RGB image I as input and generates refined *Peeled Depth* $\hat{D}$ *and RGB Maps* $\hat{D}$. The peeled segmentation maps provide dense semantic context to our generator network RefGAN. Refer Figure 4 for the architecture. We use L_1 loss over the predicted and ground-truth depth maps and RGB maps. In order to maintain 3D structure consistency, we add Chamfer loss over point-cloud obtained by back projecting the depth maps. Additionally, in order to enforce smoothness over predicted depth maps, we make first order gradients of generated depth maps $\hat{D}$ to be similar with ground-truth depth maps D_{gt}. So the combined loss function is shown as below:

$$L = w_g * l_{GAN} + w_l * l_{L_1} + w_c * l_c + w_{sm} * l_{sm} + w_r * l_{rgb} \tag{4}$$

where $w_g, w_l, w_r, w_c, w_{sm}$ are the respective weights for l_{GAN}, l_{L1}, l_{rgb}, l_c and l_{sm}.

Below, we define each loss terms:

$$l_{GAN} = E_{I_{inp},D_{gt}}[log(D(I_{inp}, D_{gt})] + E_{I_{inp},\hat{D}}[log(1 - D(I_{inp}, \hat{D}))] + E_{I_{inp},R_{gt}}[log(D(I_{inp}, R_{gt}))] + E_{I_{inp},\hat{R}}[log(1 - D(I_{inp}, \hat{R}))] \tag{5}$$

We define depth loss (l_{L_1}) as:

$$l_{L1} = \sum_{i=1}^{4} m_i \left|\hat{D^i} - D_{gt}^i\right| \tag{6}$$

where i is the i^{th} layer in peeled maps, m_i is 1 for occupied pixels (segmentation label at the pixel is not background), else 0 for depth maps ($\hat{D^1}, \hat{D^2}$) and m_i is 10 for occupied pixels, else 0 for depth maps ($\hat{D^3}, \hat{D^4}$), as these are sparse and hence more weight is given to them.

Smoothness loss (l_{sm}) is defined as :

$$l_{sm} = \sum_{i=1}^{4} \left|\nabla\hat{D^i} - \nabla D_{gt}^i\right| \tag{7}$$

To further improve the RGB peel maps conditioned on generated segmentation peel maps ($\hat{S}$), we use l_{rgb}

$$l_{rgb} = \left|\hat{R} - R_{gt}\right| \tag{8}$$

To ensure 3D consistency, we further use the chamfer loss l_c between predicted and ground truth point clouds.

$$l_c = \sum_{\hat{p}\epsilon\hat{P}} \min_{p\epsilon P_{gt}} ||\hat{p} - p||_2^2 + \sum_{p\epsilon P_{gt}} \min_{\hat{p}\epsilon\hat{P}} ||p - \hat{p}||_2^2 \tag{9}$$

3.5 Network Architecture

The peeled segmentation network (PSegGAN) consists of an encoder block, followed by residual blocks and a decoder block. The input to the network is 512x512 monocular RGB image and four peeled depth maps predicted from PeelGAN network. Dropout and Batch Normalisation are used for regularisation. The activation

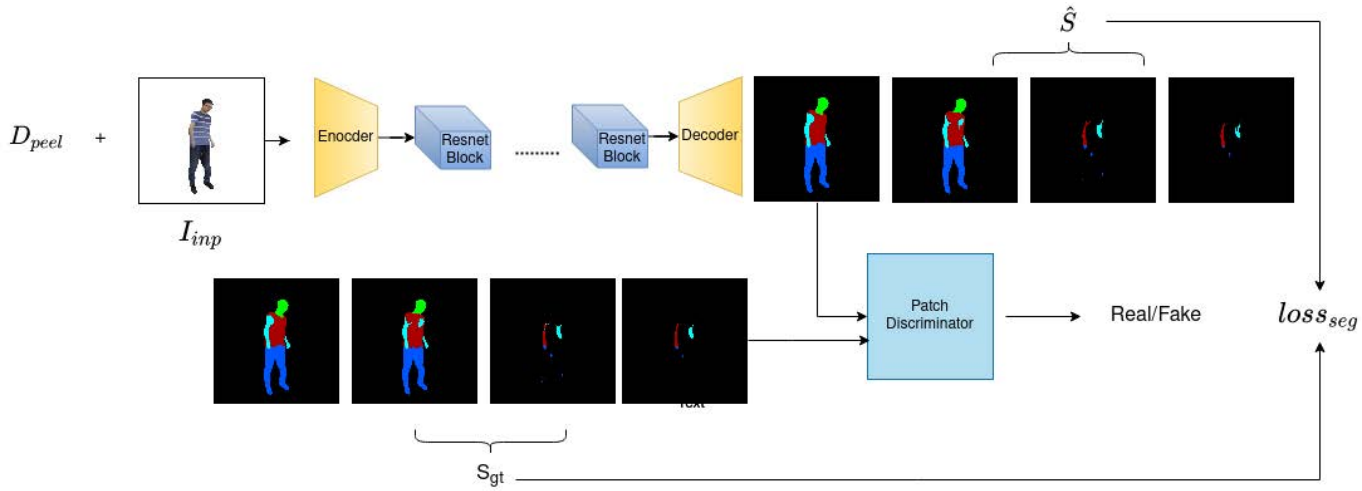

Figure 3: Architecture of the proposed PSegGAN that predicts Peeled Segmentation maps given input monocular image and coarse Peeled Depth maps predicted from PeelGAN.

function used is ReLU. The RefGAN network also use the same architecture of encoder, residual blocks but different decoder branches for depth and RGB maps prediction. The activation function used is ReLU except for the last convolution layer of decoders where sigmoid is used in depth decoder and Tanh is used in RGB decoder. The architecture of the networks is similar to [40]. The discriminator for both the networks is a patch based discriminator similar to *PatchGAN* proposed in [33].

4 EXPERIMENTS AND RESULTS

4.1 Datasets

4.1.1 Thumans. The dataset [53] consists of 6800 human meshes registered with SMPL body in varying poses and garments. We manually do semantic segmentation on the SMPL body. For each vertex in the ground-truth mesh, we find its nearest neighbour in the registered SMPL mesh and assign its label.

4.1.2 MonoPerfCap. The dataset [48] consists of 13 daily human motion sequences in relatively tight clothing. It has approximately 40,000 3D human body models.We manually segment the meshes of human subjects.

4.1.3 Cloth3D. The dataset [16] comprises of 6500 sequences of draped SMPL meshes simulated with MoCap data. Each frame of a sequence contains a garment and an SMPL body. We augment this data by capturing SMPLtexture maps with minimal clothing to simulate realistic body textures. For each sequence, five frames are randomly sampled and the naked body is subtracted from the garment. Thus, we obtain clothing occluded body as ground-truth for training. For semantic segmentation, we assign labels to clothing and manually label SMPL body. While we perform the peeling of the mesh as suggested in [40], we assign the cloth label if the naked body is occluded by cloth, else the semantic labels of SMPL are assigned to a pixel in the peeled segmentation map.

4.1.4 Buff. The dataset consists of 5 subjects with tight clothing performing simple motions. For generating segmentation labels we register an SMPL-D mesh with the original meshes. We have a segmented template SMPL mesh as prior, using this while performing peeling of the mesh we assign label to a particular vertex the label of its closest vertex in registered SMPL-D mesh. Figure 5 shows the sample mesh which is registered with SMPL and transferred labels.

4.2 Implementation Details

Our model is implemented in PyTorch and trained on 4 Nvidia RTX GPU's in parallel. In Cloth3D, BUFF and MonoPerfCap datasets, we keep subjects (which are not in training) for testing where we follow 80-20 training and test split. For THumans, we adopted test/train split from [44]. We initially train PeelGAN and freeze the network. Our segmentation network is trained for around 30 epochs with a batch size of 8 with an image resolution of 512x512 with w_g, w_{seg} set as 1 and 100. We freeze the segmentation network and pass its output to the depth refinement network. Our depth refinement network is trained for 50 epochs with a batch size of 8 and same image resolution with $w_g, w_l, w_r, w_c, w_{sm}$ as 1,500,500,100 and 10. Adam optimiser was used to train both networks, initialised with a learning rate of 0.0002, $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We train our segmentation network and depth refinement network in a sequential fashion. The ground-truth segmentation, depth and RGB maps are produced by performing ray tracing using Trimesh library. The reconstructed point-cloud has around 50,000 points.

4.3 Results

We performed both quantitative and qualitative evaluations on Thuman, MonoPerfCap, Buff and Cloth3D dataset. Figure 6 shows the qualitative results of our method on all four datasets where our method seems to perform well on variations in body shape, pose and (loose) clothing.

We also compared our method with other state-of-the-art (SOTA) methods. Table 1 report quantitative results using Chamfer distance

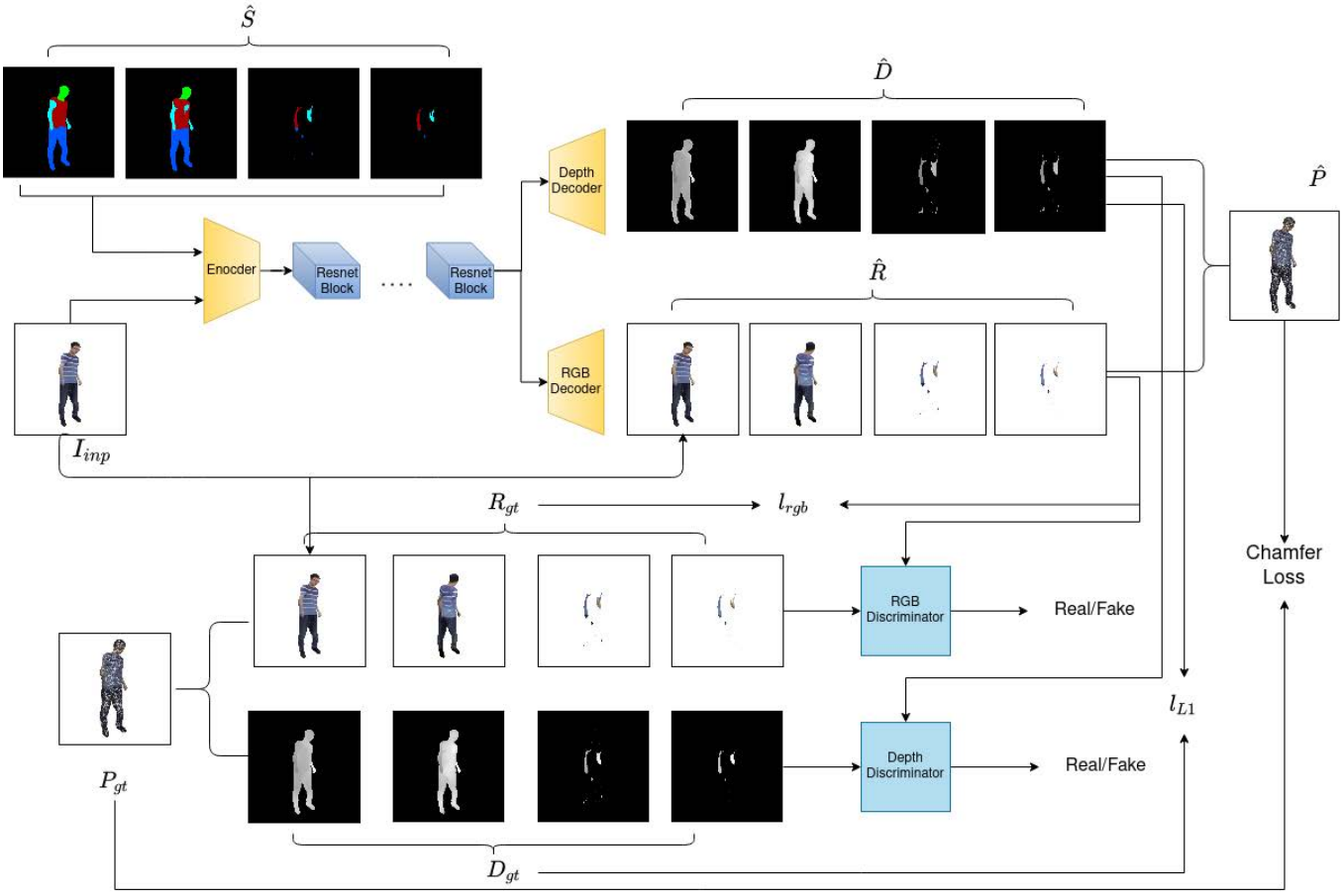

Figure 4: Architecture of the proposed RefGAN that predicts refined *Peeled Depth and RGB maps* given input monocular image and *Peeled Segmentation maps* predicted by PSegGAN.

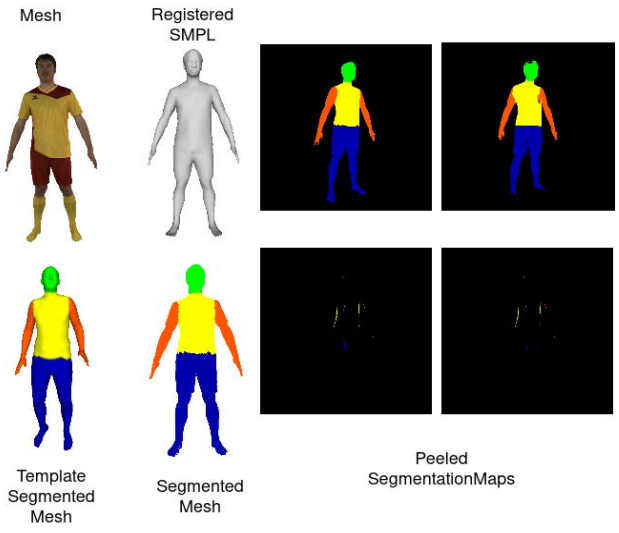

Figure 5: Generating semantic segmentation maps.

Table 1: Quantitative results on Thuman Dataset

Method	ChamferDistance
DeepHuman [51]	0.0011
PeeledHuman [40]	0.00051
GeoPiFu [44]	0.00017
Our Method	0.00037

on Thuman dataset. As we can see that our method achieves lower Chamfer distance on the test set as compared to DeepHuman [51] and PeeledHuman [40]. Although GeoPiFu [44] seems to perform superior in terms of Chamfer distance, we compare the qualitative results with some of the test samples where our method reconstructs more consistent 3D body models as compare to GeoPiFu, as shown in Figure 9. Also note that in Figure 1, we show the result on BUFF dataset where all the models are trained on THumans and tested on BUFF. This experiment proves the generalizability of our model when tested on other datasets. Additionally, GeoPIFu training takes 7 days on six GPUs while our training takes less than three days on four GPUs. Please refer to supplementary for more results.

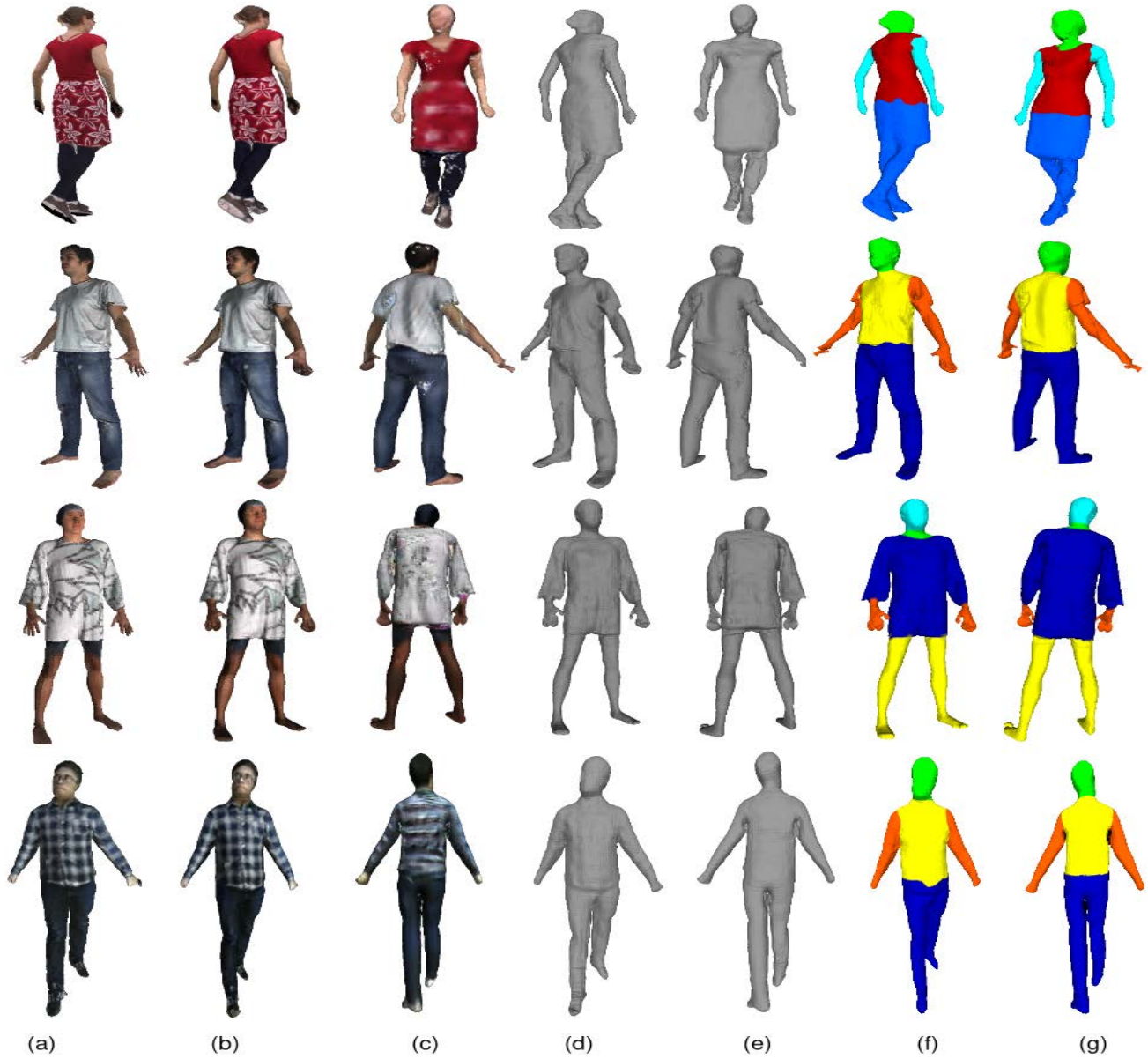

Figure 6: Qualitative results (meshes) of our method on MonoPerfCap, BUFF, Cloth3D and Thuman Datasets.(a) input RGB image, (b) & (c) predicted 3D output in different views, (d) & (e) 3D mesh without texture in different views, (f) & (g) our predicted semantic segmentation in different views.

We also compared with the most relevant PeeledHuman [40] method on Cloth3D and MonoPerfCap dataset where our method achieves lower Chamfer distance, as reported in Table 2. This implies that the PeeledHuman representation require more semantic context to reconstruct plausible shapes. Figure 8 shows the comparison with [40] where we render predicted point clouds (raw output) on all four datasets where our method consistently achieves superior quality reconstruction results. We are able to reconstruct plausible body shapes because of the inclusion of semantic context proposed in our method. Please refer to our supplementary material for video rendering of qualitative results.

We performed ablation study on the depth of network for RefGAN by varying the number of ResNet blocks. As reported in Table 3, we can conclude that a deeper network with more ResNet blocks helps improve the performance of our model on Cloth3D dataset. Additionally, we also performed ablation study of the effect of various loss terms on the final prediction. The effect of loss terms on the chamfer distance between prediction and ground truth has been reported in Table 4 and Figure 7.

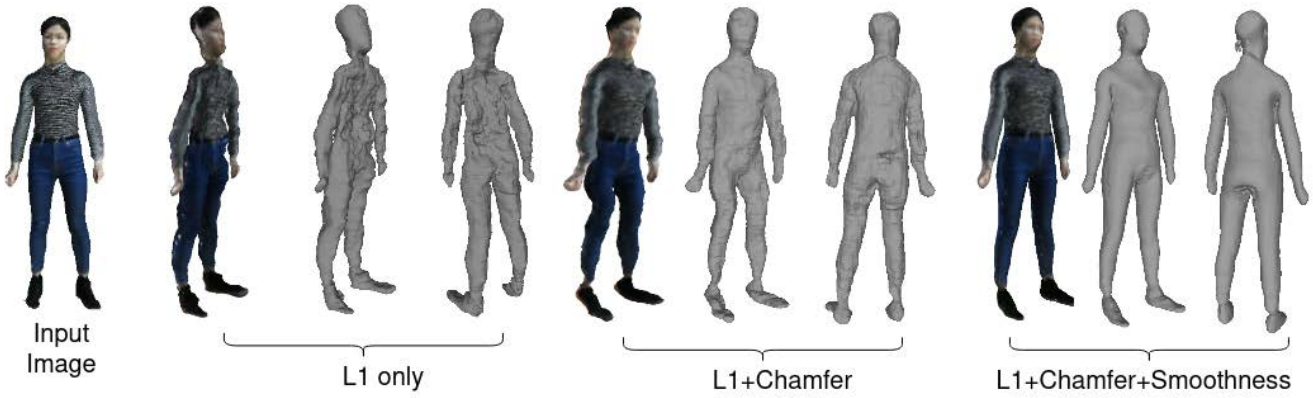

Figure 7: Qualitative visualization of ablative study on various loss terms.

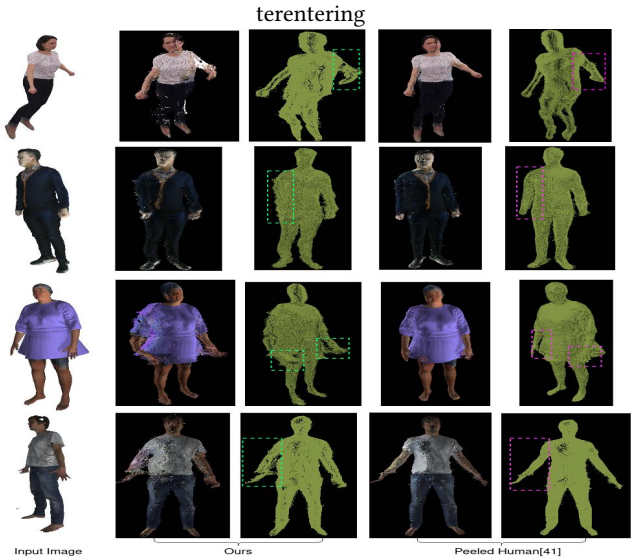

Figure 8: Qualitative comparison of generated point cloud (colored and without color) of our method (d,e) and Peeled-Human (b,c) [40].

Table 2: Quantitative results (Chamfer distance) of our method with [40]

Dataset	PeeledHuman	Our Method
Cloth3D	0.00147	0.00135
MonoPerfCap	0.00095	0.00086
Buff	0.00030	0.00025

5 CONCLUSION

We introduced *Peeled Segmentation map* representation in a coarse-to-fine refinement framework that reconstructs the 3D human body with clothing from a monocular image. Along with the reconstruction, we also provide 3D semantic segmentation of the human from

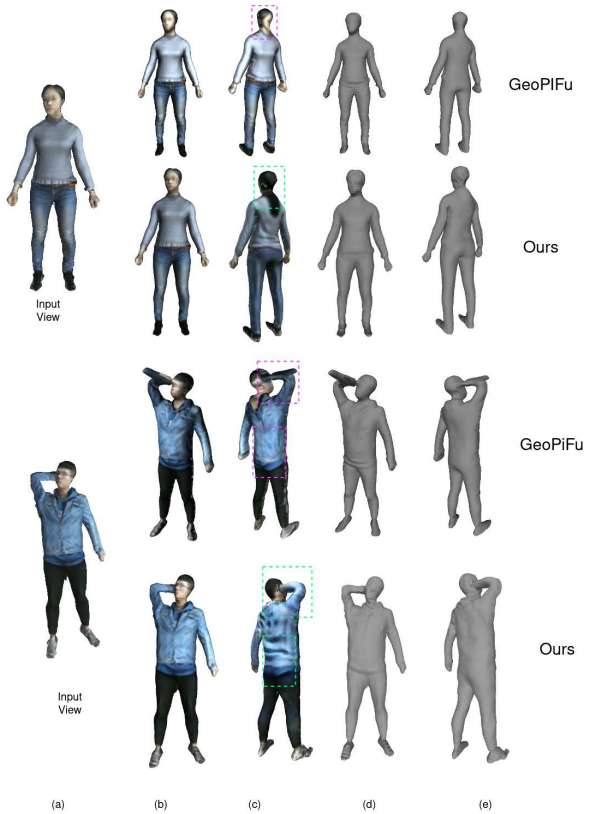

Figure 9: Comparison with GeoPIFu on THUman dataset

Table 3: Effect of varying number of ResNet blocks for Res-GAN.

ResNetBlocks	Chamfer Distance
6	0.00154
9	0.00141
18	0.00135

Table 4: Qualitaive results of ablation study performed on loss terms on the THumans Dataset

Loss function	Chamfer Distance
L1	0.000456
L1+Chamfer	0.000411
L1+Chamfer+smoothness	0.000370

monocular image. We use the semantic segmentation information as a prior to our refinement network which provides a global context of human shape. Our method performs well even in the case of partial occlusions. We evaluate our method on various datasets and report impressive results on par with state-of-the-art methods. As part of future work, we intend to extend the work for temporal reconstruction of human bodies in action.

REFERENCES

[1] D W Jacobs and J. Malik. A. Kanazawa, M. J Black. 2018. End-to-end recovery of human shape and pose. *CVPR*. (2018).

[2] C. Manafas A. S Jackson and G. Tzimiropoulos. 2018. 3d human body reconstruction from a single image via volumetric regression. *ECCV* (2018).

[3] S.S. Jinka A. Venkat and A. Sharma. 2018. Deep textured 3D reconstruction of human bodies. *BMVC* (2018).

[4] Xu W. Theobalt C. Pons-Moll G. Alldieck T., Magnor M. 2018. Detailed human avatars from monocular video. *3D Vision* (2018).

[5] Xu W. Theobalt C. Pons-Moll G. Alldieck T., Magnor M. 2018. Video based reconstruction of 3D people models. *CVPR* (2018).

[6] C. Theobalt B. L. Bhatnagar, G. Tiwari and G. Pons-Moll. 2019. Multi-Garment Net: Learning to dress 3D people from images. *ICCV* (2019).

[7] Lassner C. Gehler P. Romero J. Black M.J. Bogo F., Kanazawa A. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. *ECCV* (2016).

[8] Z. Chen and H. Zhang. 2019. Learning implicit fields for generative shape modeling. *CVPR* (2019).

[9] Levoy M. Curless B. 1996. A volumetric method for building complex models from range images. *Computer Graphics and Interactive Techniques, SIGGRAPH* (1996).

[10] Oztireli C. Ziegler R. Gross M. Dibra E., Jain H. 2017. Human shape from silhouettes using generative hks descriptors and cross-modal neural network. *CVPR* (2017).

[11] Daphne Koller Sebastian Thrun Jim Rodgers Dragomir Anguelov, Praveen Srinivasan and James Davis. 2005. Expressive body capture: 3D hands, face, and body from a single image. *In ACM Transactions on Graphics(ToG)* (2005).

[12] N. Ghorbani T. Bolkart A. A. Osman D. Tzionas G. Pavlakos, V. Choutas and M. J. Black. 2019. Expressive body capture:3D hands, face, and body from a single image. *CVPR* (2019).

[13] Nima Ghorbani T. Bolkart AA Osman Dimitrios Tzionas G. Pavlakos, V. Choutas and Michael J Black. 2019. Expressive body capture: 3D hands, face, and body from a single image. *CVPR* (2019).

[14] B. Russell J. Yang E. Yumer I. Laptev G. Varol, D. Ceylan and C. Schmid. 2018. BodyNet Volumetric inference of 3D human body shapes. *ECCV*. (2018).

[15] X. Martin N. Mahmood M. J. Black I. Laptev G. Varol, J. Romero and C. Schmid. 2017. Learning from synthetic humans. *CVPR* (2017).

[16] M. Madadi H. Bertiche and S. Escalera. 2020. CLOTH3D: Clothed 3d humans. *ECCV* (2020).

[17] Zollhofe M. Pons-Mol G. Theobal C. Habermann M., Xu W. 2019. Livecap: Real-time human performance capture from monocular video. *In ACM Transactions on Graphics(ToG)* (2019).

[18] J. Straub-R. Newcombe J. Joon Park, P. Florence and S. Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. *CVPR* (2019).

[19] Sai Sagar Jinka, Rohan Chacko, Astitva Srivastava, Avinash Sharma, and P. J. Narayanan. 2021. SHARP: Shape-Aware Reconstruction of People In Loose Clothing. *CoRR* abs/2106.04778 (2021). arXiv:2106.04778 https://arxiv.org/abs/2106.04778

[20] Sheikh Y. Joo H., Simon T. 2018. Total capture: A 3d deformation model for tracking faces, hands, and bodies. *CVPR*. (2018).

[21] P. Davidson J. Busch X. Yu M. Whalen G. Harvey S. Orts-Escolano R. Pandey J. Dourgarian et al. K. Guo, P. Lincoln. 2019. The relightables: volumetric performance capture of humans with realistic relighting. *In ACM Transactions on Graphics(ToG)* (2019).

[22] Zhang J.Y. Felsen P. Malik J. Kanazawa, A. 2019. Learning 3d human dynamics from video. *CVPR* (2019).

[23] M. Kazhdan, M. Bolitho, and M. Hoppe. 2006. Poisson surface reconstruction. In *SGP*, Vol. 7.

[24] Pons-Moll G . Keyang Z., Bhatnagar B.L. 2020. Unsupervised shape and pose disentanglement for 3d meshes. *ECCV* (2020).

[25] M. Niemeyer S. Nowozin L. Mescheder, M. Oechsle and A. Geiger. 2019. Occupancy networks: Learning 3d reconstruction in function space. *CVPR* (2019).

[26] Kiefel M. Bogo F. Black M.J. Gehler P.V. Lassner C., Romero J. 2017. Unite the people: Closing the loop between 3d and 2d human representation. *CVPR*. (2017).

[27] W. E Lorensen and H. E Cline. 1987. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics* (1987).

[28] J. Romero G. Pons-Moll M. Loper, N. Mahmood and M. J.Black. 2015. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (ToG)* (2015).

[29] M. Oechsle M. Niemeyer, L. Mescheder and A. Geiger. 2020. Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision. *CVPR* (2020).

[30] G. Pons-Moll P. Gehler M. Omran, C. Lassner and B. Schiele. 2018. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. *3D Vision* (2018).

[31] Seitz S.M. Newcombe R.A., Fox D. 2015. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. *CVPR* (2015).

[32] Seitz S.M. Newcombe R.A., Fox D. 2017. Killingfusion: Non-rigid 3d reconstruction without correspondences. *CVPR* (2017).

[33] T. Zhou P. Isola, J.-Y. Zhu and A. A. Efros. 2017. Image-to-image translation with conditional adversarial networks. *CVPR*, (2017).

[34] Zhou X. Daniilidis K. Pavlakos G., Zhu L. 2018. Learning to estimate 3d human pose and shape from a single color image. *CVPR*. (2018).

[35] Hu S. Black M.J. Pons-Moll G., Pujades S. 2017. Clothcap: Seamless 4d clothing capture and retargeting. *In ACM Transactions on Graphics(ToG)*. (2017).

[36] D. Ceylan R. Mech Q. Xu, W. Wang and U. Neumann. 2019. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *NeurIPS* (2019).

[37] Z. Huang W. Chen C. Ma H. Li R. Natsume, S. Saito and S. Morishima. 2019. SiCloPe: Silhouette-based clothed people. *CVPR* (2019).

[38] W. Chen S. Liu, S. Saito and H. Li. 2019. Learning to infer implicit surfaces without 3D supervision. *NeurIPS* (2019).

[39] T. Simon S. Lombardi, J. Saragih and Y. Sheikh. 2018. Deep appearance models for face rendering. *In ACM Transactions on Graphics(ToG)* (2018).

[40] A. Sharma S. S. Jinka, R. Chacko and PJ Narayanan. 2020. PeeledHuman: Robust shape representation for textured 3D human body reconstruction. *3DVision* (2020).

[41] J. Saragih S. Saito, S. Simon and H. Joo. 2020. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. *CVPR* (2020).

[42] R. Natsume S. Morishima A. Kanazawa S. Saito, Z. Huang and H. Li. 2019. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. *ICCV* (2019).

[43] C. Theobalt T. Alldieck, G. Pons-Moll and M. Magnor. 2019. Tex2Shape: Detailed full human body geometry from a single image. *ICCV* (2019).

[44] H. Jin T. He, J. Collomosse and S. Soatto. 2020. Geo-PIFu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *NeurIPS* (2020).

[45] Bharat Lal Bhatnagar Christian Theobalt Thiemo Alldieck, Marcus Magnor and Gerard Pons-Moll. 2019. Learning to reconstruct people in clothing from a single RGB camera. *CVPR* (2019).

[46] X. Martin C. Schmid V. Gabeur, J.-S. Franco and G. Rogez. 2019. Moulding humans: Non-parametric 3D human shape estimation from single images. *ICCV* (2019).

[47] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. 2017. Learning from Synthetic Humans. In *CVPR*.

[48] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, Dushyant M., H. Seidel, and C. Theobalt. 2018. MonoPerfCap: Human Performance Capture From Monocular Video. *ACM Trans. Graph.* 37, 2, Article 27 (2018), 15 pages.

[49] Zanfir A. Freeman W.T. Sukthankar R. Sminchisescu C. Xu H., Bazavan E.G. 2020. Ghum : Generative 3d human shape and articulated pose models. *CVPR* (2020).

[50] W. Chen Y. Zhao J. Xing C. LeGendre L Luo C. Ma Z. Huang, T. Li and H. Li. 2018. Deep volumetric video from very sparse multi-view performance capture. *ECCV* (2018).

[51] Y. Wei Q. Dai Z. Zheng, T. Yu and Y. Liu. 2019. DeepHuman: 3D human reconstruction from a single image. *ICCV* (2019).

[52] Christoph Lassner Hao Li Zeng Huang, Yuanlu Xu and Tony Tung. 2020. ARCH: Animatable reconstruction of clothed humans. *CVPR* (2020).

[53] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. 2019. DeepHuman: 3D Human Reconstruction from a Single Image. In *ICCV*.

[54] Wang S. Cao X. Yang R. Zhu H., Zuo X. 2019. Detailed human shape estimation from a single image by hierarchical mesh deformation. *CVPR* (2019).