Single-view 3D Body and Cloth Reconstruction under Complex Poses

Nicolas Ugrinovic[a], Albert Pumarola[b], Alberto Sanfeliu[c] and Francesc Moreno-Noguer[d]
Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain

Keywords: 3D Human Reconstruction, Augmented/virtual Really, Deep Networks.

Abstract: Recent advances in 3D human shape reconstruction from single images have shown impressive results, leveraging on deep networks that model the so-called implicit function to learn the occupancy status of arbitrarily dense 3D points in space. However, while current algorithms based on this paradigm, like PiFuHD (Saito et al., 2020), are able to estimate accurate geometry of the human shape and clothes, they require high-resolution input images and are not able to capture complex body poses. Most training and evaluation is performed on 1k-resolution images of humans standing in front of the camera under neutral body poses. In this paper, we leverage publicly available data to extend existing implicit function-based models to deal with images of humans that can have arbitrary poses and self-occluded limbs. We argue that the representation power of the implicit function is not sufficient to simultaneously model details of the geometry and of the body pose. We, therefore, propose a coarse-to-fine approach in which we first learn an implicit function that maps the input image to a 3D body shape with a low level of detail, but which correctly fits the underlying human pose, despite its complexity. We then learn a displacement map, conditioned on the smoothed surface and on the input image, which encodes the high-frequency details of the clothes and body. In the experimental section, we show that this coarse-to-fine strategy represents a very good trade-off between shape detail and pose correctness, comparing favorably to the most recent state-of-the-art approaches. Our code will be made publicly available.

1 INTRODUCTION

While the 3D reconstruction of the human pose (Martinez et al., 2017; Moreno-Noguer, 2017; Pavlakos et al., 2017; Rogez et al., 2019; Mehta et al., 2018; Kinauer et al., 2018) and shape of the naked body (Kanazawa et al., 2017; Pavlakos et al., 2018; Varol et al., 2018; Varol et al., 2017) from single images has been extensively studied over the past few years and led to very accurate results, doing this with clothed humans remains a difficult challenge. There exist recent works that provide very good body and cloth reconstructions, but are methods limited to mild human poses, typically standing up in front of the camera (Saito et al., 2020; Saito et al., 2019; Natsume et al., 2019; Alldieck et al., 2019b; Jackson et al., 2018). A challenge that still remains open is thus to capture diverse poses while maintaining a detailed geometry of clothes and body.

PiFu (Saito et al., 2019) and very recently (Saito et al., 2020) are the most relevant works on clothed human reconstruction, and builds upon the representation capacity of implicit functions, shown to be very effective for estimating the geometry of rigid 3D objects (Mescheder et al., 2019; Chen and Zhang, 2019; Xu et al., 2019). PiFu learns a per-pixel feature vector aligned with the 3D surface to get an implicit function based on local information. However, while this strategy provides a lot of detail, it cannot generalize to arbitrary human poses.

Other works are able to capture diverse poses but lack details of human clothing (Genova et al., 2020). There exist methods that do not use implicit functions, but introduce an additional step to the estimation of a parametric naked body model. For instance, (Alldieck et al., 2019b) learns a displacement map over the SMPL model (Loper et al., 2015), although, this approach is also limited to a small range of body poses and it needs high-quality 1024×1024 input images.

In this paper, we use implicit functions and propose an approach that, given a single image, is able to predict detailed meshes of clothed 3D humans for a wide range of poses and can work with but it is not

[a] https://orcid.org/0000-0002-1823-3780
[b] https://orcid.org/0000-0003-4185-6991
[c] https://orcid.org/0000-0003-3868-9678
[d] https://orcid.org/0000-0002-8640-684X

Ugrinovic, N., Pumarola, A., Sanfeliu, A. and Moreno-Noguer, F.
Single-view 3D Body and Cloth Reconstruction under Complex Poses.
DOI: 10.5220/0010896100003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 4: VISAPP*, pages 192-203
ISBN: 978-989-758-555-5; ISSN: 2184-4321

limited to 224×224 input images.

We argue that one of the reasons why (Saito et al., 2019; Saito et al., 2020) does not generalize well to difficult poses is that it strongly relies on local pixel features to guide the reconstruction and, thus, has no awareness of the overall topology of the mesh and therefore struggle to model unseen parts of the body.

To address this, we exploit global image features and alleviate their inherent lack in details using two strategies: First, we introduce a coarse-to-fine architecture with two modules, one building on an implicit function and global features that learns a coarse 3D shape, but with a correct body pose; and another network that learns a displacement map to add extra detail (see Fig. 1). Second, we take into account the structure of the human body by including 2D joints as inputs of our system. This enables to have overall mesh consistency and retain the details of body and clothing in complex poses.

We quantitatively evaluate our method on synthetic data and qualitatively on real and synthetic images and demonstrate that our approach can capture a wide range of poses better than previous state-of-the-art methods based on implicit functions. Thus, we claim that global reasoning combined with a refinement step leads to coherent human meshes with no disconnected body parts, even in difficult poses, while maintaining a good level of detail.

2 RELATED WORK

Single-view 3D Reconstruction of Rigid Objects. is a well studied topic in computer vision and computer graphics. The works in this realm can be mainly categorized by the representation they use, whether it is a voxel grid (Choy et al., 2016; Tulsiani et al., 2017; Wu et al., 2017), pointcloud (Pumarola et al., 2020; Fan et al., 2016), mesh (Wang et al., 2018; Gkioxari et al., 2019) or implicit function (Mescheder et al., 2019). Voxels usually require extensive memory and are time consuming to train while usually leading to reconstructions with very restricted resolution. Pointclouds require additional non-trivial post processing steps to generate the final mesh. (Wang et al., 2018; Gkioxari et al., 2019) directly work on the mesh using a graph based CNN (Scarselli et al., 2008), although they are only able to generate overly smoothed meshes with simple topology which can be genus-0 only. In contrast, we choose to work with implicit function representation due to the well known fact that they require relatively simple architectures and have the ability to obtain a greater level of detail without requiring vast amounts of memory.

Several works (Mescheder et al., 2019; Park et al., 2019; Xu et al., 2019; Chen and Zhang, 2019) have shown that implicit functions can be learned by means of deep neural networks, and it is possible to get high resolution reconstruction by applying the marching cubes (MC) algorithm. Most recent approaches for image 3D reconstruction use implicit functions. For example, (Mescheder et al., 2019) conditions the learning of occupancy probabilities to an input image, being able to reconstruct a high resolution mesh. However, they rely solely on global image features which hinders the model to learn high frequency details. We, instead, use local information about the joints and learn a displacement map to improve the reconstruction details as a result of the MC algorithm. (Chen and Zhang, 2019) also uses global features suffering from the same lack of detail needed to capture clothed humans.

Single-View 3D Human Reconstruction. While the problem of localizing the 3D position of the joints from a single image has been extensively studied (Martinez et al., 2017; Moreno-Noguer, 2017; Rogez et al., 2019; Moon et al., 2019; Mehta et al., 2018) 3D human body shape reconstruction still remains an open problem. Single-view human reconstruction requires strong priors due to the inherent ambiguity of the problem. This has been addressed by using parametric models learned from body scan repositories such as SCAPE (Anguelov et al., 2005) and SMPL (Loper et al., 2015) to represent the human body geometry by a reduced number of parameters. These parameters are then optimized to match image characteristics. For example, methods that use deep neural networks input additional information such as silhouettes (Dibra et al., 2017; Pavlakos et al., 2018) and other types of manual annotations (Lassner et al., 2017; Omran et al., 2018). Furthermore, (Vince Tan and Cipolla, 2017) uses a differential renderer along with a deep neural network to predict SMPL body parameters by directly estimating and minimizing the error of image features. Despite the usefulness of parametric models, they can only reproduce the geometry of the naked human body.

Monocular reconstruction of cloth geometry has been traditionally addressed under the Shape-from-Template (SfT) paradigm (Moreno-Noguer and Fua, 2013; Sanchez et al., 2010; Moreno-Noguer and Porta, 2011; Agudo et al., 2016), requiring 3D-to-2D point correspondences between a template mesh and the input. More recently (Pumarola et al., 2018) introduced a deep network which alleviated the need for estimating correspondences. In any event, the clothes reconstructed by these approaches, were focused to simple rectangular-like shapes, and were not applica-

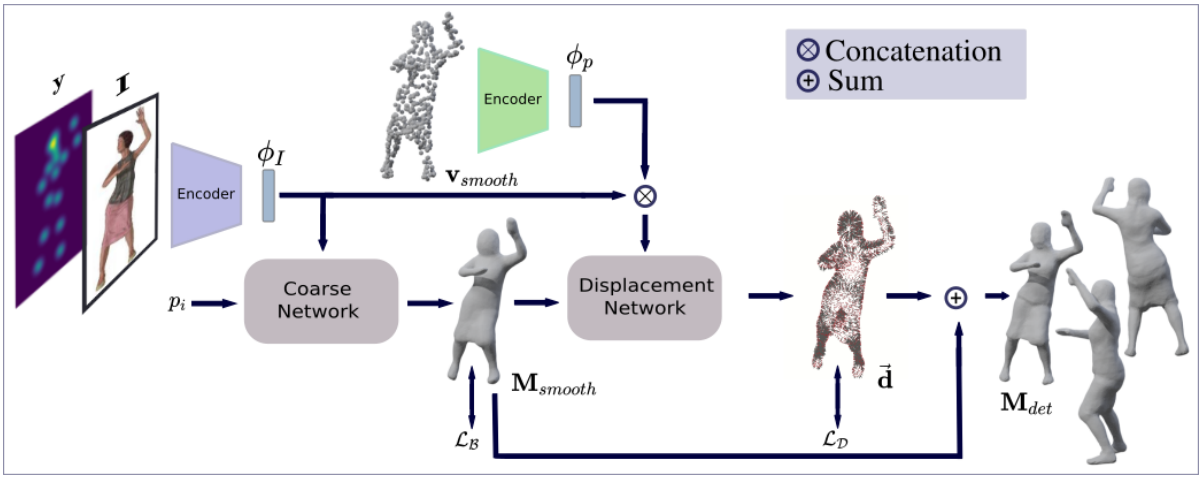

Figure 1: Overview of our pipeline to reconstruct clothed people under complex poses. Given an input RGB image, an implicit function-based network initially predicts a smoothed version of the geometry, but with an accurate body pose. The fine details of the mesh are recovered by a second network that computes a displacement field over the smooth mesh.

ble to reconstruct the shape of the garments worn by humans.

To overcome this limitation, (Alldieck et al., 2019b) proposes to learn a displacement map on top of the SMPL body model and is able to represent certain type of clothing, short hair details and hands details. However, it fails for more complex topologies such as dresses and skirts and it is limited to mild human body poses (people standing in front of the camera and looking at it). Also others use displacement maps for this purpose (Zhu et al., 2019; Onizuka et al., 2020), although mostly from videos or few image frames (Alldieck et al., 2018; Alldieck et al., 2019a). In this paper, while we also learn a displacement map, we are capable of capturing dresses and skirts while including a large diversity of body pose.

To address the limitations of parametric models, template-free methods have been used, some based on voxel representations (Varol et al., 2018; Zheng et al., 2019; Jackson et al., 2018), others based on different representations (Pumarola et al., 2019; Saito et al., 2019). BodyNet (Varol et al., 2018) infers the volumetric body shape, although, due to resolution constrains and the use of SMPL as a final fitting, it cannot recover clothing geometry. DeepHuman (Zheng et al., 2019) uses a volume-to-volume translation approach showing impressive results to capture pose and certain type of clothing, but it fails to correctly capture complex cloth geometry such as skirts and also suffers from high memory requirements of voxel representation, limiting its resolution and requiring and initial estimation of template-based model SMPL. To tackle the resolution limitation of voxels, GimNet (Pumarola et al., 2019) uses geometry images to represent the body shape and is able to capture complex poses and geometries such as dresses, although with a lack of details. Finally, PIFu (Saito et al., 2019) and PIFuHD (Saito et al., 2020) use implicit function representation which is memory efficient and results in impressive level of details even for complex cloth geometries and accessories. However, this approach can not generalize to arbitrary human poses. We also use an implicit function representation, but in contrast to previous approaches we are able to capture a large range of arbitrary poses. This is made possible thanks to a first module of our model, which is general enough and reasons in a global manner generating realistic human meshes.

Finally, most similar in concept but very different implementation from our work, (He et al., 2020) demonstrate that in order to have a better detailed reconstruction, it is first necessary to have a solid geometric prior which can be learned from a coarse voxel representation of the human body.

3D Datasets. Even though 3D reconstruction has become a popular topic in the field, there are very few publicly available datasets that contain 3D information of human body. Obtaining the 3D body shape is a complex task that requires vast amounts of effort. BUFF dataset (Zhang et al., 2017) is one of the few that contains high-quality 3D scans, nevertheless, it only includes 6 different subjects and although it has a good human body pose variation, only captures restricted actions. As an alternative, datasets with synthetically photo-realistic images have appeared in the scene (Varol et al., 2017; Pumarola et al., 2019). SURREAL (Varol et al., 2017) is the largest dataset, containing 6 million frames generated by projecting synthetic textures of clothes onto random SMPL body

shapes. However, given that clothes are projected onto a naked body model, they are only textures and have no shape of their own, making it impossible to learn clothing details from this dataset. On the contrary, 3DPeople (Pumarola et al., 2019) contains models of 80 different 3D dressed subjects that perform 70 actions and 2.5 million photorealistic rendered images in which every action sequence is captured by from 4 camera views. For this work we use 3DPeople dataset. Most recently (Caliskan et al., 2020) announced a similar dataset containing images of synthetic humans and their corresponding 3D human mesh annotations. We don't use this dataset, however, because it has not yet been made public.

3 METHOD

3.1 Problem Formulation

We aim to solve the problem of single image 3D reconstruction applied to human bodies with clothing. Our goal is to make sure that not only the inferred pose of the mesh representing the person is correct but also that we recover geometry details of the clothing.

Let $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ be an input RGB image of a single clothed person at an arbitrary pose. Our aim is to learn a mapping $\mathcal{M}$ to reconstruct the mesh $\mathbf{M}$ which is a detailed 3D representation of the clothed body of the person. We represent $\mathbf{M}$ as a mesh with N vertices v_i, where $v_i = (x_i, y_i, z_i)$ are the 3D coordinates of each vertex that explains the body of the person in the image, taking into account the body shape, pose and clothing details. We train $\mathcal{M}$ in a supervised manner.

3.2 Network Architecture

We next describe our network to generate detailed meshes under complex poses from a single image. Given the high complexity of the task, we use a coarse-to-fine approach and divide our method into two main modules, as shown in Fig. 1.

The first module, denoted *coarse network*, outputs a smoothed mesh $\mathbf{M}_{smooth}$ provided an input set of query points $\mathbf{p}$ and an observation of the 3D object, the image $\mathbf{I}$. This mesh *intentionally* lacks the level of detail we are looking for but it is enforced to accurately fit the body pose.

The second module, which we call *displacement network*, adds details to the mesh by estimating vertex displacement $\vec{d_i}$ over the direction of the normal vector $\vec{n_i}$ for each vertex v_i of $\mathbf{M}_{smooth}$, yielding to $\mathbf{M}_{det}$. For this, we learn a network that takes as inputs $\mathbf{I}$ and a set of vertices randomly sampled from $\mathbf{M}_{smooth}$, which we shall denote $\mathbf{v}_{smooth}$.

It is worth noting that, as an additional input to guide the learning of both networks, we use the 2D joints of the person in $\mathbf{I}$. Next, we explain both networks in detail.

3.2.1 Coarse Network

Given the input image $\mathbf{I}$, we use J ground truth 2D body joint locations and represent them as heatmaps $\mathbf{y} \in \mathbb{R}^{H \times W \times J}$. We use $J = 17$ body joints. This joint representation is then concatenated with $\mathbf{I}$ and fed into the network. Additionally, the network has as input a set of query points in the 3D space $\mathbf{p}_{xyz} = \{p_i\}_{i=1}^{K}$. Our goal is to learn the occupancy probability for each p_i given $\mathbf{I}$ and $\mathbf{y}$. Formally, we seek to estimate the mapping:

$$\mathcal{M} : \mathbf{I} \oplus \mathbf{y}, p_i \rightarrow [0,1] \quad (1)$$

This mapping takes the form of an implicit function and can be learned by a neural network $f_{\theta_s}(p_i, \mathbf{I}, \mathbf{y})$. Estimating $\mathcal{M}$ to account for high frequency details is, however, significantly challenging for the network, and indeed we found out that training this network to learn details resulted in meshes with incorrect body poses. For this reason, we force it to learn a smoothed version of the occupancy field of the ground truth mesh, hence its name *coarse network*. To enforce this, instead of using the detailed mesh as ground truth, we train this network with a pseudo ground truth that results from applying Laplacian smoothing (Sorkine et al., 2004).

Finally, at inference, to recover the mesh we first evaluate $f_{\theta_s}(p, \mathbf{I}, \mathbf{y})$ for all p of a discretized volumetric space. We then use an octree based algorithm MISE (Mescheder et al., 2019) and mark each p as occupied if $f_{\theta_s}(p, \mathbf{I}, \mathbf{y})$ is bigger or equal than some threshold τ. After the evaluation is complete, we apply the MC algorithm (Lorensen and Cline, 1987) to extract and approximate isosurface and estimate the faces topology of $\mathbf{M}_{smooth}$. Note that although we intentionally train f_{θ_s} to produce a smooth mesh, the body pose is expected to be correct. Also note that we build on (Mescheder et al., 2019) and, therefore, follow their formulation, however, any other reconstruction model could be used instead.

3.2.2 Displacement Network

This network has a similar architecture as the previous one with two main differences: instead of estimating occupancy probability, it regresses the magnitude for displacements $\vec{d_i}$ and takes an additional conditioning value that also serves as a query input, the vertices

$\mathbf{v}_{smooth}$. In the same fashion as before, we learn a new encoding for the image and joints representation ϕ_I but we use $\mathbf{v}_{smooth}$ to generate a point encoding ϕ_p and concatenate this to ϕ_I. This way we are able to condition the learning of the displacements on $\mathbf{I}$, $\mathbf{y}$ and $\mathbf{M}_{smooth}$. This network, denoted $h_{\theta_d}(p,\mathbf{I},\mathbf{y},\mathbf{v})$, regresses the magnitude of the displacement $\vec{d}_i$ which is then applied to $\mathbf{v}_{smooth}$ in the direction of the normals $\vec{n}_i$ of $\mathbf{M}_{smooth}$. This reduces the complexity of the problem by forcing the regressor to learn only a scalar value and not a 3-dimensional vector, helping the network to learn the proper displacements.

The final result is obtained by adding the learned displacements to the vertices estimated by the first module:

$$\mathbf{v}_{det} = \mathbf{v}_{smooth} + \vec{\mathbf{d}}\,, \tag{2}$$

where $\mathbf{v}_{det}$ are the vertices that correspond to $\mathbf{M}_{det}$ and share the same faces as $\mathbf{M}_{smooth}$ and $\vec{\mathbf{d}}$ is the estimated displacement over the direction of the normal vector.

Finally, at inference, to obtain the detailed mesh $\mathbf{M}_{det}$ we first evaluate $h_{\theta_d}(p,\mathbf{I},\mathbf{y},\mathbf{v})$, that in this case are all vertices $\mathbf{v}_{smooth}$. Then, using equation 2 we get the detail vertices for $\mathbf{M}_{det}$.

3.3 Learning the Model

3.3.1 Smooth Reconstruction Loss

To learn the parameters θ_s of the neural network $f_{\theta_s}(p,\mathbf{I},\mathbf{y})$, we randomly sample points in the 3D bounding volume of the mesh representing the person. We sample these points in three ways: *(a)* uniformly over the bounding volume, *(b)* densely over the face and hands, and *(c)* densely over the surface. For b and c we sample several points (much more than a) near the surface of the mesh, that is why we say it is a dense sampling. To automatically obtain sampling points for face and hands we only sample points within a radius r of a sphere centered at the 3D joints corresponding to hands and face. We found that hands and face require higher level of detail to be better reconstructed than feet, hence, we do not include sampling specifically corresponding to feet. For each sample image i in a training batch we sample K points $p_{ij} \in \mathbb{R}^3, j = 1,...,K$. The minibatch loss $\mathcal{L}_B$ is then is evaluated at those locations:

$$\mathcal{L}_{\mathcal{B}}(\theta_s) = \frac{1}{B}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{K}\mathcal{L}(f_{\theta_s}(p_{ij},\mathbf{I},\mathbf{y}),o_{ij})\,, \tag{3}$$

where $o_{ij} \equiv o(p_{ij})$ denotes the true occupancy at point p_{ij}, and $|\mathcal{B}|$ is the minibatch size. The loss $\mathcal{L}(\cdot,\cdot)$, different from (Mescheder et al., 2019), is a weighted binary cross-entropy (wBCE) classification loss that takes into account the unbalanced number of points that lay inside the mesh in contrast to those that are outside. This avoids losing important body parts, especially the limbs, when extracting the mesh.

In a similar fashion as in (Mescheder et al., 2019) we also introduce a generative loss that helps us capture the rich distribution of complex clothing. We do this by adding an encoder network $g_\psi(\cdot)$ that takes as inputs the points and occupancies to predict the mean u_ψ and standard deviation σ_ψ of a Gaussian distribution $q_\psi(z|(p_{ij},o_{ij})_{j=1:K})$ on a latent space $z \in \mathbb{R}^L$ as output and then optimizing the KL divergence. This way, the new loss becomes:

$$\begin{aligned}\mathcal{L}_{\mathcal{B}}(\theta) = &\frac{1}{B}\sum_{i=1}^{|\mathcal{B}|}[\sum_{i=j}^{K}\mathcal{L}(f_{\theta_s}(p_{ij},\mathbf{I},\mathbf{y}),o_{ij}) + \\ &KL(q_\psi(z|(p_{ij},o_{ij})_{j=1:K})||p_0(z))]\end{aligned} \tag{4}$$

where $p_0(z)$ is a prior distribution on the latent variable z_i and z_i is sampled according to $q_\psi(z|(p_{ij},o_{ij})_{j=1:K})$. We train this as a conditional variational autoencoder (Sohn et al., 2015).

To generate $\mathbf{M}_{smooth}$ we use a hierarchical isosurface extraction algorithm (Mescheder et al., 2019), that incrementally builds an octree to efficiently obtain a high resolution mesh, that is then forwarded to the second stage of our method.

3.3.2 Displacement Loss

In order to learn the parameters θ_d of the neural network $h_{\theta_d}(p,\mathbf{I},\mathbf{y},\mathbf{v})$, in a similar manner as we did with the coarse network, we randomly sample N points p_{ij} from $\mathbf{v}_{smooth}$ and evaluate the minibatch loss $\mathcal{L}_{\mathcal{B}}$. Yet, instead of using a wCBE loss, we use an L2 loss:

$$\mathcal{L}_{\mathcal{D}}(\theta_d) = \frac{1}{B}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{N}\|f_{\theta_d}(p_{ij},\mathbf{I},\mathbf{y},\mathbf{v}_{smooth}) - d_{ij}\|_2 \tag{5}$$

4 IMPLEMENTATION DETAILS

Our model builds upon the ONet network architecture (Mescheder et al., 2019). For the *coarse network* $f_{\theta_s}(p,\mathbf{I},\mathbf{y})$ we use 5 ResNet blocks (He et al., 2016) which are conditioned on the input using conditional batch normalization (Ioffe and Szegedy, 2015). For the image and joint encoding we use a ResNet18 architecture.

For the *displacement network* we modify the architecture by adding 5 ResNet blocks (yielding to

Table 1: Quantitative evaluation on 3DPeople. Numerical comparison of our approach with other methods that use implicit functions retrained with the same data as ours. We measure IoU, Chamfer distance, Normal Consistency and Point to Surface (see main text) to validate the different components of our model. ↑: higher the better. ↓: lower the better.

Method	IoU ↑	Chamfer ↓	Normal Consistency ↑	P2S ↓
ONet	0.516	0.280	0.793	18.135
PiFu	0.244	1.550	0.601	70.200
Ours	**0.610**	**0.100**	**0.821**	**16.200**

a total of 10 blocks) and changing the last layer to regress the displacement value. We observed that for less amount of layers, the network is not able to capture the complexities of clothes and other details. It is important for the network to understand the 3D structure of the body in order to regress the desired displacements, for this reason we also modify the conditioning input of the architecture to be able to include mesh vertices as a prior. For this we use a similar encoder as in PointNet (Qi et al., 2016) and for the network we use 10 ResNet blocks. We plan to release our code.

The model is trained with 60,000 synthetic images of cropped clothed people resized to 224 x 224 pixels as needed by the image encoder, however, this resolution could be easily changed. These images correspond to 15,000 different meshes of varying number of vertices taken from the 3DPeople dataset (Pumarola et al., 2019) and projected to 4 camera views. We use 44 subjects out of 80 to reduce training time.

In order to train f_{θ_s} we generate occupancy annotations, i.e determine which points lie in the interior of the mesh. This step requires a watertight mesh. To do this we use code provided by (Mescheder et al., 2019). We train the *coarse network* during 645 epochs, K=2048 and Adam (Kingma and Ba, 2014) optimizer with initial learning rate of $1e-4$, beta1 0.9, beta2 0.999. For weighted-BCE we use a positive weight of 25. For reconstructing the mesh, we use a threshold parameter τ=0.96 for all cases. For this network to better capture complex poses, we first normalize each mesh w.r.t. three points: hips, upper left leg and upper right leg.

To train $h_{\theta_d}(p, \mathbf{I}, \mathbf{y}, \mathbf{v})$ we generate ground truth data using the results obtained from our *coarse network* and compute the displacement over the normal by first densely sampling the surface of the ground truth mesh and then finding the distance over the normal direction from a mesh vertex to the nearest point in the ground truth mesh. We train during 1700 epochs with batch size 14, K=2,048 and N=10,000. As for the optimizer we use Adam (Kingma and Ba, 2014) with initial learning rate of $1e-4$, beta1 as 0.9, beta2 as 0.999. At epoch 170 we change the learning rate to $1e-5$ and, again, at epoch 1,200 to $1e-6$.

5 EXPERIMENTAL EVALUATION

This section provides an evaluation of our proposed method. We present quantitative and qualitative results on synthetic images from 3DPeople (Pumarola et al., 2019) and qualitative results on images in the wild. We evaluate our approach on 3,200 images randomly chosen for 5 subjects (2 female/ 3 male) from (Pumarola et al., 2019).

We compare our approach quantitatively (see Table 1) with other two prominent implicit function models for 3D reconstruction, namely, OccupancyNets (ONets) (Mescheder et al., 2019) and PiFu (Saito et al., 2019). Note that to ensure fair comparison both a re-trained with the same training data as our model and we test all models with the same test set as ours. Although one could argue that numerical comparison with SOTA should include PIFuHD (Saito et al., 2020), this was not possible as the authors have not released the training code. However, we believe that the methods in question are good representatives of powerful implicit function models for 3D reconstruction. In this sense, being ONet a good candidate for global consistency models and PIFu for hi-detail local consistent models. Qualitative comparison with both these methods on synthetic images can be found in Fig. 4.

Additionally, in Table 2 we present a quantitative ablation study to validate all the components propose in this paper and used by our final method. The table compares our method and several baselines built upon the Occupancy Net (Mescheder et al., 2019) and the losses we have defined in our system. Table 2 reports the errors for all methods and shows that our approach consistently improves all baselines. Also, notice how the addition of the wBCE and KL losses over the ONet baseline, gracefully reduce the errors.

As evaluation metrics we use volumetric IoU, Chamfer distance (CD), normal consistency score and point to surface score (P2S). Volumetric IoU is defined as the quotient of the volume of the two meshes union and the volume of their intersection. We use the same procedure as in (Mescheder et al., 2019) to obtain this value. We calculate the CD by randomly sampling 100,000 points from both the watertight ground truth and the estimated meshes. We define a normal consistency score as the mean absolute dot product of the normals in one mesh and the

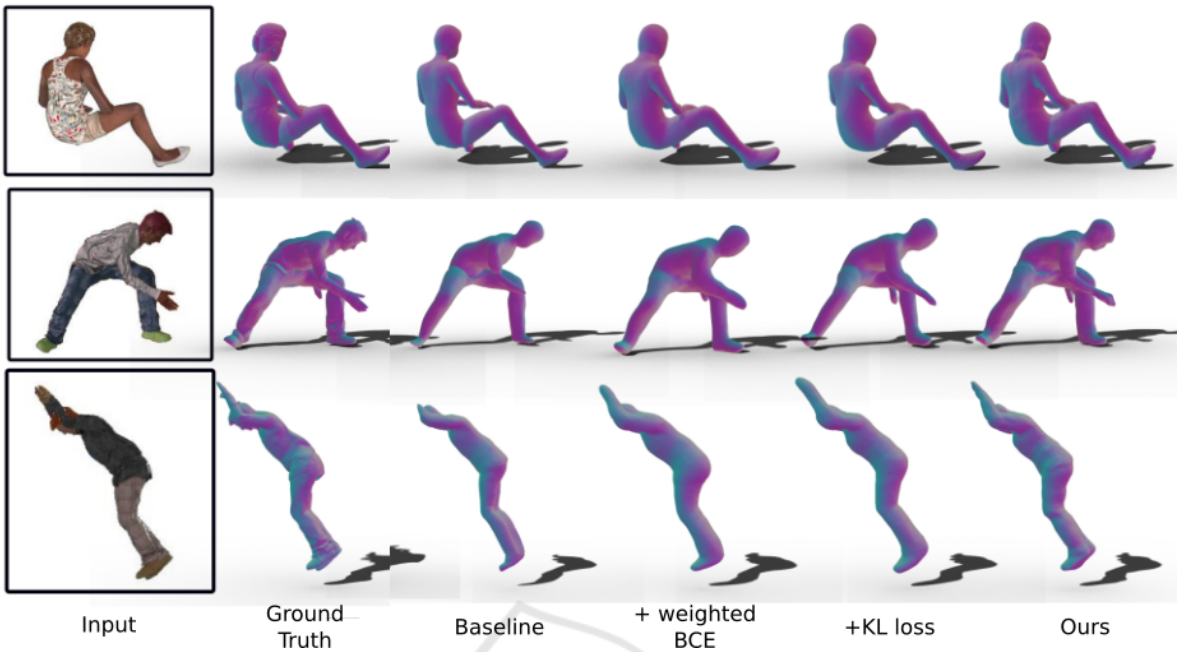

Figure 2: Comparison between the baselines used on 3DPeople. Baseline is (Mescheder et al., 2019) retrained on 3DPeople and subsequent columns are results of added components to that model validated in Table 2. The figure displays the reconstructed meshes from the camera viewpoint. The color of the meshes encodes the normal directions of the surface. Note how our approach captures the global consistency of the mesh, as the previous, and additionally presents certain clothing details.

Table 2: Quantitative ablation study on 3DPeople dataset. Note that dense sampling denotes sampling strongly in the surface of the mesh (meaning several more points than in uniform sampling).

Components							*Metrics*			
Occupancy	wBCE	KL	Joints	Uniform Samp.	Dense Samp.	Displacement	CD ↓	IoU↑	Normal Consistency↑	P2S↓
✓				✓			2.752	0.516	0.793	18.135
✓	✓			✓			1.689	0.576	0.808	18.698
✓	✓	✓		✓			1.496	0.579	0.811	18.353
✓	✓	✓		✓	✓		1.422	0.579	0.814	18.265
✓	✓	✓	✓	✓	✓		**1.051**	**0.612**	**0.829**	16.397
✓	✓	✓	✓	✓	✓	✓	1.082	0.606	0.821	**16.200**

normals at the corresponding nearest neighbors in the other mesh. As in (Saito et al., 2019), we measure the average point-to-surface Euclidean distance (P2S) in cm from the vertices on the reconstructed surface to the ground truth.

Fig. 2 shows three samples of the meshes reconstructed with each of the baselines and our final method. Regarding the three ONet baselines, note how the introduction of the losses tend to produce better reconstructions, although the sharper geometry details are more evident in our approach (Fig. 2(ours)), which includes all previous losses plus the refinement of the geometry estimated with the displacement network. Also the effect of other components of our model and the proposed training scheme if depicted qualitatively in Fig. 3.

Qualitative comparison on synthetic and real images is also presented. Fig. 4 presents sample synthetic images from our test set that none of the models have seen before. Here we present results of our method along with ONet and PIFu, note that all are re-trained with 3DPeople dataset. As shown in Fig. 4 one can note that PiFu if capable of reconstructing in a very acceptable manner all the front-view parts of the meshes, however, it fails to give a global consistency to the mesh. This can be seen in the columns depicting the side-view. We argue that this is due to PIFu's heavy reliance on local aligned features. Also we argue that PIFu is penalized by the relatively low-resolution of the input images, whereas our methods in not that sensitive to low-resolution failures. Additionally, since the 2D joints are not exploited by PiFu, the structure of the body it produces is not always consistent. Qualitative comparison with other SOTA methods on real images can be found in Fig. 5. Here we compare our method with (Saito et al., 2019; Saito

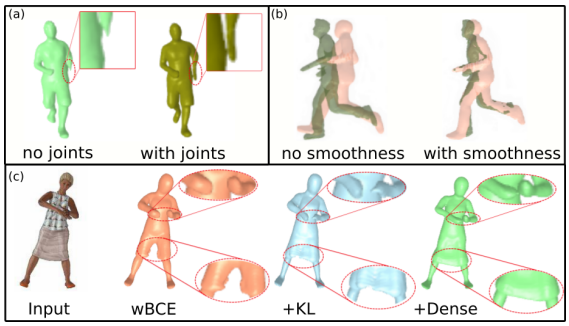

Figure 3: Visual ablation study. Note that these are outputs of the *coarse network* so they lack finer details. (a) Difference in reconstructions when using 2D joints as inputs vs. not using them. (b) Effect of enforcing smoothness. Green meshes are ground truth, pink ones are reconstructions. Here we present reconstruction when the network tries to learn the detailed ground truth mesh vs. reconstruction when forcing *coarse network* to learn a smooth version of the ground truth. (c) Impact of using a generative loss and "dense" sampling, meaning, sampling more heavily on the surface of the mesh. Adding generative loss helps to capture the richness of the 3D shape distribution. Here we can see that by adding KL loss and then sampling near the surface, especially around face and hands (dense), we obtain better results both in hands, face and skirts.(+KL=wBCE+KL, +Dense=wBCE+KL+Dense).

et al., 2020; He et al., 2020). Note that non of these methods nor ours have been train with real images and inference, in this case, is done with the trained weights provided by the authors of each method. All PIFu methods, except Geo-PIFu (to a lesser extent) show the same problem addressed before: shockingly good front views, however lacking global consistency and human body coherence. Geo-PIFu works better in these cases as this model specifically aims for global coherence just as our method does.

Impact of using 2D Joints. We found out that using 2D joints as additional input to our model improves the reconstruction quality. By adding joint information we prevent the network from generating incomplete human bodies, especially in cases where the image presents self-occlusions (see Fig. 3(a)).

Impact of Enforcing Smoothness. As stated before, we enforce the *coarse network* to learn a smooth version of the ground truth mesh. This reliefs the network from learning a more complex mapping to account for high-level details which has an impact on the correctness of the reconstructed human pose. As shown in Fig. 3(b), one can clearly see that when we do not enforce to learn a smooth version of the mesh, the pose deviates considerably from the ground truth.

Impact of using a Generative Model. The use of generative loss (equation 4) helps the model to better capture the richness and variability of the distribution of human clothing and body details such as hands and face. As it can be seen in Fig. 3(c), when adding the KL loss term to the model the skirt and hands are better reconstructed. Moreover, this is improved when combining this with the dense sampling strategy that was mentioned before.

Impact of Dense Sampling. When combined with the KL loss, the dense sampling strategy (near surface and around face and hands) helps the model to better capture the correct structure of clothing and human body. In the case of Fig. 3(c), we show how adding this sampling strategy results in better hands and skirts. Although not shown here, we also observed slight improvement in the face area.

Real Images. We finally show in Figures 4 and 5 the reconstructed shapes on synthetic and real images, respectively. Note, specially in the synthetic examples, how we are able to capture very complex body poses together with the details of the clothing (e.g. skirts). Also, note that for test and real images (given that we do not have ground truth for 2D joints) we use an off-the-shelf 2D pose detector such as (Cao et al., 2019). Another alternative is to use (Rong et al., 2021) and get all necessary joints by projecting them into the 2D space.

Although we get good results on real images, it can be perceived, in some cases, that the results are not as good as on synthetic ones. We hypothesize that this is due to a slight difference in appearance of real images in contrast to synthetic ones, especially due to lighting conditions, shadows and color. It is known that there is domain gap between real and synthetic images. We believe that by training with real images or paying more attention to the photo-realism of synthetic images we would get even better results. While we are able to capture skirts, where most of other methods fail, there is still room for improvement. However, we believe that combining global reasoning with a refinement step to add details is the right direction to obtain coherent human meshes in a wide range of poses with high enough detail.

6 CONCLUSIONS

In this paper we have made the following contributions to the problem of reconstructing the shape of dressed humans. As far as we can tell we are the first ones to do 3D reconstruction of clothed human body from single image in a wide range of poses including complex ones. In doing so, we do not require high resolution images. We demonstrate that different sampling schemes can improve the details with implicit function representation. Finally, we are able to

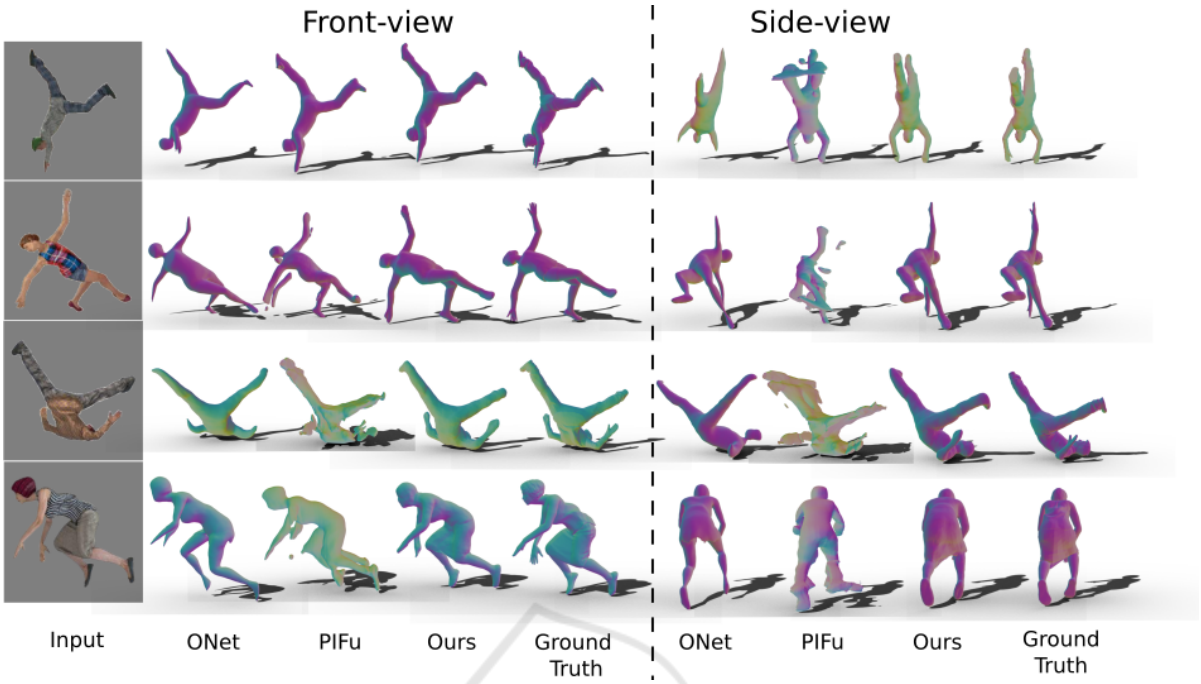

Figure 4: Results on synthetic images of the 3DPeople dataset. For every row we display the input RGB image and the mesh reconstructed using our approach and comparative approaches seen from two different viewpoints, Onets (Mescheder et al., 2019) and PIFu (Saito et al., 2019). The color of the meshes encodes the normal directions of the surface.

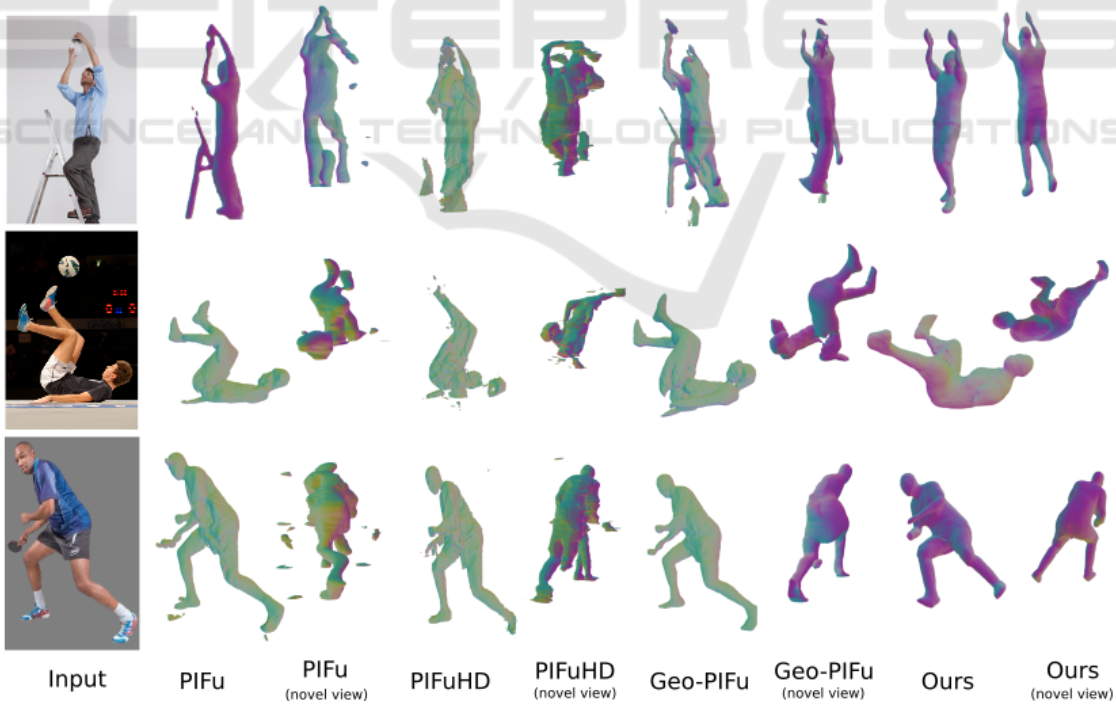

Figure 5: Qualitative results of our approach on real images. We compare with PIFu (Saito et al., 2019), PIFuHD (Saito et al., 2020) and Geo-PIFu (He et al., 2020).

capture details such as dresses and skirts while maintaining consistency of the body from all directions and not only the observed view.

ACKNOWLEDGMENTS

This work is supported by the Spanish government with the projects MoHuCo PID2020-120049RB-I00 and María de Maeztu Seal of Excellence MDM-2016-0656.

REFERENCES

Agudo, A., Moreno-Noguer, F., Calvo, B., and Montiel, J. M. (2016). Real-time 3d reconstruction of non-rigid shapes with a single moving camera. *Computer Vision and Image Understanding*, 153:37–54.

Alldieck, T., Magnor, M., Bhatnagar, B. L., Theobalt, C., and Pons-Moll, G. (2019a). Learning to reconstruct people in clothing from a single rgb camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alldieck, T., Magnor, M., Xu, W., Theobalt, C., and Pons-Moll, G. (2018). Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alldieck, T., Pons-Moll, G., Theobalt, C., and Magnor, M. (2019b). Tex2shape: Detailed full human body geometry from a single image. In *IEEE International Conference on Computer Vision (ICCV)*.

Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., and Davis, J. (2005). Scape: shape completion and animation of people. *Trans. Graph*.

Caliskan, A., Mustafa, A., Imre, E., and Hilton, A. (2020). Multi-view consistency loss for improved single-image 3d reconstruction of clothed people. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *T-PAMI*.

Chen, Z. and Zhang, H. (2019). Learning implicit fields for generative shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Choy, C. B., Xu, D., Gwak, J., Chen, K., and Savarese, S. (2016). 3d-R2n2: A Unified Approach for Single and Multi-view 3d Object Reconstruction. In *European Conference on Computer Vision*.

Dibra, E., Jain, H., Öztireli, C., Ziegler, R., and Gross, M. (2017). Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fan, H., Su, H., and Guibas, L. (2016). A Point Set Generation Network for 3d Object Reconstruction from a Single Image. *TOG*.

Genova, K., Cole, F., Sud, A., Sarna, A., and Funkhouser, T. (2020). Local deep implicit functions for 3d shape. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gkioxari, G., Malik, J., and Johnson, J. (2019). Mesh r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9785–9795.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, T., Collomosse, J., Jin, H., and Soatto, S. (2020). Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Conference on Neural Information Processing Systems (Conferenceon Neural Information Processing Systems (NeurIPS))*.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on International Conference on Machine Learning (ICML)*.

Jackson, A. S., Manafas, C., and Tzimiropoulos, G. (2018). 3d Human Body Reconstruction from a Single Image via Volumetric Regression. In *European Conference on Computer Vision*.

Kanazawa, A., Black, M. J., Jacobs, D. W., and Malik, J. (2017). End-to-end Recovery of Human Shape and Pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kinauer, S., Güler, R. A., Chandra, S., and Kokkinos, I. (2018). Structured output prediction and learning for deep monocular 3d human pose estimation. In Pelillo, M. and Hancock, E., editors, *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 34–48, Cham. Springer International Publishing.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*.

Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M. J., and Gehler, P. V. (2017). Unite the People: Closing the Loop Between 3d and 2d Human Representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Lorensen, W. E. and Cline, H. E. (1987). Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH*.

Martinez, J., Hossain, R., Romero, J., and Little, J. J. (2017). A simple yet effective baseline for 3d human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*.

Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., and Theobalt, C. (2018). Single-shot multi-person 3D pose estimation from monocular RGB. In *3DV 2018 , International Conference on 3D Vision*, pages 120–130, Verona, Italy. IEEE.

Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., and Geiger, A. (2019). Occupancy Networks: Learn-

ing 3d Reconstruction in Function Space. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Moon, G., Chang, J., and Lee, K. M. (2019). Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *The IEEE Conference on International Conference on Computer Vision (IEEE International Conference on Computer Vision (ICCV))*.

Moreno-Noguer, F. (2017). 3d human pose estimation from a single image via distance matrix regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Moreno-Noguer, F. and Fua, P. (2013). Stochastic exploration of ambiguities for nonrigid shape recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(2):463–475.

Moreno-Noguer, F. and Porta, J. M. (2011). Probabilistic simultaneous pose and non-rigid shape. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (IEEE Conference on Computer Vision and Pattern Recognition (CVPR))*, pages 1289–1296.

Natsume, R., Saito, S., Huang, Z., Chen, W., Ma, C., Li, H., and Morishima, S. (2019). Siclope: Silhouette-based clothed people. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., and Schiele, B. (2018). Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE.

Onizuka, H., Hayirci, Z., Thomas, D., Sugimoto, A., Uchiyama, H., and Taniguchi, R.-i. (2020). Tetratsdf: 3d human reconstruction from a single image with a tetrahedral outer shell. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6011–6020.

Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S. (2019). DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pavlakos, G., Zhou, X., Derpanis, K. G., and Daniilidis, K. (2017). Coarse-to-fine volumetric prediction for single-image 3d human pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pavlakos, G., Zhu, L., Zhou, X., and Daniilidis, K. (2018). Learning to Estimate 3d Human Pose and Shape from a Single Color Image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pumarola, A., Agudo, A., Porzi, L., Sanfeliu, A., Lepetit, V., and Moreno-Noguer, F. (2018). Geometry-aware network for non-rigid shape prediction from a single view. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (IEEE Conference on Computer Vision and Pattern Recognition (CVPR))*.

Pumarola, A., Popov, S., Moreno-Noguer, F., and Ferrari, V. (2020). C-flow: Conditional generative flow models for images and 3d point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pumarola, A., Sanchez-Riera, J., Choi, G. P. T., Sanfeliu, A., and Moreno-Noguer, F. (2019). 3dpeople: Modeling the geometry of dressed humans. In *IEEE International Conference on Computer Vision (ICCV)*.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2016). Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*.

Rogez, G., Weinzaepfel, P., and Schmid, C. (2019). LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Rong, Y., Shiratori, T., and Joo, H. (2021). Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *IEEE International Conference on Computer Vision Workshops*.

Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., and Li, H. (2019). Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Saito, S., Simon, T., Saragih, J., and Joo, H. (2020). Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sanchez, J., Östlund, J., Fua, P., and Moreno-Noguer, F. (2010). Simultaneous pose, correspondence and non-rigid shape. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (IEEE Conference on Computer Vision and Pattern Recognition (CVPR))*, pages 1189–1196.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

Sohn, K., Yan, X., and Lee, H. (2015). Learning structured output representation using deep conditional generative models. In *Conferenceon Neural Information Processing Systems (NeurIPS)*.

Sorkine, O., Cohen-Or, D., Lipman, Y., Alexa, M., Rössl, C., and Seidel, H.-P. (2004). Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, SGP '04, page 175–184, New York, NY, USA. Association for Computing Machinery.

Tulsiani, S., Zhou, T., Efros, A. A., and Malik, J. (2017). Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., and Schmid, C. (2018). Bodynet: Volumetric inference of 3d human body shapes. In *European Conference on Computer Vision*.

Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M. J., Laptev, I., and Schmid, C. (2017). Learning from synthetic humans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vince Tan, I. B. and Cipolla, R. (2017). Indirect deep structured learning for 3d human body shape and pose prediction. In *British Machine Vision Conference (BMVC)*.

Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., and Jiang, Y.-G. (2018). Pixel2mesh: Generating 3d Mesh Models from Single RGB Images. In *European Conference on Computer Vision*.

Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, B., and Tenenbaum, J. (2017). Marrnet: 3d shape reconstruction via 2.5 d sketches. In *Advances in neural information processing systems*, pages 540–550.

Xu, Q., Wang, W., Ceylan, D., Mech, R., and Neumann, U. (2019). DISN: Deep Implicit Surface Network for High-quality Single-view 3d Reconstruction. In *Conferenceon Neural Information Processing Systems (NeurIPS)*.

Zhang, C., Pujades, S., Black, M., and Pons-Moll, G. (2017). Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zheng, Z., Yu, T., Wei, Y., Dai, Q., and Liu, Y. (2019). Deephuman: 3d human reconstruction from a single image. In *The IEEE International Conference on Computer Vision (IEEE International Conference on Computer Vision (ICCV))*.

Zhu, H., Zuo, X., Wang, S., Cao, X., and Yang, R. (2019). Detailed human shape estimation from a single image by hierarchical mesh deformation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.