TetraTSDF: 3D human reconstruction from a single image with a tetrahedral outer shell

Hayato Onizuka
Kyushu University
mound028810@gmail.com

Zehra Hayırcı
Technical University of Munich
zehrahayirci@gmail.com

Diego Thomas
Kyushu University
thomas@ait.kyushu-u.ac.jp

Akihiro Sugimoto
National Institute of Informatics
sugimoto@nii.ac.jp

Hideaki Uchiyama
Kyushu University
uchiyama@limu.ait.kyushu-u.ac.jp

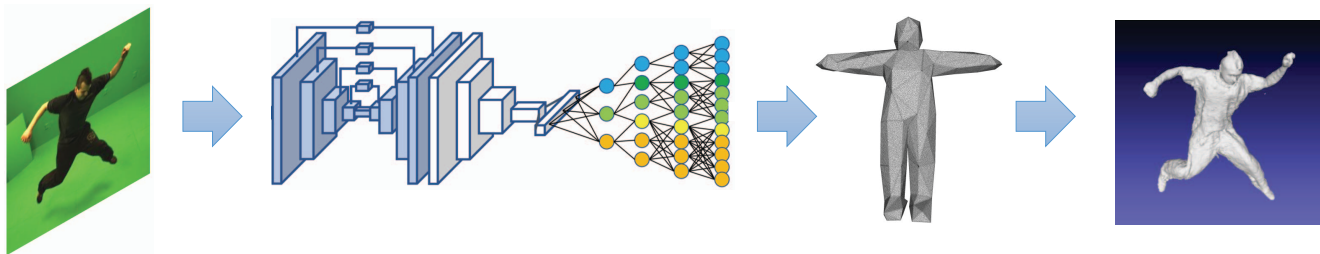Rin-ichiro Taniguchi
Kyushu University
rin@kyudai.jp

Figure 1: We present a tetrahedral volumetric representation of the human body and a method called TetraTSDF that is able to retrieve the detailed 3D shape of a person wearing loose clothes from a single 2D image.

Abstract

Recovering the 3D shape of a person from its 2D appearance is ill-posed due to ambiguities. Nevertheless, with the help of convolutional neural networks (CNN) and prior knowledge on the 3D human body, it is possible to overcome such ambiguities to recover detailed 3D shapes of human bodies from single images. Current solutions, however, fail to reconstruct all the details of a person wearing loose clothes. This is because of either (a) huge memory requirement that cannot be maintained even on modern GPUs or (b) the compact 3D representation that cannot encode all the details. In this paper, we propose the tetrahedral outer shell volumetric truncated signed distance function (TetraTSDF) model for the human body, and its corresponding part connection network (PCN) for 3D human body shape regression. Our proposed model is compact, dense, accurate, and yet well suited for CNN-based regression task. Our proposed PCN allows us to learn the distribution of the TSDF in the tetrahedral volume from a single image in an end-to-end manner. Results show that our proposed method allows to reconstruct detailed shapes of humans wearing loose clothes from single RGB images.

1. Introduction

Detailed 3D shapes of the human body reveal personal characteristics that cannot be captured with standard 2D pictures. Such information is crucial for many applications in the entertainment industry (3D video), business (virtual try-on) or medical use (self-awareness or rehabilitation). The first systems that built 3D models of the human body were designed to work in controlled settings using laser scanners, multi-view calibrated camera arrays or markers. These systems were hard to set up, not affordable, and offering limited application areas.

In the last decade, consumer-grade depth cameras have been successfully used to build 3D models of the human body [45]. However, high quality depth cameras are still not available to a large part of the consumers (most of smartphones are still not equipped with depth cameras). Moreover, consumer-grade depth cameras do not work well in outdoor environments. As a consequence, methods that can efficiently reconstruct detailed 3D human shapes in unconstrained environments are needed. In this work, we focus on the task of detailed 3D human body reconstruction in a single shot with a standard RGB camera.

In the literature, there are two strategies to generate a 3D

model from a single color image: (1) parametric model fitting and (2) 3D model regression. Methods that fall into the first strategy fit a parametric human template 3D model (such as the SMPL model [19]) to the input color image. To fit the template model, various cost functions have been proposed that consider silhouette, skeleton and feature points ([4]). The latter strategy takes advantage of the recent advances in convolutional neural networks (CNN). Thereby, depth image regression is followed by volumetric fusion [33], or end-to-end RGB to 3D model techniques [1, 30, 36] have been proposed.

CNN-based methods are promising for reconstructing 3D human bodies from a single color image because they have the potential to capture detailed and complex features (such as clothes wrinkles). However, several limitations exist. The main problem stems from the volumetric TSDF representation which is used to regress the 3D shape. The resolution of the volumetric representation has to meet the fine details of the human body, meaning that a considerable amount of memory is required. However, such memory constraint is hard to be maintained even on modern GPUs.

The key challenge for achieving higher accuracy in 3D human body shape reconstruction is to define a more compact 3D human body representation in memory that still allows casting the problem as a well-adapted regression task. One solution may be displacement mapping [1]. However, this compact representation inevitably loses some details in shape because its dependency to the SMPL mesh. For instance, occluded parts in non-convex areas or garments like shoes or gloves cannot be reconstructed by using displacement mapping.

In this paper, we propose a new volumetric 3D body representation for end-to-end 3D body shape regression from a single color image. Our proposed 3D body representation is based on a tetrahedral TSDF field embedded into a human-specific outer shell. The outer-shell is built from a coarse version of the SMPL model [19] and can be fitted to a human body using the SMPL pose and shape parameters. The tetrahedral TSDF field is built at the summits of a tetrahedral volumetric grid defined by the outer shell. We also propose a new network to estimate the tetrahedral TSDF field from a single color image that combines CNN and our proposed Part Connection Network (PCN).

Our contributions are three fold: (1) a new 3D body volumetric representation that is compact, dense, accurate and yet well suited for CNN-based regression tasks; (2) a method to generate high quality TSDF fields from ground truth (GT) 3D human body scans; and (3) a new CNN-PCN based hourglass network for end-to-end regression of 3D human body shape from a single color image.

2. Related work

Fitting a parametric 3D model to the input 2D color image has been the standard way to reconstruct a 3D shape from a single 2D image for a long time. Recently, using CNNs has proven to be a powerful alternative. Here, we review related works that use both of these strategies, with a particular focus on the human body shape reconstruction problem.

2.1. Template model fitting

The classic approach to estimate a 3D shape of an object from a single color image is to fit a template 3D model so that it matches its 2D projection while satisfying some constraints (*e.g.*, [4, 5, 19]). Landmark-guided non-rigid registration of 3D templates to 2D or 2.5D inputs have been widely studied. Lu *et.al.* [21], for example proposed to use facial landmarks to fit a deformable face model to 2.5D data. Cashman *et.al.* [5], proposed to represent a 3D deformable model with a linear combination of subdivision surfaces. Such model can be fitted to a collection of 2D images by manually providing some key-points and the silhouette of the object to be reconstructed.

Recovering the 3D human shape from a single RGB image is an open-problem in the field. There are only few proposed techniques that deal with complex poses and deformations. Most of the techniques rely either on a pre-scanned model of the subject [8, 13, 18, 32, 38] or a template model [2, 3]. In [13], the authors propose to first scan the 3D model of a person using a multi-view reconstruction system. Then, the reconstructed 3D model is non-rigidly aligned to the RGB video in real-time. Similarly, in [12], Bogo *et.al.* [4], propose to optimize the parameters of a parametric template model given input images and poses by using many cues like silhouette overlap, height constraint and smooth shading. The authors use the SMPL model [19] to recover various parameters such as pose and shape from a single RGB image. Recently, Kolotouros *et al.* [17] proposed a method to estimate SMPL parameters by developing a self-improving loop combining CNN and optimization method. However, template-based methods fail to capture loose clothing and thus only reconstruct bare human body.

2.2. Convolutional Neural Network regression

Recently, CNNs have brought new possibilities to many domains in computer vision. 3D shape reconstruction from a single image is one of those areas that strongly developed with the availability of new CNN tools.

Inspired by the extraordinary performance of CNNs for segmentation tasks, several methods have been proposed that represent the 3D shapes as binary occupancy maps [6, 11, 35, 40, 41]. If the task of estimating the 3D surface is expressed as a segmentation problem, then CNNs

can predict outside and inside voxels. For example, Wu et al. [42] proposed an extension of 2D CNNs for the case of volumetric outputs. Further optimization improvements were proposed in [44] and [46]. In the case of 3D face regression, Jackson *et.al.*, [15] proposed a method for direct regression of a volumetric representation of the face using CNN. By using probabilistic assignment, smooth surfaces could be obtained. Varol *et.al.* [36], extended this method to full body shape regression. All these methods share the common limitation that the memory consumption scales cubically with the shape resolution. Even with modern GPUs, volumetric regression networks only work with low resolution grid. Then, only coarse 3D models can be generated.

Riegler *et al.* [28], proposed to use octrees to reduce the memory usage and adapt the CNNs to predict high-resolution occupancy maps providing the tree structure is known in advance. However, the method can not be applied to reconstruct 3D human bodies with different poses because the tree structure changes with every new input. In [34], Tararchenko *et.al.* proposed a technique to overcome this problem by also predicting the tree structure. However, training the network to learn the sparse structure of octrees is effortful. Recently, Saito *et al.* [30] proposed a memory-efficient method by handling each 3D point individually. They reported high accuracy 3D reconstruction results on a private dataset while using reasonable amount of memory. However, this method jointly estimates the 3D shape and body pose while 3D body pose estimators have made significant progress recently and achieved high accuracy results (*e.g.*, [14] reported an average error of less than 2 cm for the 3D joint position). We reason that the tasks should be separated: one task for body pose estimation and one task dedicated to the 3D shape estimation.

Recently, Alldieck *et al.* [1] have proposed to use displacement mapping on top of a template human model to represent the 3D human body with loose clothes. This not only fits well to the CNN formulation but it also requires low memory. However, it has a severe limitation that it can not reconstruct convex parts like inside cloth wrinkles. In addition, because the template human model is naked with fingers in hands and feet and deviation mapping can only encode displacement in the normal vector direction, reconstructing shoes or gloves for example is not possible. In the meantime Gabeur *et al.* [10] proposed a method that predicts depth maps from visible and hidden side by using GAN to "mould" the 3D human body.

We reason that when reconstructing the human body with loose clothes, the implicit volumetric TSDF is the best representation to handle various shapes. We observe that in the regular grid, many of the voxels in the 3D bounding box around the person are actually unnecessary. Therefore, we propose a new tetrahedral 3D shape representation that is able to reduce the memory consumption drastically while
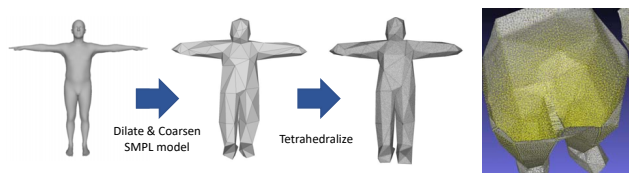

Figure 2: Our tetrahedral outer shell for the human body is built from a coarsen version of SMPL template model.
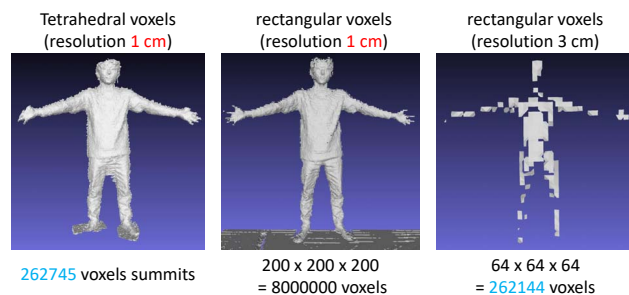

Figure 3: Our representation allows us to reconstruct detailed 3D shapes while using a small amount of voxels.

allowing to reconstruct high resolution 3D shapes.

3. Proposed tetrahedral representation

The Truncated Signed Distance Function was first introduced by Curless and Levoy in [7] to represent 3D surfaces and has been extensively used in modern RGB-D simultaneous localisation and mapping (SLAM) systems [23, 24, 27, 29]. At any point in the 3D space, the TSDF function takes the signed distance to the 3D surface as value. These values are truncated between -1 and 1 for practical implementation reasons. In general the TSDF is sampled in a regular grid of (rectangular) voxels and the 3D mesh of the surface can be extracted using well established algorithms such as the Marching Cubes [20] or ray tracing.

Our objective here is to reduce the irrelevant field around the human body while covering the meaningful space, which we call the outer shell. We propose to modify the well known SMPL model [19] to create this outer shell. The SMPL model has well defined pose and shape parameters that can be fitted to any 3D human dataset, skeleton or even RGB image (using CNNs for example).

3.1. Coarse human outer shell

We propose to inflate the SMPL neutral body model so that once fitted to the input it covers the entire body as well as the loose clothes. We also propose to remove shape details of the SMPL model (such as nose, mouth etc ...). Our pipeline is illustrated in Figure. 2. We reason that we do not need details at the surface of the outer-shell because the details will be encoded into the TSDF field. Therefore we
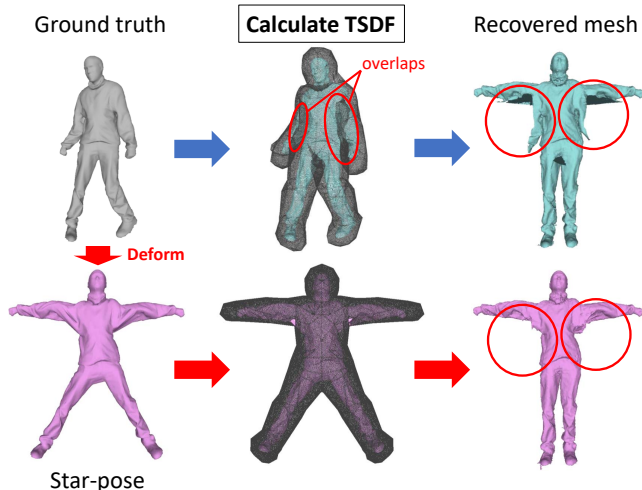
6012

Figure 4: After fitting the coarse SMPL model to the 3D scan we calculate the TSDF values from the star-pose warped ground truth 3D mesh.

down-sample the SMPL model, and then up-sample to generate a coarse outer shell with high density vertices[1].

Our proposed outer shell allows us to define a compact 3D space around the person. However, the shape of the outer shell is highly irregular and the standard uniform grid discretization of the 3D space in rectangular voxels is not possible anymore. Thus we propose using tetrahedral discretization of the 3D space instead of the standard rectangular voxels. Figure 2 illustrates the tetrahedral grid created from the outer shell. Note that the outer shell is built only once and then fitted to any person. As a consequence, the same tetrahedral grid can be used for any person.

Figure 3 shows the advantage of our proposed tetrahedral representation over the standard uniform grid for human body 3D representation. Our proposed representation requires significantly less amount of voxels to represent a 3D surface while keeping the same density of points in the extracted surface. Note that the amount of voxels is directly related to the number of parameters in the CNN network. In practice, there are two main differences with the standard rectangular TSDF representation: (1) the TSDF value in the standard volumetric representation is stored at the center of each voxel, whereas, in out case, the value is stored at the summits of each tetrahedra. (2) Naturally, we prefer Marching Tetrahedra algorithm [31] instead of Marching Cubes to extract the 3D surface.

3.2. TSDF computation inside tetrahedra

In order to train a CNN to regress the TSDF values inside our proposed outer shell from a given 2D image, we must

---

[1]We used the Blender software (https://www.blender.org) to do the dilation, down-sampling and up-sampling processes

first create a ground truth dataset to supervise the learning process. In our case, the training dataset consists of a set of pairs of one 2D image and one corresponding dense tetrahedral TSDF field (*i.e.*, the set of TSDF values for each voxel summit). To build a large amount of dense TSDF fields from a set of publicly available ground truth human 3D scans we need an efficient algorithm. In this section we detail our proposed algorithm to generate such ground truth training dataset.

Firstly, we fit the coarse model to the GT 3D scan by optimizing the SMPL pose and shape parameters given the GT 3D skeleton. Then we compute the TSDF value for each voxel of the outer shell. Here arise two problems: (1) the GT 3D mesh is sparse, so the standard signed distance between a voxel summit and its closest point is significantly different than the signed distance to the surface. (2) In some poses parts of the coarse model may overlap, which may result in ghost effects (for example, a part of the torso may be encoded into the voxels that correspond to the arm).

To solve the first problem we compute the TSDF values as the point-to-plane distance between the voxel summit and the tangent plane of the closest point. This is a reasonable approximation of the point-to-surface signed distance (and it can be computed quickly) when the voxel is close to the surface. However, the approximation may become completely wrong when the voxel is far from the surface. To overcome this, we truncate the TSDF values based on a threshold on the euclidean point-to-point distance between the voxel summit and its closest point.

$$TSDF(v) = \begin{cases} \frac{\hat{n} \cdot (v - \hat{v})}{\tau} & \text{if } \|v - \hat{v}\|_2 \leq \tau \\ \sigma(\hat{n} \cdot (v - \hat{v})) & \text{if } \|v - \hat{v}\|_2 > \tau \end{cases}, \quad (1)$$

where v is a voxel summit, $\hat{v}$ is the closest point of v in the GT 3D mesh, and $\hat{n}$ the normal vector at $\hat{v}$. The threshold τ is set to 3 cm in the experiments and the function $\sigma(\cdot)$ returns the sign of its argument.

In order to solve the second problem we compute the TSDF values from the 3D scan warped into a star pose, where the body parts of the outer shell do not penetrate each other (see fig.4). To this end we first connect each vertex of the GT 3D mesh to its corresponding skeletal nodes with appropriate weight. Then we warp the 3D vertex coordinates to the T-shape by blending the 3D transformation of the attached skeletal joints. Once all vertices have been warped, we compute the dense TSDF field of the outer-shell as explained above.

4. Detailed 3D shape regression

We propose an end-to-end regression network to estimate the TSDF values in the tetrahedral volume from a single image. Our proposed network takes as input a single 2D
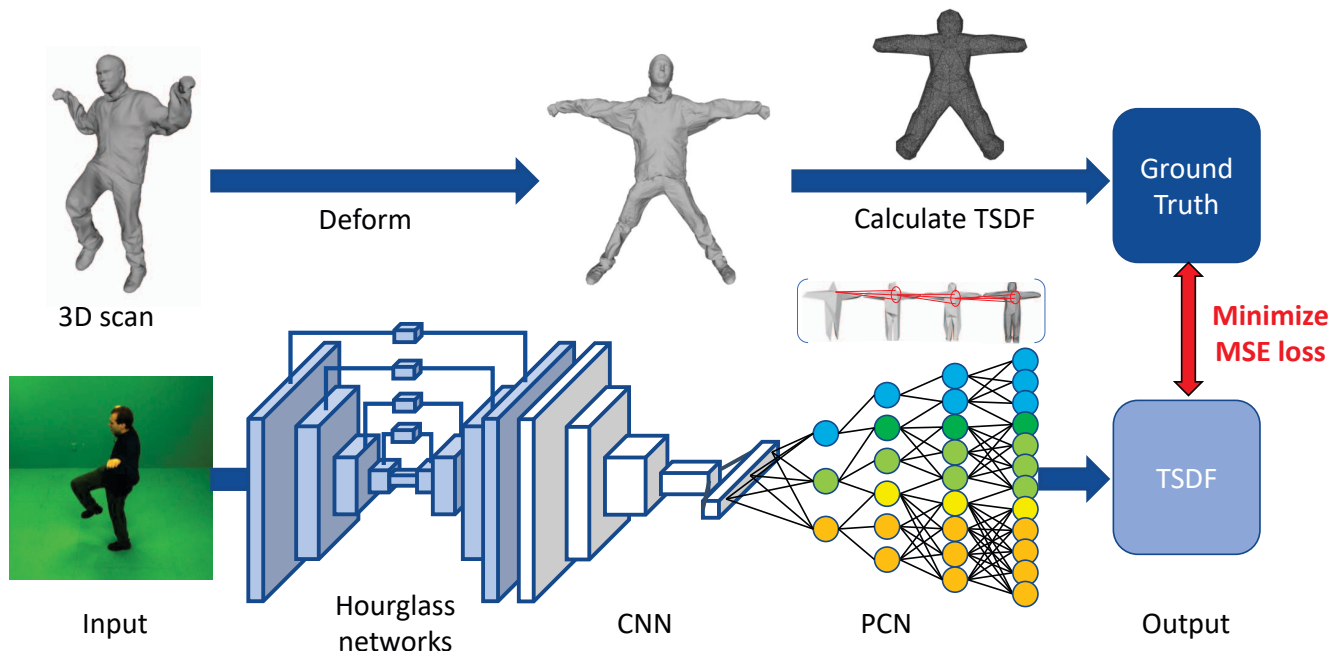
Figure 5: Our proposed network is a combination of CNN and PCN. Given a 3D pose data, it allows us to regress the 3D shape of a person wearing clothes in a end-to-end fashion.

image and outputs the dense TSDF field. From this TSDF field a detailed 3D surface of the human can be extracted by using the Marching Tetrahedra algorithm. The estimated 3D mesh can be re-posed using SMPL pose parameters obtained by any state-of-the-art 3D pose estimation method.

The standard approach to regress TSDF in volumetric data is to use stacked hourglass networks [25]. In this way, the input image is encoded into a feature vector, which is then decoded into the volumetric grid. To this end, CNNs are used to build the network with the help of the well organized data into uniform grids of pixels and voxels. In our case the volumetric data is embedded into the tetrahedral mesh, which does not have the uniform grid organization. As a consequence, state-of-the-art CNN hourglass architecture cannot be used directly.

We propose a new hourglass structure to regress the TSDF value of each vertex in the tetrahedral voxel model. Our proposed network combines CNN to encode the input image and a new Part Connection Network (PCN) to decode the feature vector (Figure 5 illustrates our proposed network). The originality of our network resides in the later part. We propose to build several tetrahedral layers by down-sampling the full resolution outer shell volumetric model (see fig. 5), and then we propagate the features in the upper layer to output the TSDF field. Note that the number of voxels in the tetrahedral model is too large to directly use a fully connected network (the number of parameters near the final layer consumes about 90GB of memory in the case

of a fully connected network). Instead we propose a Part Connected Network with partially connected layers, where connections between successive layers are done only in-between the same body parts (which consumes only 0.025 GB of memory).

For each layer l we define the partial connections using the following sparse adjacency matrices.

$$A_1^l = \begin{bmatrix} A(1)^l \\ A(2)^l \\ \vdots \\ A(n_{\text{out}})^l \end{bmatrix}. \qquad (2)$$

$$A(n)^l = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n_{\text{in}}} \\ a_{21} & a_{22} & \cdots & a_{2n_{\text{in}}} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn_{\text{in}}} \end{bmatrix}, \qquad (3)$$

$$\text{where} \quad a_{ij} = \begin{cases} 1 \text{ if } j = adj(n)[i] \\ 0 \text{ otherwise.} \end{cases}$$

$$A_2^l = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & & & & \ddots & & & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix}, \qquad (4)$$

where $n_{\mathrm{in}}, n_{\mathrm{out}}$ is the number of input and output nodes, m is the number of adjacent nodes for the n-th output node, and $adj(n)$ is the list of indices of input nodes that are connected to the n-th output node (and $adj(n)[i]$ is the ith element in the list). There is no guarantee that all of the output nodes have the same number of adjacent nodes, so we flatten the input features by using the matrix A_1^l, and reshape it to the shape of the output features by using A_2^l.

With these matrices, the features from two successive layers $(l, l+1)$ are transmitted as follows.

$$f_{l+1} = \sigma(A_2^l(A_1^l f_l \circ W_l)), \tag{5}$$

where f_l, f_{l+1} are the input and output features of the layer and W_l is the variable weight matrix for all edges. Also, $\circ$ denotes the element-wise product of two matrices and σ denotes the activation function.

To define the adjacency matrices A_1^l and A_2^l we need to identify the adjacency lists L that connect the input nodes and the output nodes of all layers. To create these adjacency lists, we focus on the locality of the nodes in the tetrahedral model. Concretely, each node in the lth layer is connected to the k-nearest neighbors in the $l+1^{th}$ layer (we used $k = 5$ in our experiments). We reason that the tetrahedral voxel model has a human body shaped graph structure and that we can consider that distant nodes have a weak relationship with each other (like between toes and fingertips). As a consequence, by connecting only the near nodes to the next layer and not connecting the distant nodes, the number of learning parameters is drastically reduced without loosing much information from the features from the previous layer.

We design the network structure so as not to lose features while reducing the number of variable parameters by connecting only the adjacent nodes. However, when connecting the last CNN layer of the first half of the network to the first PCN layer of the latter half, there is no notion of adjacency because nodes from the CNN layer do not have the shape of the human body. Therefore, we use a fully connected layer between the CNN and the PCN networks.

5. Experiments

We evaluate qualitatively and quantitatively the ability of our proposed method to reconstruct 3D shapes of the human body from a single image by using publicly available datasets.

In all our experiments, we used a volumetric resolution (*i.e.*, average distance between adjacent summits in the tetrahedral volume) of about 1 cm (which corresponds to about 2.6×10^5 voxel summits). We evaluated our network on SURREAL [37] and Articulated [38] datasets. In the evaluation using SURREAL we strictly followed the protocol as explained by the authors in [37]. Our network was tested on subjects not appearing in the training dataset.

Table 1: Quantitative comparison of results obtained with our method, BodyNet and Tex2Shape on the SURREAL [37] (naked) and Articulated [38] (clothed) datasets.

Chamfer (cm)	SURREAL	Articulated
SMPLify-x	n.a.	9.61
BodyNet	6.38	7.22
Tex2Shape	n.a.	0.72
Ours	**5.14**	**0.43**

This allows us to test the generalizability of our proposed method. In Articulated, we trained our network on 80% of the data and tested on the remaining 20%. Our method was tested on poses (and thus deformations) not appearing in the training dataset. This allows us to test the ability of our proposed methods to reconstruct detailed 3D shapes such as clothes wrinkles. In the training, we employed the mean square error loss function in the last layer and reLU as an activation function in the hidden layers. It took about 3-4 hours to train the network for Articulated dataset, with using a batch size of 5 and a single GTX 1080 GPU.

We compared our proposed method with other recent works that have made their code publicly available ([1, 26, 36]). Note that for Tex2Shape [1] only the code for testing is available and so we built a discriminator network for the training of Tex2Shape referring to their paper.

5.1. Comparative evaluation on SURREAL dataset

We compare our proposed method with BodyNet [36] to confirm the advantage of using our proposed tetrahedral volumetric representation over the classic uniform rectan-
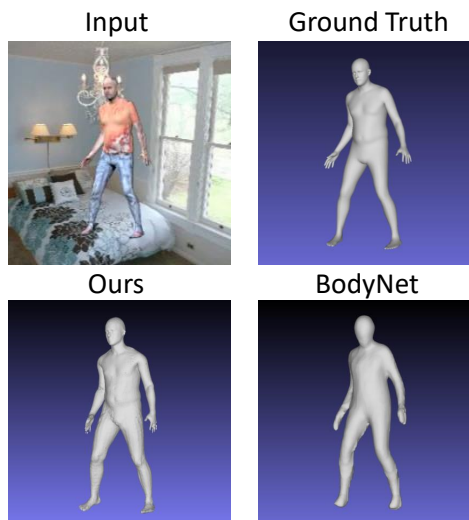

Figure 6: An example of a 3D mesh obtained with our method and with BodyNet on the SURREAL dataset.
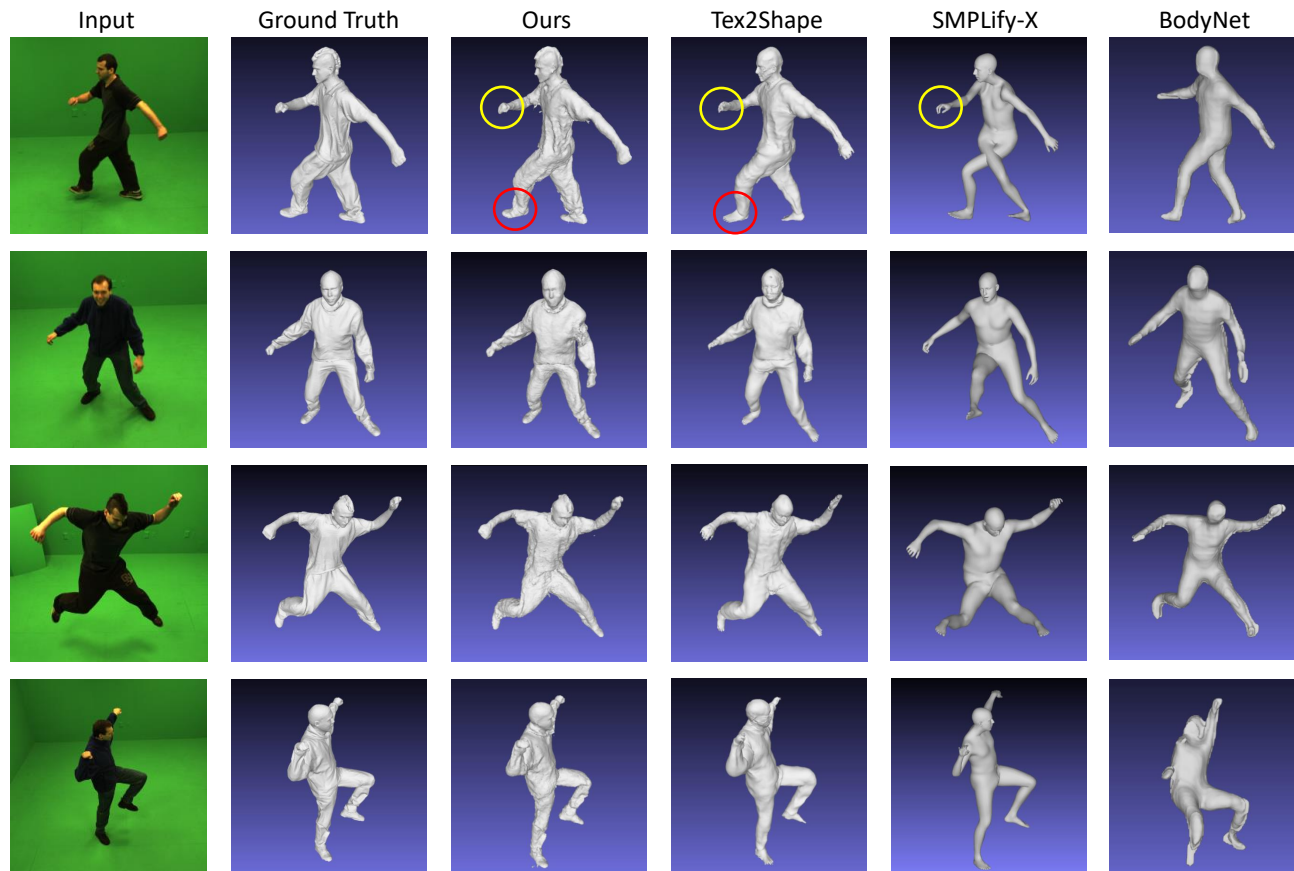
6015

Figure 7: Comparative reconstruction results obtained with our proposed method and other related works. From left to right: input image, ground truth, our method, Tex2Shape [1], SMPLiFy-X [26], BodyNet [36]

gular grid. We trained and tested our network on the SUR-REAL dataset [37], which was used in [36]. Figure 6 shows the qualitative comparative results obtained with our proposed method and BodyNet. As we can see our proposed CNN-PCN network was able to successfully reconstruct the detailed 3D shapes of the human body from only one single image. Note that BodyNet estimates both pose and shape while our method only estimates the 3D shape. We used HMR[16] pose estimation results as pose parameters to show our results in the same pose as the input image.

Table 1 shows the quantitative comparison. For the metric we used the Chamfer distance between the ground truth 3D scans and the reconstructed 3D meshes. As we can see from these results our proposed method was able to reconstruct accurate dense 3D shapes of the human body from only one single image. We obtained better results that BodyNet because our proposed tetrahedral representation allows for reconstruction at higher resolution.

5.2. Comparative evaluation on Articulated dataset

To confirm the advantages of our proposed method for reconstructing detailed 3D shapes of humans with loose clothes from a single image we compared our method with the most recent state-of-the-art method Tex2Shape [1]. For qualitative and quantitative evaluation, we used the publicly available dataset called Articulated [38] that contains sequences of human wearing loose clothes in many poses with ground truth 3D scans.

Unfortunately, the dataset used in [1] is not publicly available so we could not directly compare our proposed method with [1] on their own dataset. We built a discriminator network for the training of Tex2Shape referring to their paper [1] and trained their network on ARTICULATED dataset using the exact same train/test split as used with our proposed method. The comparative results shown in Figure 7 and Table 1 shows advantages of our proposed method compared with Tex2Shape. Note that as in [1] we show the results re-posed with the ground truth 3D pose. However, any third party 3D pose estimation could also be used to estimate the 3D pose from the input image ([22, 39]). Figure
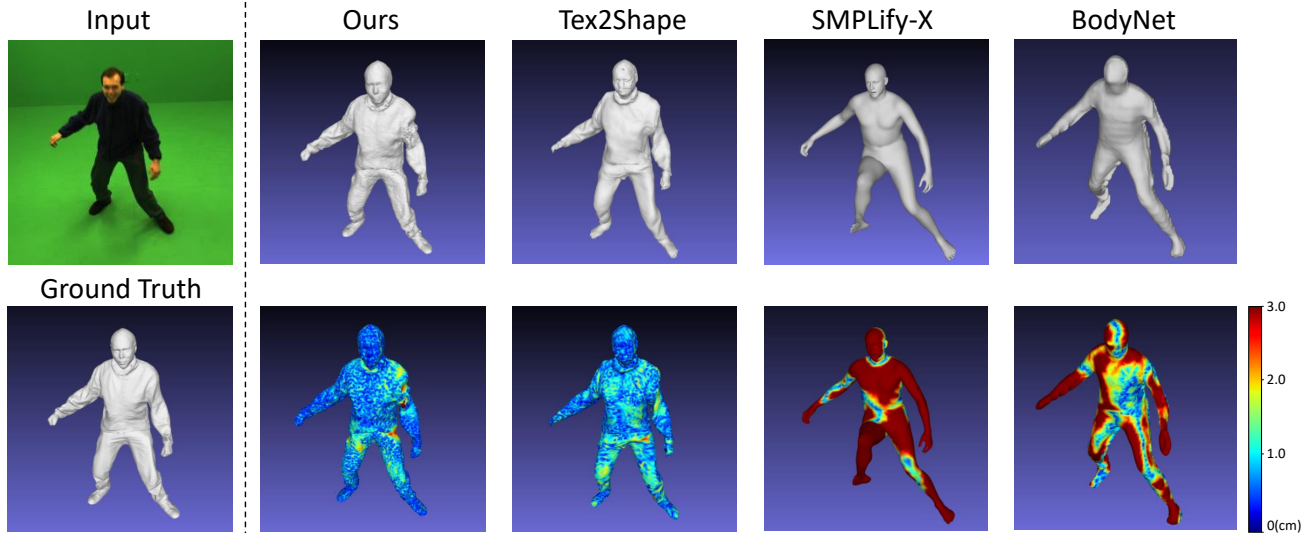
Figure 8: Visualization of the Euclidean distances between the vertices of reconstructed 3D mesh and the closest points in the GT scan. These errors are represented with heatmaps and mapped over the reconstructed 3D mesh for each method.

7 shows some representative results from the test dataset.

We also compared the results obtained with our method with those obtained with other closely related works. As seen in figures 7 and 8, our proposed method outperforms all other previous works. Figure 8 shows the heat maps of errors of the 3D reconstructed models obtained by our method and other related works. As we can see our proposed network could successfully recover the details of the loose clothes, even in occluded areas. As we can see in Table 1 we observed an average error around 0.5 cm in the 3D meshes reconstructed with our proposed method.

As we can see in the circled areas of figure 7, the strongest and clearer advantage of our proposed method over the related works is that our proposed method can retrieve both hand details such as fingers in the correct position (no details can be obtained with BodyNet) or in the inverse reconstruct the shoes around the feet (fingers still remain with Tex2Shape, giving an unpleasant effect for people wearing shoes). Moreover, with our proposed method we do not need to estimate finger poses. As we can see in Figure 7 although the hand pose is unknown (this is why the hand is always open in the results obtained with Tex2Shape) our proposed method could successfully reconstruct the detailed hand in the correct pose.

5.3. Network analysis

We analysed the performance of our network by changing the parameters. Figure 9 (a) shows the convergence speed of our network during training. Figure 9 (b) shows the performance of our network when changing the number of connections in the PCN. As expected, there is a trade-off between the number of connections (*i.e.*, memory consumption) and accuracy.
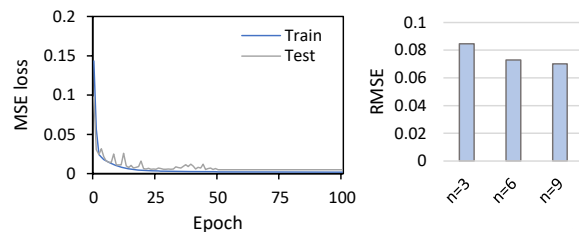

Figure 9: Left: learning curve of our network on ARTIC-ULATED dataset. Right: we reduced the connection of the nodes from original (n=9) to n=3, n=6.

6. Conclusion

We proposed a method for reconstructing fine detailed 3D shapes of human body wearing loose clothes from a single image. We introduced a novel 3D human body representation based on a tetrahedral TSDF field embedded into a coarse human outer shell. We also designed CNN-PCN network to regress the tetrahedral TSDF field in an end-to-end fashion. Our qualitative and quantitative comparative experiments using public datasets confirmed the ability of our proposed method to reconstruct detailed shapes with loose clothes. The results demonstrated out-performance of our method against the current state-of-the-art on these datasets. Several possible improvements are left for future work, such as combining our PCN with the Graph Convolutional Network [9][43].

Acknowledgement

References

[1] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 3, 6, 7

[2] Alexandru O Bălan and Michael J Black. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision*, pages 15–29. Springer, 2008. 2

[3] Alexandru O Balan, Leonid Sigal, Michael J Black, James E Davis, and Horst W Haussecker. Detailed human shape and pose from images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[4] 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*, Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016. 2

[5] Thomas J Cashman and Andrew W Fitzgibbon. What shape are dolphins? building 3d morphable models from 2d images. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):232–244, 2012. 2

[6] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016. 2

[7] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. 1996. 3

[8] Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. *Performance capture from sparse multi-view video*, volume 27. ACM, 2008. 2

[9] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS, pages 3844–3852, USA, 2016. Curran Associates Inc. 8

[10] Valentin Gabeur, Jean-Sébastien Franco Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3

[11] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision*, pages 484–499. Springer, 2016. 2

[12] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1381–1388. IEEE, 2009. 2

[13] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 38(2):14, 2019. 2

[14] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[15] Aaron S. Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1031–1039, 2017. 3

[16] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 7

[17] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2252–2261, 2019. 2

[18] Vladislav Kraevoy, Alla Sheffer, and Michiel van de Panne. Modeling from contour drawings. In *Proceedings of the 6th Eurographics Symposium on Sketch-Based interfaces and Modeling*, pages 37–44. ACM, 2009. 2

[19] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2, 3

[20] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *ACM siggraph computer graphics*, volume 21, pages 163–169. ACM, 1987. 3

[21] Xiaoguang Lu and Anil Jain. Deformation modeling for robust 3d face matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1346–1357, 2008. 2

[22] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2823–2832, 2017. 7

[23] 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. 3

[24] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew W Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, volume 11, pages 127–136, 2011. 3

[25] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 5

[26] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6, 7

[27] Victor Adrian Prisacariu, Olaf Kähler, Stuart Golodetz, Michael Sapienza, Tommaso Cavallari, Philip HS Torr, and

David W Murray. Infinitam v3: a framework for large-scale 3d reconstruction with loop closure. *arXiv preprint arXiv:1708.00783*, 2017. 3

[28] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3577–3586, 2017. 3

[29] Henry Roth and Marsette Vona. Moving volume kinectfusion. In *BMVC*, volume 20, pages 1–11, 2012. 3

[30] 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 *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 3

[31] Peter Shirley and Allan Tuchman. A polygonal approximation to direct scalar volume rendering. In *Proceedings of the 1990 Workshop on Volume Visualization*, VVS '90, pages 63–70, New York, NY, USA, 1990. ACM. 4

[32] Guang-hua Tan, Wei Chen, and Li-gang Liu. Image driven shape deformation using styles. *Journal of Zhejiang University SCIENCE C*, 11(1):27, 2010. 2

[33] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network. In *European Conference on Computer Vision*, pages 322–337. Springer, 2016. 2

[34] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2088–2096, 2017. 3

[35] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2626–2634, 2017. 2

[36] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. 2, 3, 6, 7

[37] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 6, 7

[38] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *ACM Transactions on Graphics (TOG)*, volume 27, page 97. ACM, 2008. 2, 6, 7

[39] Chunyu Wang, Yizhou Wang, Zhouchen Lin, and Alan L Yuille. Robust 3d human pose estimation from single images or video sequences. *IEEE transactions on pattern analysis and machine intelligence*, 41:1227–1241, 2018. 7

[40] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. In *Advances in neural information processing systems*, pages 540–550, 2017. 2

[41] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in neural information processing systems*, pages 82–90, 2016. 2

[42] 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. 3

[43] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 8

[44] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Advances in Neural Information Processing Systems*, pages 1696–1704, 2016. 3

[45] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7287–7296, 2018. 1

[46] Rui Zhu, Hamed Kiani Galoogahi, Chaoyang Wang, and Simon Lucey. Rethinking reprojection: Closing the loop for pose-aware shape reconstruction from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 57–65, 2017. 3