3DPeople: Modeling the Geometry of Dressed Humans

A. Pumarola[1] J. Sanchez-Riera[1] G. P. T. Choi[2] A. Sanfeliu[1] F. Moreno-Noguer[1]
[1]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
[2]John A. Paulson School of Engineering and Applied Sciences, Harvard University, USA

Figure 1: **3DPeople Dataset.** We present a synthetic dataset with 2 Million frames of 80 subjects (40 female/40 male) performing 70 different actions. The dataset contains a large range of distinct body shapes, skin tones and clothing outfits, and provides 640×480 RGB images under different viewpoints, 3D geometry of the body and clothing, 3D skeletons, depth maps, optical flow and semantic information (body parts and cloth labels). In this paper we use the 3DPeople dataset to model the geometry of dressed humans.

Abstract

Recent advances in 3D human shape estimation build upon parametric representations that model very well the shape of the naked body, but are not appropriate to represent the clothing geometry. In this paper, we present an approach to model dressed humans and predict their geometry from single images. We contribute in three fundamental aspects of the problem, namely, a new dataset, a novel shape parameterization algorithm and an end-to-end deep generative network for predicting shape.

First, we present 3DPeople*, a large-scale synthetic dataset with 2 Million photo-realistic images of 80 subjects performing 70 activities and wearing diverse outfits. Besides providing textured 3D meshes for clothes and body we annotated the dataset with segmentation masks, skeletons, depth, normal maps and optical flow. All this together makes 3DPeople suitable for a plethora of tasks.*

We then represent the 3D shapes using 2D geometry images. To build these images we propose a novel spherical area-preserving parameterization algorithm based on the optimal mass transportation method. We show this approach to improve existing spherical maps which tend to shrink the elongated parts of the full body models such as the arms and legs, making the geometry images incomplete.

Finally, we design a multi-resolution deep generative network that, given an input image of a dressed human, predicts his/her geometry image (and thus the clothed body shape) in an end-to-end manner. We obtain very promising results in jointly capturing body pose and clothing shape, both for synthetic validation and on the wild images.

1. Introduction

With the advent of deep learning, the problem of predicting the geometry of the human body from single images has experienced a tremendous boost. The combination of Convolutional Neural Networks with large Mo-

Cap datasets [44, 21], resulted in a substantial number of works that robustly predict the 3D position of the body joints [29, 30, 32, 36, 40, 49, 52, 56, 64].

In order to estimate the full body shape a standard practice adopted in [12, 14, 19, 24, 54, 66] is to regress the parameters of low rank parametric models [10, 28]. Nevertheless, while these parametric models describe very accurately the geometry of the naked body, they are not appropriate to capture the shape of clothed humans.

Current trends focus on proposing alternative representations to the low rank models. Varol *et al*. [55] advocate for a direct inference of volumetric body shape, although still without accounting for the clothing geometry. Very recently, [35] uses 2D silhouettes and the visual hull algorithm to recover shape and texture of clothed human bodies. Despite very promising results, this approach still requires frontal-view input images of the person with no background, and under relatively simple body poses.

In this paper, we introduce a general pipeline to estimate the geometry of dressed humans which is able cope with a wide spectrum of clothing outfits and textures, complex body poses and shapes, and changing backgrounds and camera viewpoints. For this purpose, we contribute in three key areas of the problem, namely, the data collection, the shape representation and the image-to-shape inference.

Concretely, we first present 3DPeople a new large-scale dataset with 2 Million photorealistic synthetic images of people under varying clothes and apparel. We split the dataset 40 male/40 female with different body shapes and skin tones, performing 70 distinct actions (see Fig. 1). The dataset contains 3D geometry of both the naked and dressed body, and additional annotations including skeletons, depth and normal maps, optical flow and semantic segmentation masks. This additional data is indeed very similar to SURREAL [56] which was built for similar purposes. The key difference between SURREAL and 3DPeople, is that in SURREAL the clothing is directly mapped as a texture on top of the naked body, while in 3DPeople the clothing does have its own geometry.

As essential as gathering a rich dataset, is the question of what is the most appropriate geometry representation for a deep network. In this paper we consider the "geometry image" proposed originally in [18] and recently used to encode rigid objects in [46, 47]. The construction of the geometry image involves two steps, first a mapping of a genus-0 surface onto a spherical domain, and then to a 2D grid resembling an image. Our contribution here is on the spherical mapping. We found that existing algorithms [13, 46] were not accurate, especially for the elongated parts of the body. To address this issue we devise a novel spherical area-preserving parameterization algorithm that combines and extends the FLASH [13] and the optimal mass transportation methods [33].

Our final contribution consists of designing a generative network to map input RGB images of a dressed human into his/her corresponding geometry image. Since we consider $128 \times 128 \times 3$ geometry images, learning such a mapping is highly complex. We alleviate the learning process through a coarse-to-fine strategy, combined with a series of geometry-aware losses. The full network is trained in an end-to-end manner, and the results are very promising in variety of input data, including both synthetic and real images.

2. Related work

3D Human shape estimation. While the problem of localizing the 3D position of the joints from a single image has been extensively studied [29, 30, 32, 36, 40, 45, 49, 52, 56, 64, 67], the estimation of the 3D body shape has received relatively little attention. This is presumably due to the existence of well-established datasets [44, 21], uniquely annotated with skeleton joints.

The inherent ambiguity for estimating human shape from a single view is typically addressed using shape embeddings learned from body scan repositories like SCAPE [10] and SMPL [28]. The body geometry is described by a reduced number of pose and shape parameters, which are optimized to match image characteristics [11, 12, 27]. Dibra *et al*. [14] are the first in using a CNN fed with silhouette images to estimate shape parameters. In [50, 54] SMPL body parameters are predicted by incorporating differential renders into the deep network to directly estimate and minimize the error of image features. On top of this, [24] introduces an adversarial loss that penalizes non-realistic body shapes. Very recently [6, 8] extended the SMPL parametric representation to model cloth and [7] used shape from shading and better texture merging to predict higher details.

Non-parametric representations for 3D objects. What is the most appropriate 3D object representation to train a deep network remains an open question, especially for non-rigid bodies. Standard non-parametric representations for rigid objects include voxels [16, 63], octrees [51, 59, 60] and point-clouds [53]. [46, 47] uses 2D embeddings computed with geometry images [18] to represent rigid objects. Interestingly, very promising results for the reconstruction of non-rigid hands were also reported. DeformNet [38] proposes the first deep model to reconstruct the 3D shape non-rigid surfaces from a single image. Bodynet [55] explores a network that predicts voxelized human body shape. Very recently, [35] introduces a pipeline that given a single image of a person in frontal position predicts the body silhouette as seen from different views, and then uses a visual hull algorithm to estimate 3D shape.

Generative Adversarial Networks. Originally introduced by [17], GANs have been used to model human body distributions and generate novel images of a person under arbi-

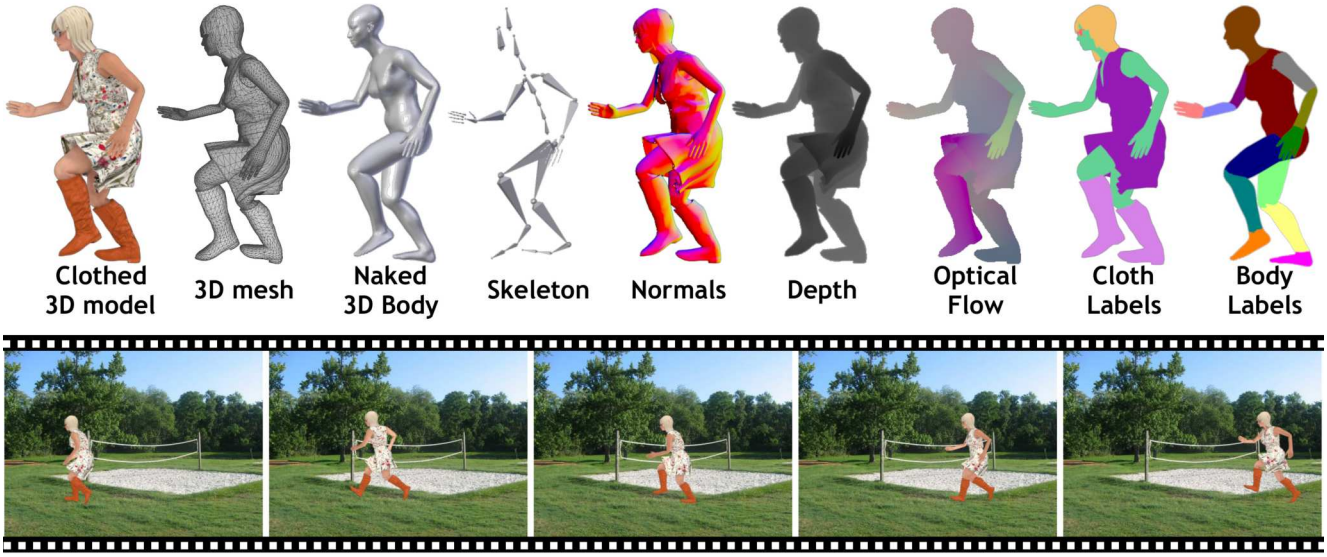


Figure 2. **Annotations of the 3D People Dataset.** For each of the 80 subjects of the dataset, we generate 280 video sequences (70 actions seen from 4 camera views). The bottom of the figure shows 5 sample frames of the *Running* sequence. Every RGB frame is annotated with the information reported in the top of the figure. 3DPeople is the first large-scale dataset with geometric meshes of body and clothes.

trary poses [39]. Kanazawa *et al*. [24] explicitly learned the distribution on SMPL parameters. DVP [25], paGAN [34] and GANimation [37] presented models for continuous face animation and manipulation. GANs have also been applied to edit [20, 48, 58] and generate [15] talking faces.

Datasets for body shape analysis. Datasets are fundamental in the deep-learning era. While obtaining annotations is quite straightforward for 2D poses [43, 9, 23], it requires using sophisticated MoCap systems for the 3D case. Additionally, the datasets acquired this way [44, 21, 21] are mostly indoors. Even more complex is the task of obtaining 3D body shape, which requires expensive setups with muti-cameras or 3D scanners. Marcard *et al*. [57] proposed solution based on IMUs and a moving camera but still does not provide perfect ground-truth annotation. To overcome this situation, datasets with synthetic but photo-realistic images have emerged as a tool to generate massive amounts of training data. SURREAL [56] is the largest and more complete dataset so far, with more than 6M frames generated by projecting synthetic textures of clothes onto random SMPL body shapes. The dataset is further annotated with body masks, optical flow and depth. However, since clothes are projected onto the naked SMPL shapes just as textures, they cannot be explicitly modeled. To fill this gap, we present the 3DPeople dataset of 3D dressed humans in motion.

3. 3DPeople dataset

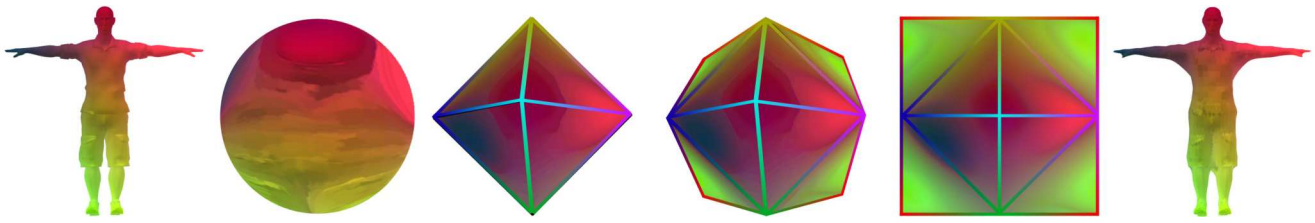
We next introduce 3DPeople, the first dataset of dressed humans with specific geometry representation for the clothes. The dataset contains 2 Million photorealistic 640×480 images split into 40 male/40 female performing 70 actions. For every subject-action sequence we randomly change the texture of the clothes, the lighting direction and the background, and capture it from 4 camera views. Each frame is annotated with (see Fig. 2): 3D textured mesh of the naked and dressed body; 3D skeleton; normals; body part and cloth segmentation masks; depth map; optical flow; and camera parameters. In the following we describe the generation process:

Body models: We have generated fully textured triangular meshes for 80 human characters using Adobe Fuse [1] and MakeHuman [2]. The distribution of the subjects physical characteristics cover a broad spectrum of body shapes, skin tones and hair geometry (see Fig. 1).

Clothing models: Each subject is dressed with a different outfit including a variety of garments, combining tight and loose clothes. Additional apparel like sunglasses, hats and caps are also included. The final rigged meshes of the body and clothes contain approximately 20K vertices.

Mocap sequences: We gather 70 realistic motion sequences from Mixamo [3]. These include human movements with different complexity, from *drinking* and *typing* actions that produce small body motions to actions like *breakdance* or *backflip* that involve very complex patterns. The mean length of the sequences is of 110 frames. While these are relatively short sequences, they have a large expressivity, which we believe make 3DPeople also appropriate for exploring action recognition tasks.

Textures, camera, lights and background: We then use Blender [4] to apply the 70 MoCap animation sequences to



(a) (b) (c) (d) (e) (f)

Figure 3. **Geometry image representation of the reference mesh**. (a) Reference mesh in a tpose configuration color coded using the xyz position. (b) Spherical parameterization; (c) Octahedral parameterization; (d) Unwarping the octahedron to a planar configuration; (e) Geometry image, resulting from the projection of the octahedron onto a plane; (f) mesh reconstructed from the geometry image. Colored edges in the octahedron and in the geometry image represent the symmetry that is later exploited by the mesh regressor Φ.

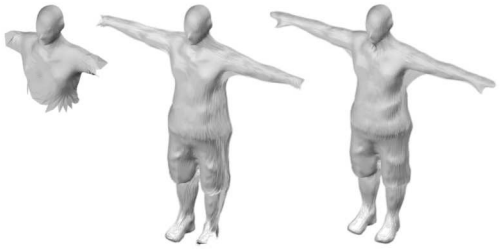


Figure 4. **Comparison of spherical mapping methods.** Shape reconstructed from a geometry image obtained with three different algorithms. Left: FLASH [13]; Center: [46]; Right: SAPP algorithm we propose. Note that SAPP is the only method that can effectively recover feet and hands.

each character. Every sequence is rendered from 4 camera views, yielding a total of 22,400 clips. We use a projective camera with a 700 mm focal length and 640×480 pixel resolution. The 4 viewpoints correspond approximately to orthogonal directions aligned with the ground. The distance to the subject changes for every sequence to ensure a full view of the body in all frames. The textures of the clothes are randomly changed for every sequence (see again Fig. 1). The illumination is composed of an ambient lighting plus a light source at infinite, which direction is changed per sequence. As in [56] we render the person on top of a static background image, randomly taken from the LSUN dataset [65].

Semantic labels: For every rendered image, we provide segmentation labels of the clothes (8 classes) and body (14 parts). Observe in Fig. 2-top-right that the former are aligned with the dressed human, while the body parts are aligned with the naked body.

4. Problem formulation

Given a single image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ of a person wearing an arbitrary outfit, we aim at designing a model capable of directly estimating the 3D shape of the clothed body. We represent the body shape through the mesh associated to a geometry image with N^2 vertices $\mathbf{X} \in \mathbb{R}^{N \times N \times 3}$ where $\mathbf{x}_i = (x_i, y_i, z_i)$ are the 3D coordinates of the i-th vertex, expressed in the camera coordinates system and centered on the root joint $\mathbf{x}_r$. This representation is a key ingredient of our design, as it maps the 3D mesh to a regular 2D grid structure that preserves the neighborhood relations, fulfilling thus the locality assumption required in CNN architectures. Furthermore, the geometry image representation allows uniformly reducing/increasing the mesh resolution by simply uniformly downsampling/upsampling. This will play an important role in our strategy of designing a coarse-to-fine shape estimation approach.

We next describe the two main steps of our pipeline: 1) the process of constructing the geometry images, and 2) the deep generative model we propose for predicting 3D shape.

5. Geometry image for dressed humans

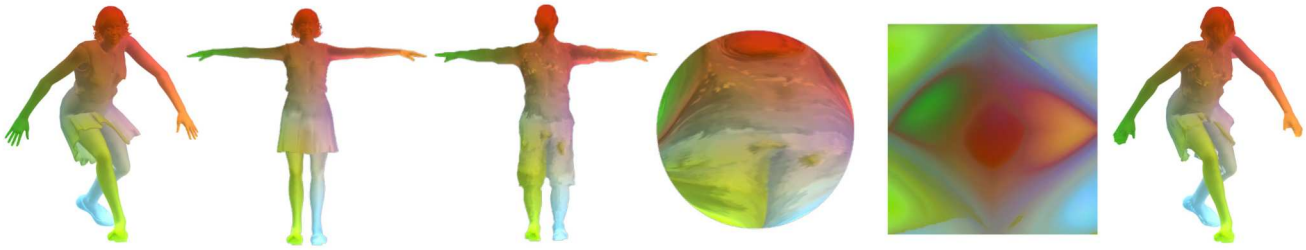
The deep network we describe later will be trained using pairs $\{\mathbf{I}, \mathbf{X}\}$ of images and their corresponding geometry image. For creating the geometry images we consider two different cases, one for a reference mesh in a tpose configuration, and another for any other mesh of the dataset.

5.1. Geometry image for a reference mesh

One of the subjects of our dataset in a tpose configuration is chosen as a reference mesh. The process for mapping this mesh into a planar regular grid is illustrated in Fig. 3. It involves the following steps:

Repairing the mesh. Let $\mathbf{R}^{\text{tpose}} \in \mathbb{R}^{N_R \times 3}$ be the reference mesh with N_R vertices in a tpose configuration (Fig. 3-a). We assume this mesh to be a manifold mesh and to be genus-0. Most of the meshes in our dataset, however, do not fulfill these conditions. In order to fix the mesh we follow the heuristic described in [46] which consists of a voxelization, a selection of the largest connected region of the α-shape, and subsequent hole filling using a medial axis approach. We denote by $\tilde{\mathbf{R}}^{\text{tpose}}$ the repaired mesh.

Spherical parameterization. Given the repaired genus-0 mesh $\tilde{\mathbf{R}}^{\text{tpose}}$, we next compute the spherical parameterization $\mathcal{S} : \tilde{\mathbf{R}}^{\text{tpose}} \rightarrow \mathbf{S}$ that maps every vertex of $\tilde{\mathbf{R}}^{\text{tpose}}$ onto the unit sphere $\mathbf{S}$ (Fig. 3-b). Details of the algorithm we use are explained below.



(a) (b) (c) (d) (e) (f)

Figure 5. **Geometry image estimation for an arbitrary mesh**. (a) Input mesh $\mathbf{Q}$ in an arbitrary pose color coded using the xyz position of the vertices; (b) Same mesh in a tpose configuration ($\mathbf{Q}^{\text{tpose}}$). The color of the mesh is mapped from $\mathbf{Q}$; (c) Reference tpose $\mathbf{R}^{\text{tpose}}$. The colors again correspond from those transferred from $\mathbf{Q}$ through the non-rigid map between $\mathbf{Q}^{\text{tpose}}$ and $\mathbf{R}^{\text{tpose}}$; (d) Spherical mapping of $\mathbf{Q}$; (e) Geometry image of $\mathbf{Q}$; (f) Mesh reconstructed from the geometry image. Note that while being computed through a non-rigid mapping between the two reference poses, the recovered shape is a very good approximation of the input mesh $\mathbf{Q}$.

Unfolding the sphere. The sphere $\mathbf{S}$ is mapped onto an octahedron and then cut along edges to output a flat geometry image $\mathbf{X}$. Let us formally denote by $\mathcal{U} : \mathbf{S} \rightarrow \mathbf{X}$, and by $\mathcal{G}^R = \mathcal{U} \circ \mathcal{S} : \tilde{\mathbf{R}}^{\text{tpose}} \rightarrow \mathbf{X}$ the mapping from the reference mesh to the geometry image. The unfolding process is shown in Fig. 3-(c,d,e). Color lines in the geometry image correspond to the same edge in the octahedron, and are split after the unfolding operation. We will later enforce this symmetry constraint when predicting geometry images.

5.2. Spherical area-Preserving parameterization

Although there exist several spherical parameterization schemes (*e.g.* [13, 46]) we found that they tend to shrink the elongated parts of the full body models such as the arms and legs, making the geometry images incomplete (see Fig. 4). In this work, we develop a spherical area-preserving parameterization algorithm for genus-0 full body models by combining and extending the FLASH method [13] and the optimal mass transportation method [33]. Our algorithm is particularly advantageous for handling models with elongated parts. The key idea is to begin with an initial parameterization onto a planar triangular domain with a suitable rescaling correcting the size of it. The area distortion of the initial parameterization is then reduced using quasi-conformal composition. Finally, the spherical area-preserving parameterization is produced using optimal mass transportation followed by the inverse stereographic projection. We provide further details in the supplemental material.

5.3. Geometry image for arbitrary meshes

The approach for creating the geometry image described in the previous subsection is quite computationally demanding (up to 15 minutes for complex meshes). To compute the geometry image for several thousand training meshes we have devised an alternative approach. Let $\mathbf{Q} \in \mathbb{R}^{N_Q \times 3}$ be the mesh of any subject of the dataset under an arbitrary pose (Fig. 5-a), and let $\mathbf{Q}^{\text{tpose}} \in \mathbb{R}^{N_Q \times 3}$ be its tpose configuration (Fig. 5-b). We assume there is a 1-to-1 vertex correspondence between both meshes, that is, $\exists\ \mathcal{I} : \mathbf{Q} \rightarrow \mathbf{Q}^{\text{tpose}}$ where $\mathcal{I}$ is a known bijective function[1]. We then compute dense correspondences between $\mathbf{Q}^{\text{tpose}}$ and the reference tpose $\tilde{\mathbf{R}}^{\text{tpose}}$, using a nonrigid icp algorithm [5]. We denote this mapping as $\mathcal{N} : \mathbf{Q}^{\text{tpose}} \rightarrow \tilde{\mathbf{R}}^{\text{tpose}}$ (see Fig. 5-c). We can then finally compute the geometry image for the input mesh $\mathbf{Q}$ by concatenating mappings:

$$\mathcal{G}^Q = \mathcal{G}^R \circ \mathcal{N} \circ \mathcal{I} : \mathbf{Q} \rightarrow \mathbf{X} \tag{1}$$

where $\mathcal{G}^R$ is the mapping from the reference mesh to the geometry image domain estimated in Sec. 5.1. It is worth pointing that the nonrigid icp between the pairs of tposes is also highly computationally demanding, but it only needs to be computed once per every subject of the dataset. Once this is done, the geometry image for a new input mesh $\mathbf{Q}$ can be created in a few seconds.

An important consequence of this procedure is that all geometry images of the dataset will be semantically aligned, that is, every uv entry in $\mathbf{X}$ will correspond to (approximately) the same semantic part of the model. This will significantly alleviate the learning task of the deep network.

6. GimNet

We next introduce *GimNet*, our deep generative network to estimate geometry images (and thus 3D shape) of dressed humans from a single image. An overview of the model is shown in Fig. 6. Given the input image, we first extract the 2D joint locations $\mathbf{p}$ represented as heatmaps [62, 38], which are then fed into a mesh regressor $\Phi(\mathbf{I}, \mathbf{p})$ trained to reconstruct the shape $\hat{\mathbf{X}}$ of the person in $\mathbf{I}$ employing a geometry image based representation. Due to the high complexity of the mapping (both $\mathbf{I}$ and $\hat{\mathbf{X}}$ are of size $128 \times 128 \times 3$), the regressor operates in a coarse-to-fine manner, progressively reconstructing meshes at higher resolution. To further enforce the reconstruction to lie on the manifold of anthropomorphic shapes, an adversarial scheme with two discriminators D is applied.

[1] This is guaranteed in our dataset, with all meshes of the same subject having the same number of vertices.

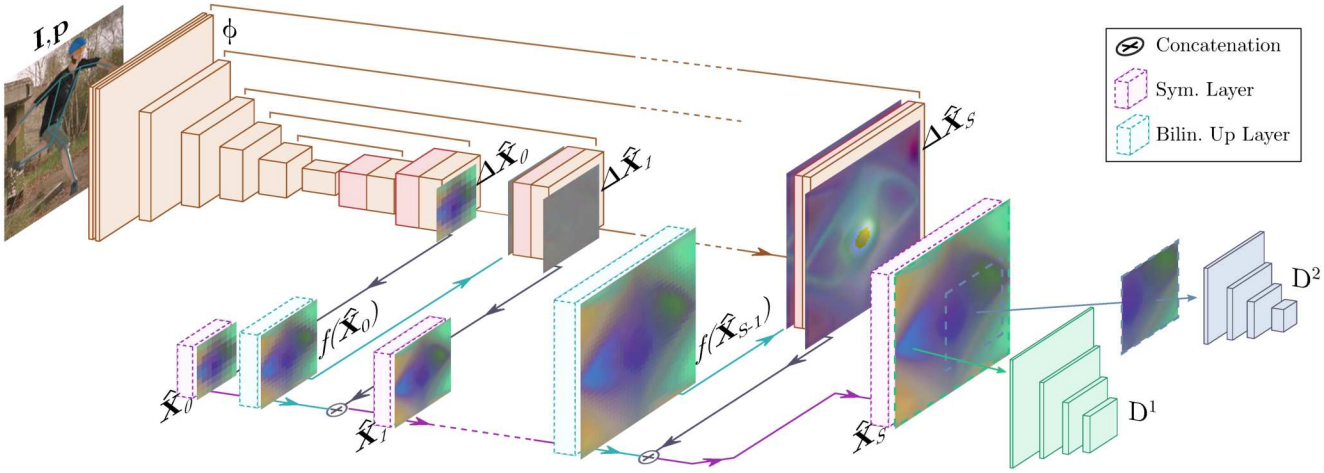


Figure 6. **Overview GimNet**. The proposed architecture consists of two main blocks: a multiscale geometry image regressor Φ and a multiscale discriminator D to evaluate the local and global consistency of the estimated meshes.

6.1. Model architecture

Mesh regressor. Given the input image $\mathbf{I}$ and the estimated 2D body joints $\mathbf{p}$, the mesh regressor Φ aims to predict the geometry image $\mathbf{X}$, *i.e.* we seek to estimate the mapping $\mathcal{M} : \mathbf{I}, \mathbf{p} \rightarrow \mathbf{X}$. Instead of directly learning the complex mapping $\mathcal{M}$, we break the process into a sequence of more manageable steps. Φ initially estimates a low-resolution mesh, and then progressively increases its resolution (see Fig. 6). This coarse-to-fine approach allows the regressor to first focus on the basic shape configuration and then shift attention to finer details, while also providing more stability compared to a network that learns the direct mapping.

As shown in Fig. 3-e, the geometry images have symmetry properties derived from unfolding the octahedron into a square, specifically, each side of the geometry image is symmetric with respect to its midpoint. We force this property using a differentiable layer that linearly operates over the edges of the estimated geometry images.

Multi-Scale Discriminator. Evaluating high-resolution meshes poses a significant challenge for a discriminator, as it needs to simultaneously guarantee local and global mesh consistency on very high dimensional data. We therefore use two discriminators with the same architecture, but that operate in different geometry image scales: (i) a discriminator with a large receptive field that evaluates the shape coherence as a whole; and (ii) a local discriminator that focuses on small patches and enforces the local consistency of the surface triangle faces.

6.2. Learning the model

3D reconstruction error. We first define a supervised multi-level L1 loss for 3D reconstruction $\mathcal{L}_{\text{R}}$ as:

$$\mathcal{L}_{\text{R}} = \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_r,\hat{\mathbf{X}}\sim\mathbb{P}_g} \frac{1}{S}\sum_{s=1}^{S} \lambda_s \left\| \mathbf{X}_s - \hat{\mathbf{X}}_s \right\|_1, \qquad (2)$$

being $\mathbb{P}_r$ and $\mathbb{P}_g$ the real and generated data distribution of clothed human geometry images respectively, S the number of scales, $\mathbf{X}_s$ the ground-truth reconstruction at scale s and $\hat{\mathbf{X}}_s = \Phi_s(\mathbf{I})$ the estimated reconstruction. The error at each scale is weighted by $\lambda_s = \frac{1}{r}$ where r is the ratio between $\hat{\mathbf{X}}_S$ and $\hat{\mathbf{X}}_s$ sizes. During initial experimentation L1 loss reported better reconstructions than mean squared error.

2D Projection Error. To encourage the mesh to correctly project onto the input image we penalize, at every scale s, its projection error $\mathcal{L}_{\text{P}}$ computed as:

$$\mathcal{L}_{\text{P}} = \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_r,\hat{\mathbf{X}}\sim\mathbb{P}_g} \frac{1}{S}\sum_{s=1}^{S} \lambda_s \left\| \mathcal{P}(\mathbf{X}_s) - \mathcal{P}(\hat{\mathbf{X}}_s) \right\|_1$$

where $\mathcal{P}$ is the differentiable projection equation and λ_s is calculated as above.

Adversarial loss. In order to further enforce the mesh regressor Φ to generate anthropomorphic shapes we perform a min-max strategy game [17] between the regressor and two discriminators operating at different scales. It is well-known that non-overlapping support between the true data distribution and model distributions can cause severe training instabilities. As proven by [42, 31], this can be addressed by penalizing the discriminator when deviating from the Nash-equilibrium, ensuring that its gradients are non-zero orthogonal to the data manifold. Formally, being D^k the k^{th} discriminator, the $\mathcal{L}_{\text{adv}}$ loss is defined as:

$$\sum_{k=1}^{K} \Big[\mathbb{E}_{\hat{\mathbf{X}}\sim\mathbb{P}_g}[\log(1 - D^k(\hat{\mathbf{X}}_S))] + \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_r}\left[\log(D^k(\mathbf{X}_S))\right] + \frac{\lambda_{\text{dgp}}}{2}\mathbb{E}_{\mathbf{X}\sim\mathbb{P}_r}(\|\nabla D^k(\mathbf{X}_S)\|_1^2 \Big], \qquad (3)$$

where λ_{dgp} is a penalty regularization for discriminator gradients, only considered on the true data distribution.

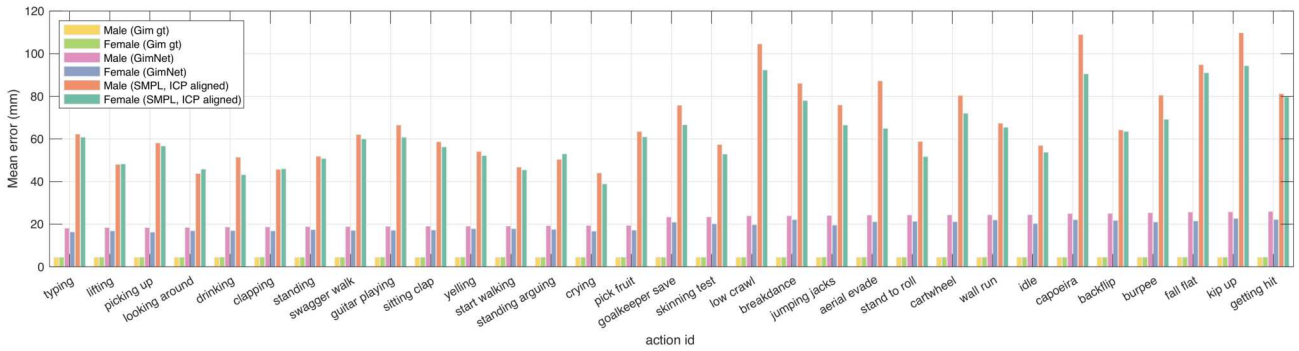


Figure 7. **Mean Error Distance on the test set.** We plot the results for the 15 worst and 15 best actions. Besides the results of GimNet, we report the results obtained by the ground truth GIM (recall that it is an approximation of the actual ground truth mesh). We also display the results obtained by the recent parametric approach of [24]. The results of this method, however are merely indicative, as we did not retrain the network with our dataset.

Feature matching loss. To improve training stabilization we penalize higher level features on the discriminators [61]. Similar to a perception loss, the estimated geometry image is compared with the ground truth at multiple feature levels of the discriminators. Being D_l^k the l^{th} layer of the k^{th} discriminator, $\mathcal{L}_{\text{F}}$ is defined as:

$$\mathbb{E}_{\mathbf{X}\sim\mathbb{P}_r,\hat{\mathbf{X}}\sim\mathbb{P}_g}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{N_l^k}\left\|D_l^k(\mathbf{X}_S)-D_l^k(\hat{\mathbf{X}}_S)\right\|_1, \quad (4)$$

where N_l^k is a weight regularizer denoting the number of elements in the l^{th} layer of the k^{th} discriminator.

Total Loss. Finally, we to solve the min-max problem:

$$\Phi^\star = \arg\min_{\Phi}\max_{D}\mathcal{L}_{\text{adv}}+\lambda_{\text{R}}\mathcal{L}_{\text{R}}+\lambda_{\text{P}}\mathcal{L}_{\text{P}}+\lambda_{\text{F}}\mathcal{L}_{\text{F}} \quad (5)$$

where λ_{R}, λ_{P} and λ_{F} are the hyper-parameters that control the relative importance of every loss term.

6.3. Implementation details

For the mesh regressor Φ we build upon the U-Net architecture [41] consisting on an encoder-decoder structure with skip connections between features at the same resolution extended to estimate geometry images at multiple scales. Detailed explanation of its architecture can be found in the supplemental material.

Both discriminator networks operate at different mesh resolutions [61] but have the same PatchGan [22] architecture mapping from the geometry image $\mathbf{X}$ to a matrix $\mathbf{Y}\in\mathbb{R}^{H/8\times W/8}$, where $\mathbf{Y}[i,j]$ represents the probability of the patch ij to be close to a real geometry image distribution. The global discriminator evaluates the final mesh resolution at scale S and the local discriminator the downsampled mesh at scale $S-1$. Detailed explanation of their architecture can be found in the supplemental material.

The model is trained with 170,000 synthetic images of cropped clothed people resized to 128×128 pixels and geometry images of $128\times128\times3$ (meshes with 16,384 vertices) during 60 epochs and $S=4$. As for the optimizer, we use Adam [26] with learning rate of $2e-4$, beta1 0.5, beta2 0.999 and batch size 110. Every 40 epochs we decay the learning rate by a factor of 0.5. The weight coefficients for the loss terms are set to $\lambda_{\text{R}}=20$, $\lambda_{\text{P}}=0.1$, $\lambda_{\text{F}}=10$ and $\lambda_{\text{dgp}}=0.01$.

7. Experimental evaluation

We next present quantitative and qualitative results on synthetic images of our dataset and on images in the wild.

Synthetic Results. We evaluate our approach on 25,000 test images randomly chosen for 8 subjects (4 male/ 4 female) of the test split. For each test sample we feed GimNet with the RGB image and the ground truth 2D pose, corrupted by Gaussian noise with 2 pixel std. For a given test sample, let $\hat{\mathbf{Y}}$ be the $N^2\times3$ estimated mesh, resulting from a direct reshaping of its estimated geometry image $\hat{\mathbf{X}}$. Also, let $\mathbf{Y}$ be the ground truth mesh, which does not need to have neither the same number of vertices as $\hat{\mathbf{Y}}$, nor necessarily the same topology. Since there is no direct 1-to-1 mapping between the vertices of the two meshes we propose using the following metric:

$$\text{dist}(\hat{\mathbf{Y}},\mathbf{Y})=\frac{1}{2}(\text{KNN}(\hat{\mathbf{Y}}\rightarrow\mathbf{Y})+\text{KNN}(\mathbf{Y}\rightarrow\hat{\mathbf{Y}})) \quad (6)$$

where $\text{KNN}(\hat{\mathbf{Y}}\rightarrow\mathbf{Y})$ represents the average Euclidean distance for all vertices of $\hat{\mathbf{Y}}$ to their nearest neighbor in $\mathbf{Y}$. Note that $\text{KNN}(\cdot,\cdot)$ is not a true distance measure because it is not symmetric. This is why we compute it bidirectionally.

The quantitative results are summarized in Fig. 7. We report the average error (in mm) of GimNet for 30 actions (the 15 with the highest and lowest error). Note that the error of GimNet is bounded between 15 and 35mm. Recall, however, that we do not consider outlier 2D detections in our experiments, but just 2D noise. We also evaluate the error of

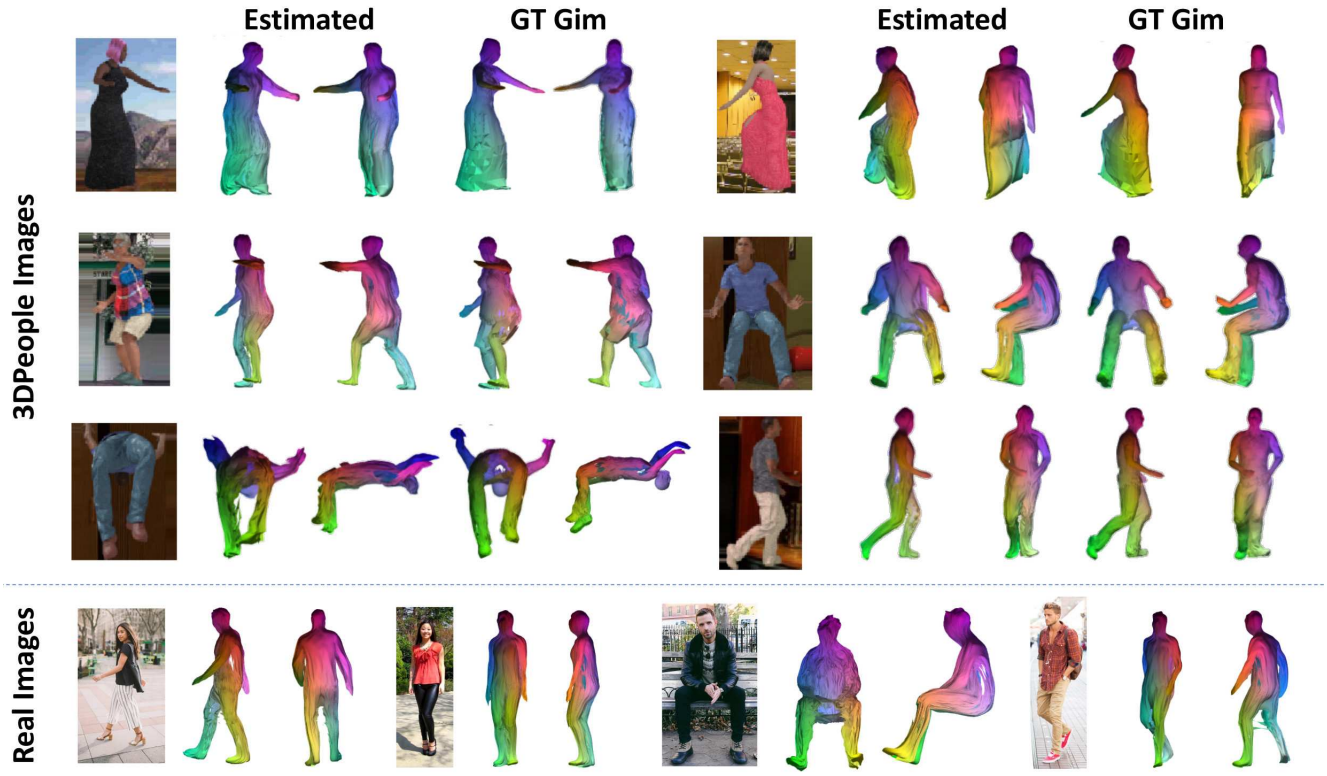


Figure 8. **Qualitative results.** For the synthetic images we plot our estimated results and the shape reconstructed directly from the ground truth geometry image. In all cases we show two different views. The color of the meshes encodes the xyz vertex position.

the ground truth geometry image, as it is an approximation of the actual ground truth mesh. This error is below 5mm, indicating that the geometry image representation does indeed capture very accurately the true shape. Finally, we also provide the error of the recent parametric approach of [24], that fits SMPL parameters to the input images. Nevertheless, these results are just indicative, and cannot be directly compared with our approach, as we did not retrain [24]. We add them here just to demonstrate the challenge posed by the new 3DPeople dataset. Indeed, the distance error in [24] was computed after performing a rigid-icp of the estimated mesh with the ground truth mesh (there was no need of this for GimNet).

Qualitative Results. We finally show in Fig. 8 qualitative results on synthetic images from 3DPeople and real fashion images downloaded from Internet. Remarkably, note how our approach is able to reconstruct long dresses (top row images), known to be a major challenge [35]. Note also that some of the reconstructed meshes have spikes. This is one of the limitations of the non-parametric models, that the reconstructions tend to be less smooth than when using parametric fittings. However, non-parametric models have also the advantage that, if properly trained, can span a much larger configuration space.

8. Conclusions

In this paper we have made three contributions to the problem of reconstructing the shape of dressed humans: 1) we have presented the first large-scale dataset of 3D humans in action in which cloth geometry is explicitly modelled; 2) we have proposed a new algorithm to perform spherical parameterizations of elongated body parts, to later model rigged meshes of human bodies as geometry images; and 3) we have introduced an end-to-end network to estimate human body and clothing shape from single images, without relying on parametric models. While the results are very promising, there are still several avenues to explore. For instance, extending the problem to video, exploring new regularization schemes on the geometry images, or combining segmentation and 3D reconstruction are all open problems that can benefit from the proposed 3DPeople dataset.

9. Acknowledgements

This work is supported in part by an Amazon Research Award, the Croucher Foundation and the Spanish MiNeCo under projects HuMoUR TIN2017-90086-R, ColRobTransp DPI2016-78957-R and María de Maeztu Seal of Excellence MDM-2016-0656. We also thank Nvidia for hardware donation under the GPU Grant Program.

References

[1] https://www.adobe.com/es/products/fuse.html. 3

[2] http://www.makehumancommunity.org/. 3

[3] https://www.mixamo.com/. 3

[4] Blender - a 3d modelling and rendering package. https://www.blender.org/. 3

[5] Nonrigid ICP, MATLAB Central File Exchange, 2019. https://www.mathworks.com/matlabcentral/fileexchange/41396-nonrigidicp/, 2019. 5

[6] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *CVPR*, 2019. 2

[7] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *3DV*, 2018. 2

[8] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, 2018. 2

[9] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *CVPR*, June 2014. 3

[10] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape Completion and Animation of People. *ACM*, 2005. 2

[11] Alexandru O Balan, Leonid Sigal, Michael J Black, James E Davis, and Horst W Haussecker. Detailed human shape and pose from images. In *CVPR*, 2007. 2

[12] 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 *ECCV*, 2016. 2

[13] Pui Tung Choi, Ka Chun Lam, and Lok Ming Lui. FLASH: Fast landmark aligned spherical harmonic parameterization for genus-0 closed brain surfaces. *SIAM J. Imaging Sci.*, 8(1):67–94, 2015. 2, 4, 5

[14] Endri Dibra, Himanshu Jain, Cengiz ztireli, Remo Ziegler, and Markus Gross. Human Shape from Silhouettes using Generative HKS Descriptors and Cross-modal Neural Networks. In *CVPR*, 2017. 2

[15] Amanda Duarte, Francisco Roldan, Miquel Tubau, Janna Escur, Santiago Pascual, Amaia Salvador, Eva Mohedano, Kevin McGuinness, Jordi Torres, and Xavier Giro-i Nieto. Wav2pix: Speech-conditioned face generation using generative adversarial networks. In *ICASSP*, 2019. 3

[16] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016. 2

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *NIPS*, 2014. 2, 6

[18] Xianfeng Gu, Steven J Gortler, and Hugues Hoppe. Geometry images. In *TOG*, volume 21, pages 355–361. ACM, 2002. 2

[19] Peng Guan, Alexander Weiss, Alexandru O. Balan, and Michael Black. Estimating Human Shape and Pose from a Single Image. In *ICCV*, 2009. 2

[20] Ziwei Liu Ping Luo Xiaogang Wang Hang Zhou, Yu Liu. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, 2019. 3

[21] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *PAMI*, 36(7):1325–1339, 2014. 2, 3

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image Translation with Conditional Adversarial Networks. In *CVPR*, 2017. 7

[23] Sam Johnson and Mark Everingham. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *BMVC*, 2010. 3

[24] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end Recovery of Human Shape and Pose. In *CVPR*, 2018. 2, 3, 7, 8

[25] Hyeongwoo Kim, Pablo Carrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *TOG*, 2018. 3

[26] Diederik Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *ICLR*, 2015. 7

[27] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the People: Closing the Loop between 3D and 2D human representations. In *CVPR*, 2017. 2

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *TOG*, 34(6):248:1–248:16, Oct. 2015. 2

[29] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 2

[30] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *TOG*, 36(4), 2017. 2

[31] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which Training Methods for GANs do actually Converge? In *ICML*, 2018. 6

[32] Francesc Moreno-Noguer. 3D Human Pose Estimation from a Single Image via Distance Matrix Regression. In *CVPR*, 2017. 2

[33] Saad Nadeem, Zhengyu Su, Wei Zeng, Arie Kaufman, and Xianfeng Gu. Spherical Parameterization Balancing Angle and Area Distortions. *TVCG*, 23(6):1663–1676, 2017. 2, 5

[34] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. pagan: real-time avatars using dynamic textures. In *SIGGRAPH Asia*, page 258. ACM, 2018. 3

[35] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. SiCloPe: Silhouette-Based Clothed People. In *CVPR*, 2019. 2, 8

[36] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 2017. 2

[37] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *ECCV*, 2018. 3

[38] Albert Pumarola, Antonio Agudo, Lorenzo Porzi, Alberto Sanfeliu, Vincent Lepetit, and Francesc Moreno-Noguer. Geometry-aware network for non-rigid shape prediction from a single view. In *CVPR*, 2018. 2, 5

[39] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *CVPR*, 2018. 3

[40] Grégory Rogez and Cordelia Schmid. MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. In *NIPS*, 2016. 2

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 7

[42] Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *NIPS*, 2017. 6

[43] Ben Sapp and Ben Taskar. Modec: Multimodal Decomposable Models for Human Pose Estimation. In *CVPR*, 2013. 3

[44] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. *IJCV*, 2010. 2, 3

[45] Edgar Simo-Serra, Ariadna Quattoni, Carme Torras, and Francesc Moreno-Noguer. A joint model for 2d and 3d pose estimation from a single image. In *CVPR*, 2013. 2

[46] Ayan Sinha, Jing Bai, and Karthik Ramani. Deep learning 3D Shape Surfaces using Geometry Images. In *ECCV*, 2016. 2, 4, 5

[47] Ayan Sinha, Asim Unmesh, Qixing Huang, and Karthik Ramani. SurfNet: Generating 3D shape surfaces using deep residual networks. In *CVPR*, 2017. 2

[48] Yang Song, Jingwen Zhu, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*, 2018. 3

[49] Xiao Sun, Bin Xiao, Shuang Liang, and Yichen Wei. Integral Human Pose Regression. In *ECCV*, 2018. 2

[50] Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect Deep structured Learning for 3D Human Body Shape and Pose Prediction. In *BMVC*, 2017. 2

[51] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree Generating Networks: Efficient Convolutional Architectures for High-resolution 3D Outputs. In *ICCV*, 2017. 2

[52] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3D pose estimation from a single image. In *CVPR*, 2017. 2

[53] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. A Point Set Generation Network for 3D Object Reconstruction from a Single Image. In *CVPR*, 2017. 2

[54] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *NIPS*, 2017. 2

[55] Gul 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

[56] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2, 3, 4

[57] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 3

[58] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans. In *BMVC*, 2018. 3

[59] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-CNN: Octree-based Convolutional Neural Networks for 3D Shape Analysis. *TOG*, 36(4), 2017. 2

[60] Peng-Shuai Wang, Chun-Yu Sun, Yang Liu, and Xin Tong. Adaptive O-CNN: A Patch-based Deep Representation of 3D Shapes. *TOG*, 37(6), 2018. 2

[61] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 7

[62] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional Pose Machines. In *CVPR*, 2016. 5

[63] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective Transformer Nets: Learning Single-view 3D object Reconstruction without 3D Supervision. In *NIPS*, 2016. 2

[64] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *CVPR*, 2018. 2

[65] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *arXiv:1506.03365*, 2015. 4

[66] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. In *NeurIPS*, 2018. 2

[67] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *ICCV*, 2017. 2