Learning to Dress 3D People in Generative Clothing

Qianli Ma[1], Jinlong Yang[1], Anurag Ranjan[1,2], Sergi Pujades[4], Gerard Pons-Moll[5],
Siyu Tang[*3], and Michael J. Black[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany   [2]University of Tübingen, Germany
[3]ETH Zürich, Switzerland   [4]Université Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, France
[5]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
{qma, jyang, aranjan, black}@tue.mpg.de
sergi.pujades-rocamora@inria.fr   gpons@mpi-inf.mpg.de   siyu.tang@inf.ethz.ch
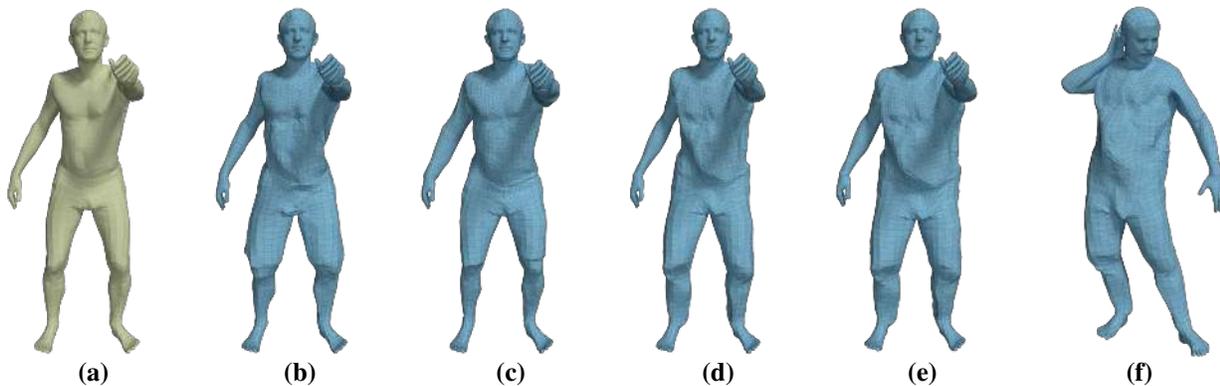
Figure 1: **CAPE** model for clothed humans. Given a SMPL body shape and pose **(a)**, CAPE adds clothing by randomly sampling from a learned model **(b, c)**, can generate different clothing types — shorts in **(b, c)** vs. long-pants in **(d)**. The generated clothed humans can generalize to diverse body shapes **(e)** and body poses **(f)**. Best viewed zoomed-in on screen.

Abstract

Three-dimensional human body models are widely used in the analysis of human pose and motion. Existing models, however, are learned from minimally-clothed 3D scans and thus do not generalize to the complexity of dressed people in common images and videos. Additionally, current models lack the expressive power needed to represent the complex non-linear geometry of pose-dependent clothing shapes. To address this, we learn a generative 3D mesh model of clothed people from 3D scans with varying pose and clothing. Specifically, we train a conditional Mesh-VAE-GAN to learn the clothing deformation from the SMPL body model, making clothing an additional term in SMPL. Our model is conditioned on both pose and clothing type, giving the ability to draw samples of clothing to dress different body shapes in a variety of styles and poses. To preserve wrinkle detail, our Mesh-VAE-GAN extends patchwise discriminators to 3D meshes. Our model, named CAPE, represents global shape and fine local structure, effectively extending the SMPL body model to clothing. To our knowl-edge, this is the first generative model that directly dresses 3D human body meshes and generalizes to different poses. The model, code and data are available for research pur-poses at https://cape.is.tue.mpg.de.

1. Introduction

Existing generative human models [6, 22, 33, 38] suc-cessfully capture the statistics of human shape and pose de-formation, but still miss an important component: clothing. This leads to several problems in various applications. For example, when body models are used to generate synthetic training data [20, 42, 43, 49], the minimal body geome-try results in a significant domain gap between synthetic and real images of humans. Deep learning methods re-construct human shape from images, based on minimally dressed human models [5, 23, 26, 27, 30, 36, 38, 39]. Al-though the body pose matches the image observation, the minimal body geometry does not match clothed humans in

---

most cases. These problems motivate the need for a parametric clothed human model.

Our goal is to create a generative model of clothed human bodies that is low-dimensional, easy to pose, differentiable, can represent different clothing types on different body shapes and poses, and produces geometrically plausible results. To achieve this, we extend SMPL [33] and factorize clothing shape from the undressed body, treating clothing as an additive displacement in the canonical pose (see Fig. 2). The learned clothing layer is compatible with the SMPL body model by design, enabling easy re-posing and animation. The mapping from a given body shape and pose to clothing shape is one-to-many. However, existing regression-based clothing models [16, 52] produce *deterministic* results that fail to capture the stochastic nature of clothing deformations. In contrast, we formulate clothing modeling as a *probabilistic generation task*: for a single pose and body shape, multiple clothing deformations can be sampled. Our model, called CAPE for "Clothed Auto Person Encoding", is conditioned on clothing types and body poses, so that it captures different types of clothing, and can generate *pose-dependent deformations*, which are important for realistically modeling clothing.

We illustrate the key elements of our model in Fig. 1. Given a SMPL body shape, pose and clothing type, CAPE can generate different structures of clothing by sampling a learned latent space. The resulting clothing layer plausibly adapts to different body shapes and poses.

Technical approach. We represent clothing as a displacement layer using a graph that inherits the topology of SMPL. Each node in this graph represents the 3-dimensional offset vector from its corresponding vertex on the underlying body. To learn a generative model for such graphs, we build a graph convolutional neural network (Sec. 4), under the framework of a VAE-GAN [7, 29], using graph convolutions [11] and mesh sampling [41] as the backbone layers. This addresses the problem with existing generative models designed for 3D meshes of human bodies [32, 50] or faces [41] that tend to produce over-smoothed results; such smoothing is problematic for clothing where local details such as wrinkles matter. Specifically, the GAN [15] module in our system encourages visually plausible wrinkles. We model the GAN using a patch-wise discriminator for mesh-like graphs, and show that it effectively improves the quality of the generated fine structures.

Dataset. We introduce a dataset of 4D captured people performing a variety of pose sequences, in different types of clothing (Sec. 5). Our dataset consists of over 80K frames of 8 male and 3 female subjects captured using a 4D scanner. We use this dataset to train our network, resulting in a parametric generative model of the clothing layer.

Versatility. CAPE is designed to be "plug-and-play" for many applications that already use SMPL. Dressing SMPL with CAPE yields 3D meshes of people in clothing, which can be used for several applications such as generating training data, parametrizing body pose in a deep network, having a clothing "prior", or as part of a generative analysis-by-synthesis approach [20, 43, 49]. We demonstrate this on the task of image fitting by extending SMPLify [9] with our model. We show that using CAPE together with SMPLify can improve the quality of reconstructed human bodies in clothing.

In summary, our key contributions are: (1) We propose a probabilistic formulation of clothing modeling. (2) Under this formulation, we learn a conditional Mesh-VAE-GAN that captures both global shape and local detail of a mesh, with controlled conditioning based on human pose and clothing types. (3) The learned model can generate pose-dependent deformations of clothing, and generalizes to a variety of garments. (4) We augment the SMPL 3D human body model with our clothing model, and show an application of the enhanced "clothed-SMPL". (5) We contribute a dataset of 4D scans of clothed humans performing a variety of motion sequences. Our dataset, code, and trained model are available for research purposes at https://cape.is.tue.mpg.de.

2. Related Work

The capture, reconstruction and modeling of clothing has been widely studied. Table 1 shows recent methods categorized into two major classes: (1) reconstruction and capture methods, and (2) parametric models, detailed as follows.

Reconstructing 3D humans. Reconstruction of 3D humans from 2D images and videos is a classical computer vision problem. Most approaches [9, 17, 23, 26, 27, 30, 36, 38, 46] output 3D body meshes from images, but *not clothing*. This ignores image evidence that may be useful. To reconstruct clothed bodies, methods use volumetric [34, 44, 48, 54] or bi-planar depth representations [12] to model the body and garments as a whole. We refer to these as Group 1 in Table 1. While these methods deal with arbitrary clothing topology and preserve a high level of detail, the reconstructed clothed body is not parametric, which means the pose, shape, and clothing of the reconstruction can not be controlled or animated.

Another group of methods are based on SMPL [1, 2, 3, 4, 8, 55]. They represent clothing as an offset layer from the underlying body as proposed in ClothCap [40]. We refer to these approaches as Group 2 in Table 1. These methods can change the pose and shape of the reconstruction using the deformation model of SMPL. This assumes clothing deforms like an undressed human body; i.e. that clothing shape and wrinkles *do not change as a function of pose*. We also use a body-to-cloth offset representation to learn

Table 1: Selection of related methods. Two main 3D clothing method classes exist: (1) Image-based reconstruction and capture methods, and (2) Clothing models that predict deformation as a function of pose. Within each class, methods differ according to the criteria in the columns.

Method Class	Methods	Parametric Model	Pose-dep. Clothing	Full-body Clothing	Clothing Wrinkles	Captured Data*	Code Public	Probabilistic Sampling
Image	Group 1 [†]	No	No	Yes	Yes	Yes	Yes	No
Reconstruction	Group 2 [‡]	Yes	No	Yes	Yes	Yes	Yes	No
Capture	ClothCap [40]	Yes	No	Yes	Yes	Yes	No	No
	DeepWrinkles [28]	Yes	Yes	No	Yes	Yes	No	No
	Yang et al. [52]	Yes	Yes	Yes	No	Yes	No	No
Clothing	Wang et al. [51]	Yes	No	No	No	No	Yes	Yes
Models	DRAPE [16]	Yes	Yes	Yes	Yes	No	No	No
	Sanesteban et al. [45]	Yes	Yes	Yes	Yes	No	No	No
	GarNet [18]	Yes	Yes	Yes	Yes	No	No	No
	Ours	Yes	Yes	Yes	Yes	Yes	Yes	Yes

* As opposed to simulated / synthetic data.
[†] Group 1: BodyNet [48], DeepHuman [54], SiCloPe [34], PIFu [44], MouldingHumans [12].   [‡] Group 2: Octopus [1], MGN [8], Tex2Shape [4].

our model, but critically, we learn a neural function mapping from pose to multi-modal clothing offset deformations. Hence, our work differs from these methods in that we learn a parametric model of how clothing deforms with pose.

Parametric models for 3D bodies and clothes. Statistical 3D human body models learned from 3D body scans, [6, 22, 33, 38] capture body shape and pose and are an important building block for multiple applications. Most of the time, however, people are *dressed* and these models do not represent clothing. In addition, clothes deform as we move, producing changing wrinkles at multiple spatial scales. While clothing models learned from real data exist, few generalize to new poses. For example, Neophytou and Hilton [35] learn a layered garment model on top of SCAPE [6] from dynamic sequences, but generalization to novel poses is not demonstrated. Yang et al. [52] train a neural network to *regress* a PCA-based representation of clothing, but show generalization on the same sequence or on the same subject. Lähner et al. [28] learn a garment-specific pose-deformation model by *regressing* low-frequency PCA components and high frequency normal maps. While the visual quality is good, the model is garment-specific and does not provide a solution for full-body clothing. Similarly, Alldieck et al. [4] use displacement maps with a UV-parametrization to represent surface geometry, but the result is only static. Wang et al. [51] allow manipulation of clothing with sketches in a static pose. The Adam model [22] can be considered clothed but the shape is very smooth and not pose-dependent. Clothing models have been learned from physics simulation of clothing [16, 18, 37, 45], but visual fidelity is limited by the quality of the simulations. Furthermore, the above methods are *regressors* that produce single point estimates. In contrast, our model is *generative*, which allows us to *sample* clothing.

A conceptually different approach infers the parameters of a physical clothing model from 3D scan sequences [47]. This generalizes to novel poses, but the inference problem is difficult and, unlike our model, the resulting physics simulator is not differentiable with respect to the parameters.

Generative models on 3D meshes. Our model predicts clothing displacements on the graph defined by the SMPL mesh using graph convolutions [10]. There is an extensive recent literature on methods and applications of graph convolutions [11, 25, 32, 41, 50]. Most relevant here, Ranjan et al. [41] learn a convolutional autoencoder using graph convolutions [11] with mesh down- and up-sampling layers [13]. Although it works well for faces, the mesh sampling layer makes it difficult to capture the local details, which are key in clothing. In our work, we capture local details by extending the PatchGAN [21] architecture to 3D meshes.

3. Additive Clothed Human Model

To model clothed human bodies, we factorize them into two parts: the minimally-clothed body, and a clothing layer represented as displacements from the body. This enables us to naturally extend SMPL to a class of clothing types by treating clothing as an additional additive shape term. Since SMPL is in wide use, our goal is to extend it in a way that is consistent with current uses, making it effectively a "drop in" replacement for SMPL.

3.1. Dressing SMPL

SMPL [33] is a generative model of human bodies that factors the surface of the body into shape (β) and pose (θ) parameters. As shown in Fig. 2 (a), (b), the architecture of SMPL starts with a triangulated template mesh, $\bar{T}$, in rest pose, defined by $N = 6890$ vertices. Given shape and pose parameters (β, θ), 3D offsets are added to the template, corresponding to shape-dependent deformations ($B_S(\beta)$) and pose dependent deformations ($B_P(\theta)$). The resulting mesh is then *posed* using the skinning function W. Formally:

$$T(\beta, \theta) = \bar{T} + B_S(\beta) + B_P(\theta) \quad (1)$$
$$M(\beta, \theta) = W(T(\beta, \theta), J(\beta), \theta, \mathcal{W}) \quad (2)$$

(a) $\bar{T}$ **(b)** $T(\beta) = \bar{T} + B_S(\beta)$ **(c)** $T(\beta,\theta) = T(\beta) + B_P(\theta)$ **(d)** $T_{\text{clo}} = T(\beta,\theta) + S_{\text{clo}}(z,\theta,c)$ **(e)** $W(T_{\text{clo}}, J(\beta), \theta, \mathcal{W})$
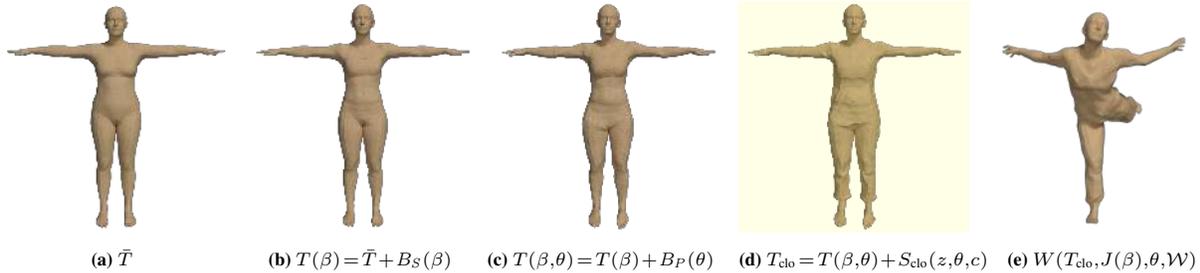
Figure 2: Additive clothed human model. Our contribution is highlighted with yellow background. Following SMPL, our model **(a)** starts from a template mesh, and linearly adds offsets contributed by **(b)** individual body shape β, and **(c)** pose θ; note the deformation on the hips and feet caused by the ballet pose. **(d)** We further add a clothing layer parametrized by pose θ, clothing type c and a clothing shape variable z. **(e)** The vertices are then posed using the skinning equation of SMPL.

where the blend skinning function $W(\cdot)$ rotates the rest pose vertices T around the 3D joints J (computed from β), linearly smoothes them with the blend weights $\mathcal{W}$, and returns the posed vertices M. The pose $\theta \in \mathbb{R}^{3 \times (23+1)}$ is represented by a vector of relative 3D rotations of the 23 joints and the global rotation in axis-angle representation.

SMPL adds *linear deformation layers* to an initial body shape. Following this, we define clothing as an extra offset layer from the body and add it on top of the SMPL mesh, Fig. 2 (d). In this work, we parametrize the clothing layer by the body pose θ, clothing type c and a low-dimensional latent variable z that encodes clothing shape and structure.

Let $S_{\text{clo}}(z,\theta,c)$ be the clothing displacement layer. We extend Eq. (1) to a clothed body template in the rest pose:

$$T_{\text{clo}}(\beta,\theta,c,z) = T(\beta,\theta) + S_{\text{clo}}(z,\theta,c). \qquad (3)$$

Note that the clothing displacements, $S_{\text{clo}}(z,\theta,c)$, are pose-dependent. The final clothed template is then posed with the SMPL skinning function, Eq. (2):

$$M(\beta,\theta,c,z) = W(T_{\text{clo}}(\beta,\theta,c,z), J(\beta), \theta, \mathcal{W}). \qquad (4)$$

This differs from simply applying blend skinning with *fixed* displacements, as done in e.g. [1, 8]. Here, we train the model such that pose-dependent clothing displacements in the template pose are correct once posed by blend skinning.

3.2. Clothing representation

Vertex displacements are not a physical model for clothing and cannot represent all types of garments, but this approach achieves a balance between expressiveness and simplicity, and has been widely used in deformation modeling [16], 3D clothing capture [40] and recent work that reconstructs clothed humans from images [1, 8, 55].

The displacement layer is a graph $\mathcal{G}_d = (\mathcal{V}_d, \mathcal{E}_d)$ that inherits the SMPL topology: the edges $\mathcal{E}_d = \mathcal{E}_{\text{SMPL}}$. $\mathcal{V}_d \in \mathbb{R}^{3 \times N}$ is the set of vertices, and the feature on each vertex is the 3-dimensional offset vector, (d_x, d_y, d_z), from its corresponding vertex on the underlying body mesh.

We train our model on 3D scans of people in clothing. From data pairs $(\mathcal{V}_{\text{clothed}}, \mathcal{V}_{\text{minimal}})$ we compute displace-

ments, where $\mathcal{V}_{\text{clothed}}$ stands for the vertices of a clothed human mesh, and $\mathcal{V}_{\text{minimal}}$ the vertices of a minimally-clothed mesh. Therefore, we first scan subjects in both clothed and minimally-clothed conditions, then use the SMPL model with free deformation [1, 53] to register the scans. As a result, we obtain SMPL meshes capturing the geometry of the scans, the corresponding pose parameters, and vertices of the *unposed meshes*[1]. For each $(\mathcal{V}_{\text{clothed}}, \mathcal{V}_{\text{minimal}})$ pair, the displacements are then calculated as $\mathcal{V}_d = \mathcal{V}_{\text{clothed}} - \mathcal{V}_{\text{minimal}}$, where the subtraction is performed per-vertex along the feature dimension. Ideally, $\mathcal{V}_d$ has non-zero values only on body parts covered with clothes.

In summary, CAPE extends the SMPL body model to clothed bodies. Compared to volumetric representations of clothed people [34, 44, 48, 54], our combination of the body model and the garment layer is superior in the ease of reposing and garment retargeting: the former uses the same blend skinning as the body model, while the latter is a simple addition of the displacements to a minimally-clothed body shape. In contrast to similar models that also dress SMPL with offsets [1, 8], our garment layer is parametrized, low-dimensional, and pose-dependent.

4. CAPE

Our clothing term $S_{\text{clo}}(z,\theta,c)$ in Eq. (3) is a function of z, a code in a learned low-dimensional latent space that encodes the shape and structure of clothing, body pose θ, and clothing type c. The function outputs the clothing displacement graph $\mathcal{G}_d$ as described in Sec. 3.2. We parametrize this function using a graph convolutional neural network (Graph-CNN) as a VAE-GAN framework [15, 24, 29].

4.1. Network architecture

As shown in Fig. 3, our model consists of a generator G with an encoder-decoder architecture and a discriminator D. We also use auxiliary networks C_1, C_2 to handle the conditioning. The network is differentiable and is trained end-to-end.

---

[1] We follow SMPL and use the T-pose as the zero-pose. For the mathematical details of registration and unposing, we refer the reader to [53].

(a)   **(b)**
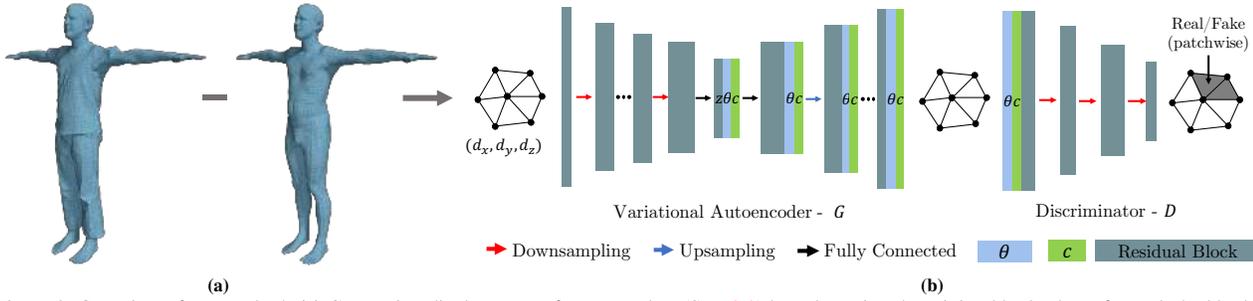
Figure 3: Overview of our method. **(a)** Computing displacements from scan data (Sec. 3.2) by subtracting the minimal body shape from clothed body mesh in the unposed space. **(b)** Schematic of our network architecture.

For simplicity, we use the following notation in this section. x: the vertices $\mathcal{V}_d$ of the input displacement graph; $\hat{x}$: vertices of the reconstructed graph; θ and c: the pose and clothing type condition vector; z: the latent code.

Graph generator. We build the graph generator following a VAE-GAN framework. During training, an encoder $Enc(\cdot)$ takes in the displacement x, extract its features through multiple graph convolutional layers, and maps it to the low-dimensional latent code z. A decoder is trained to reconstruct the input graph $\hat{x} = Dec(z)$ from z. Both the encoder and decoder are feed-forward neural networks built with mesh convolutional layers. Linear layers are used at the end of the encoder and the beginning of the decoder. The architecture is shown in the supplemental materials.

Stacking graph convolution layers causes a loss of local features [31] in the deeper layers. This is undesirable for clothing generation because fine details, corresponding to wrinkles, are likely to disappear. Therefore, we improve the standard graph convolution layers with residual connections, which enable the use of low-level features from the layer input if necessary.

At test time, the encoder is not needed. Instead, z is sampled from the Gaussian prior distribution, and the decoder serves as the graph generator: $G(z) = Dec(z)$. We detail different use cases below.

Patchwise discriminator. To further enhance fine details in the reconstructions, we introduce a patchwise discriminator D for graphs, which has shown success in the image domain [21, 56]. Instead of looking at the entire generated graph, the discriminator only classifies whether a graph patch is real or fake based on its local structure. Intuitively this encourages the discriminator to only focus on fine details, and the global shape is taken care of by the reconstruction loss. We implement the graph patchwise-discriminator using four graph convolution-downsampling blocks [41]. We add a discriminative real / fake loss for each of the output vertices. This enables the discriminator to capture a patch of neighboring nodes in the reconstructed graph and classify them as real / fake (see Fig. 3).

Conditional model. We condition the network with body pose θ and clothing type c. The SMPL pose parameters are in axis-angle representation, and are difficult for the neural network to learn [27, 30]. Therefore, following previous work [27, 30], we transform the pose parameters into rotational matrices using the Rodrigues equation. The clothing types are discrete by nature, and we represent them using one-hot labels. Both conditions are first passed through a small fully-connected embedding network, $C_1(\theta), C_2(c)$, respectively, so as to balance the dimensionality of learned graph features and of the condition features. We also experiment with different ways of conditioning the mesh generator: concatenation in the latent space; appending the condition features to the graph features at all nodes in the generator; and the combination of the two. We find that the combined strategy works better in terms of network capability and the effect of conditioning.

4.2. Losses and learning

For reconstruction, we use an L1 loss over the vertices of the mesh x, because it encourages less smoothing compared to L2, given by

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{x \sim p(x), z \sim q(z|x)} \left[\|G(z, \theta, c) - x\|_1 \right]. \quad (5)$$

Furthermore, we apply a loss on the mesh edges to encourage the generation of wrinkles instead of smooth surfaces. Let e be an edge in the set of edges, $\mathcal{E}$, of the ground truth graph, and $\hat{e}$ the corresponding edge in the generated graph. We penalize the mismatch of all corresponding edges by

$$\mathcal{L}_{\text{edge}} = \mathbb{E}_{e \in \mathcal{E}, \ \hat{e} \in \hat{\mathcal{E}}} \left[\|e - \hat{e}\|_2 \right]. \quad (6)$$

We also apply a KL divergence loss between the distribution of latent codes and the Gaussian prior

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_{x \sim p(x)} \left[\text{KL}(q(z|x) \ \| \ \mathcal{N}(0, \boldsymbol{I})) \right]. \quad (7)$$

Moreover, the generator and the discriminator are trained using an adversarial loss

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{x \sim p(x)} \left[\log(D(x, \theta, c)) \right] + \\ \mathbb{E}_{z \sim q(z|x)} \left[\log(1 - D(G(z, \theta, c))) \right], \quad (8)$$

where G tries to minimize this loss against the D that aims to maximize it.

The overall objective is a weighted sum of these loss terms given by

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \gamma_{\text{edge}}\mathcal{L}_{\text{edge}} + \gamma_{\text{kl}}\mathcal{L}_{\text{KL}} + \gamma_{\text{gan}}\mathcal{L}_{\text{GAN}}. \quad (9)$$

Training details are provided in the supplemental materials.

5. CAPE Dataset

We build a dataset of 3D clothing by capturing temporal sequences of 3D human body scans with a high-resolution body scanner (3dMD LLC, Atlanta, GA). Approximately 80K 3D scan frames are captured at 60 FPS, and a mesh with SMPL model topology is registered to each scan to obtain surface correspondences. We also scanned the subjects in a minimally-clothed condition to obtain an accurate estimate of their body shape under clothing. We extract the clothing as displacements from the minimally-clothed body as described in Sec. 3.2. Noisy frames and failed registrations are removed through manual inspection.

The dataset consists of 8 male subjects and 3 female subjects, performing a wide range of motions. The subjects gave informed written consent to participate and to release the data for research purposes. "Clothing type" in this work refers to the 4 types of full-body outfits, namely *shortlong*: short-sleeve upper body clothing and long lower body clothing; and similarly *shortshort, longshort, longlong*. These outfits comprise 8 types of common garments. We refer to the supplementary material for the list of garments, further details, and examples from the dataset.

Compared to existing datasets of 3D clothed humans, our dataset provides captured data and *alignments* of SMPL to the scans, separates the clothing from body, and provides accurate, *captured* ground truth body shape under clothing. For each subject and outfit, our dataset contains large pose variations, which induces a wide variety of wrinkle patterns. Since our dataset of 3D meshes has a consistent topology, it can be used for the quantitative evaluation of different Graph-CNN architectures. The dataset is available for research purposes at https://cape.is.tue.mpg.de.

6. Experiments

We first show the representation capability of our model and then demonstrate the model's ability to generate new examples by probabilistic sampling. We then show an application to human pose and shape estimation.

6.1. Representation power

3D mesh auto-encoding errors. We use the reconstruction accuracy to measure the capability of our VAE-based model for geometry encoding and preserving. We compare

with a recent convolutional mesh autoencoder, CoMA [41], and a linear (PCA) model. We compare to both the original CoMA with a $4\times$ downsampling (denoted as "CoMA-4"), and without downsampling (denoted "CoMA-1") to study the effect of downsampling on over-smoothing. We use the same latent space dimension $n_z = 18$ (number of principal components in the case of PCA) and hyper-parameter settings, where applicable, for all models.

Table 2 shows the per-vertex Euclidean error when using our network to reconstruct the clothing displacement graphs from a held-out test set in our CAPE dataset. The model is trained and evaluated on male and female data separately. Body parts such as head, fingers, toes, hands and feet are excluded from the accuracy computation, as they are not covered with clothing.

Our model outperforms the baselines in the auto-encoding task; additionally, the reconstructed shape from our model is *probabilistic* and *pose-dependent*. Note that, CoMA here is a deterministic auto-encoder with a focus on reconstruction. Although the reconstruction performance of PCA is on par with our method on male data, PCA can not be used directly in the inference phase with a pose parameter as input. Furthermore, PCA assumes a Gaussian distribution of the data, which does not hold for complex clothing deformations. Our method addresses both of these issues.

Fig. 4 shows a qualitative comparison of the methods. PCA keeps wrinkles and boundaries, but the rising hem on the left side disappears. CoMA-1 and CoMA-4 are able to capture global correlation, but the wrinkles tend to be smoothed. By incorporating all the key components, our model manages to model both local structures and global correlations more accurately than the other methods.

Ablation study. We remove key components from our model while keeping all the others, and evaluate the model performance; see Table 2. We observe that the discriminator, residual block and edge loss all play important roles in the model performance. Comparing the performance of CoMA-4 and CoMA-1, we find that the mesh the downsampling layer causes a loss of fidelity. However, even without any spatial downsampling, CoMA-1 still underperforms our model. This shows the benefits of adding the discriminator, residual block, and edge loss in our model.

Table 2: Per-vertex auto-encoding error in millimeters. Upper section: comparison with baselines; lower section: ablation study.

	Male		Female	
Methods	Error mean	median	Error mean	median
PCA	$5.65 \pm \mathbf{4.81}$	4.30	4.82 ± 3.82	3.78
CoMA-1	6.23 ± 5.45	4.66	4.69 ± 3.85	3.61
CoMA-4	6.87 ± 5.62	5.29	4.86 ± 3.96	3.75
Ours	$\mathbf{5.54} \pm 5.09$	**4.03**	$\mathbf{4.21} \pm \mathbf{3.76}$	**3.08**
Ablated Components	Error mean	median	Error mean	median
Discriminator	5.65 ± 5.18	4.11	4.31 ± 3.78	3.18
Res-block	5.60 ± 5.21	4.05	4.27 ± 3.76	3.15
Edge loss	5.93 ± 5.40	4.32	4.32 ± 3.78	3.19

| Ground Truth | PCA | COMA-4 | COMA-1 | w/o Discr. | w/o edge loss | CAPE full model |
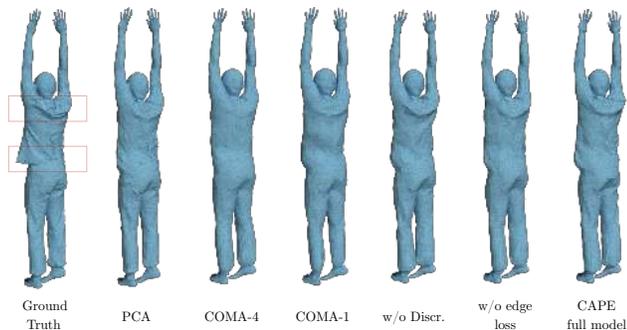
Figure 4: Example of reconstruction by the baselines, ablated version of our model, and our full model. Pay attention to the hem and wrinkles on upper back. Our model is able to recover both long-range correlations and local details. Best viewed zoomed-in on screen.

6.2. Conditional generation of clothing

As a generative model, CAPE can be sampled and generates new data. The model has three parameters: z, c, θ (see Eq. (3)). By sampling one of them while keeping the other two fixed, we show how the conditioning affects the generated clothing shape.

Sampling. Fig. 5 presents the sampled clothing dressed on unseen bodies, in a variety of poses that are not used in training. For each subject, we fix the pose θ and clothing type c, and sample z several times to generate varied clothing shapes. The sampling trick in [14] is used. Here we only show untextured rendering to highlight the variation in the generated *geometry*. As CAPE inherits the SMPL topology, the generated clothed body meshes are compatible with all existing SMPL texture maps. See supplemental materials for a comparison between a CAPE sample and a SMPL sample rendered with the same texture.

As shown in the figure, our model manages to capture long-range correlations within a mesh, such as the elevated hem for a subject with raised arms, and the lateral wrinkle on the back with raised arms. The model also synthesizes local details such as wrinkles in the armpit area, and boundaries at cuffs and collars.

Pose-dependent clothing deformation. Another practical use case of CAPE is to animate an existing clothed body. This corresponds to fixing the clothing shape variable z and clothing type c, and reposing the body by changing θ. The challenge here is to have a clothing shape that is consistent across poses, yet deforms plausibly. We demonstrate the pose-dependent effect on a test pose in Fig. 6. The difference of the clothing layer between the two poses is calculated in the canonical pose, and shown with color coding. The result shows that the clothing type is consistent while local deformation changes along with pose. We refer to the supplemental video for a comparison with traditional rig-and-skinning methods that use fixed clothing offsets.

User study of generated examples. To test the realism of the generated results from our method, we performed a user study on Amazon Mechanical Turk (AMT). We dress virtual avatars in 3D and render them into front-view images. Following the protocol from [21], raters are presented with a series of "real vs fake" trials. On each trial, the rater is presented with a "real" mesh render (randomly picked from our dataset), and a "fake" render (generated by our model), shown side-by-side. The raters are asked to pick the one that they think is real. Each pair of renderings is evaluated by 10 raters. More strictly than [21], we present both real and fake renderings *simultaneously*, do not set a time limit for raters and allow zoom-in for detailed comparison. In this setting, the best score that a method can obtain is 50%, meaning that the real and fake examples are indistinguishable.

We evaluate with two test cases. In test case 1, we fix the clothing type to be "shortlong" (the most common clothing type in training), and generate 300 clothed body meshes with various poses for the evaluation. In test case 2, we fix the pose to be an A-pose (the most frequent pose in training), and sample 100 examples per clothing type for evaluation. On average, in the direct comparison with real data, our synthesized data "fools" participants $35.1\% \pm 15.7\%$ and $38.7\% \pm 16.5\%$ of the time respectively (i.e. these participants labeled our result as "real").

6.3. Image fitting

CAPE is fully differentiable with respect to the clothing shape variable z, body pose θ and clothing type c. Therefore, it can also be used in optimization frameworks. We show an application of CAPE on the task of reconstructing a body mesh from a single image, by entending the optimization-based method, SMPLify [9]. Assuming c is known, we dress the minimally-clothed output mesh from SMPLify using CAPE, project it back to the image using a differentiable renderer [19] and optimize for β, θ, z, with respect to the silhouette discrepancy.

We evaluate our image fitting pipeline on renderings of 120 randomly selected unseen test examples from the CAPE dataset. To compare, we measure the reconstruction error of SMPLify and our results against ground truth meshes using mean square vertex error (MSE). To eliminate the error introduced by the ambiguity of human scale and distance to the camera, we optimize the global scaling and translation of predictions for both methods on each test sample. A mask is applied to exclude error in the non-clothed regions such as head, hands and feet. We report the errors of both methods in Table 3. Our model performs 18% better than SMPLify due to its ability to capture clothing shape.

More details about the objective function, experimental setup and qualitative results of the image fitting experiment

Table 3: Vertex MSE of image fitting results measured in meters.

Method	SMPLify [9]	Ours
Per-vertex MSE	0.0223	0.0189
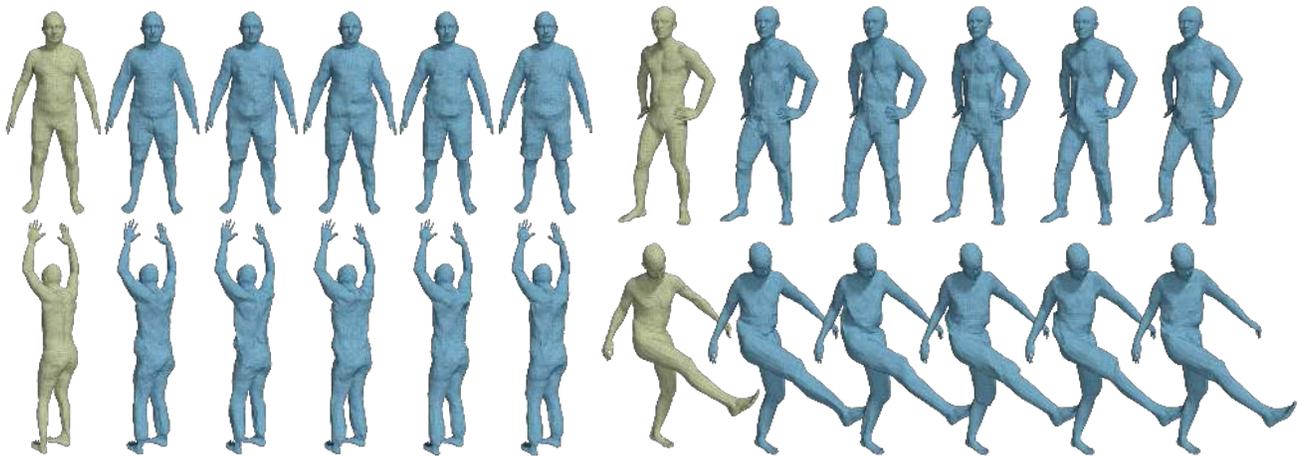
Figure 5: Clothing sampled and generated from our CAPE model applied to four unseen body shapes (green colored) with different poses. Each body in green is followed by five blue examples (generated by sampling the latent vector z) showing different versions of the same outfit type. The four groups are wearing outfit types "longshort", "longlong", "shortlong" and "shortshort", respectively. Best viewed zoomed-in on screen.
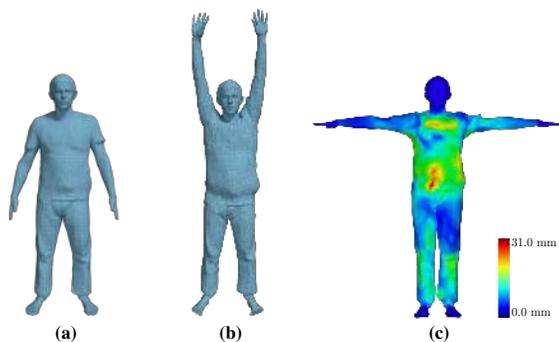
Figure 6: Pose-dependent clothing shape. **(a)** and **(b)**: two clothing shapes generated from CAPE, with different pose parameters. **(c)**: color-coded difference of the offset clothing layers in **(a)** and **(b)**, in the canonical pose.

are provided in supplementary materials.

Furthermore, once a clothed human is reconstructed from the image, our model can repose and *animate* it, as well as change the subject's clothes by re-sampling z or clothing type c. This shows the potential for several applications. We show examples in the supplemental video.

7. Conclusions, Limitations, Future Work

We have introduced a novel graph-CNN-based generative shape model that enables us to condition, sample, and preserve fine shape detail in 3D meshes. We use this to model clothing deformations from a 3D body mesh and condition the latent space on body pose and clothing type. The training data represents 3D displacements from the SMPL body model for varied clothing and poses. This design means that our generative model is compatible with SMPL in that clothing is an additional *additive* term applied to the SMPL template mesh. This makes it possible to sample clothing, dress SMPL with it, and then animate the body with pose-dependent clothing wrinkles. A

clothed version of SMPL has wide applicability in computer vision. As shown, we can apply it to fitting the body to images of clothed humans. Another application would use the model to generate training data of 3D clothed people to train regression-based pose-estimation methods.

There are a few limitations of our approach that point to future work. First, CAPE inherits the limitation of the offset representation for clothing: (1) Garments such as skirts and open jackets differ from the body topology and cannot be trivially represented by offsets. Consequently, when fitting CAPE to images containing such garments, it could fail to explain the image evidence; see discussions on the skirt example in the supplementary material. (2) Mittens and shoes: they can technically be modeled by the offsets, but their geometry is sufficiently different from fingers and toes, making this impractical. A multi-layer model can potentially overcome these limitations. Second, the level of geometric details that CAPE can achieve is upper-bounded by the mesh resolution of SMPL. To produce finer wrinkles, one can resort to higher resolution meshes or bump maps. Third, while our generated clothing depends on pose, it does not depend on *dynamics*. This does not cause severe problem for most slow motions but does not generalize to faster motions. Future work will address modeling clothing deformation on temporal sequence and dynamics.

References

[1] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 4

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *International Conference on 3D Vision (3DV)*, 2018. 2

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[4] 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)*, 2019. 2, 3

[5] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[6] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 408–416. ACM, 2005. 1, 3

[7] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. CVAE-GAN: fine-grained image generation through asymmetric training. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[8] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-Garment Net: Learning to Dress 3D People from Images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 4

[9] 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 *European Conference on Computer Vision (ECCV)*. Springer, 2016. 2, 7

[10] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann Lecun. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations (ICLR)*, 2014. 3

[11] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, pages 3844–3852, 2016. 2, 3

[12] Valentin Gabeur, Jean-Sébastien Franco, Xavier MARTIN, Cordelia Schmid, and Gregory Rogez. Moulding Humans: Non-parametric 3D Human Shape Estimation from Single Images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[13] Michael Garland and Paul S Heckbert. Surface simplification using quadric error metrics. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 209–216. ACM Press/Addison-Wesley Publishing Co., 1997. 3

[14] P. Ghosh, M. S. M. Sajjadi, A. Vergari, M. J. Black, and B. Schölkopf. From variational to deterministic autoencoders. In *International Conference on Learning Representations (ICLR)*, 2020. 7

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2, 4

[16] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. DRAPE: DRessing Any PErson. *ACM Transactions on Graphics (TOG)*, 31(4):35–1, 2012. 2, 3, 4

[17] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *The IEEE International Conference on Computer Vision (ICCV)*, 2009. 2

[18] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. GarNet: A two-stream network for fast and accurate 3D cloth draping. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[19] Paul Henderson and Vittorio Ferrari. Learning single-image 3D reconstruction by generative modelling of shape, pose and shading. *International Journal of Computer Vision (IJCV)*, 2019. 7

[20] David T. Hoffmann, Dimitrios Tzionas, Michael J. Black, and Siyu Tang. Learning to train with synthetic humans. In *German Conference on Pattern Recognition (GCPR)*, 2019. 1, 2

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 5, 7

[22] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total Capture: A 3D deformation model for tracking faces, hands, and bodies. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3

[23] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014. 4

[25] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017. 3

[26] 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 *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[27] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 5

[28] Zorah Lähner, Daniel Cremers, and Tony Tung. DeepWrinkles: Accurate and realistic clothing modeling. In *European Conference on Computer Vision*. Springer, 2018. 3

[29] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International Conference on Machine Learning (ICML)*, 2016. 2, 4

[30] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 5

[31] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 5

[32] Or Litany, Alex Bronstein, Michael Bronstein, and Ameesh Makadia. Deformable shape completion with graph convolutional autoencoders. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015. 1, 2, 3

[34] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. SiCloPe: Silhouette-based clothed people. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3, 4

[35] Alexandros Neophytou and Adrian Hilton. A layered model of human body and garment deformation. In *International Conference on 3D Vision (3DV)*, 2014. 3

[36] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *International Conference on 3D Vision (3DV)*, 2018. 1, 2

[37] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. The Virtual Tailor: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[38] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3

[39] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[40] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4):73, 2017. 2, 3, 4

[41] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3D faces using convolutional mesh autoencoders. In *The European Conference on Computer Vision (ECCV)*, 2018. 2, 3, 5, 6

[42] Anurag Ranjan, David T Hoffmann, Dimitrios Tzionas, Siyu Tang, Javier Romero, and Michael J. Black. Learning multi-human optical flow. *International Journal of Computer Vision (IJCV)*, Jan. 2020. 1

[43] Anurag Ranjan, Javier Romero, and Michael J Black. Learning human optical flow. In *British Machine Vision Conference (BMVC)*, 2018. 1, 2

[44] 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)*, 2019. 2, 3, 4

[45] Igor Santesteban, Miguel A. Otaduy, and Dan Casas. Learning-Based Animation of Clothing for Virtual Try-On. *Computer Graphics Forum (Proc. Eurographics)*, 2019. 3

[46] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. FACSIMILE: Fast and Accurate Scans From an Image in Less Than a Second. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[47] Carste Stoll, Jürgen Gall, Edilson de Aguiar, Sebastian Thrun, and Christian Theobalt. Video-based reconstruction of animatable human characters. In *ACM SIGGRAPH ASIA*, 2010. 3

[48] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3D human body shapes. In *The European Conference on Computer Vision (ECCV)*. Springer, 2018. 2, 3, 4

[49] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[50] Nitika Verma, Edmond Boyer, and Jakob Verbeek. Feastnet: Feature-steered graph convolutions for 3D shape analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[51] Tuanfeng Y Wang, Duygu Ceylan, Jovan Popović, and Niloy J Mitra. Learning a shared shape space for multimodal garment design. In *ACM SIGGRAPH ASIA*, 2018. 3

[52] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Analyzing clothing layer deformation statistics of 3D human motions. In *The European Conference on Computer Vision (ECCV)*. Springer, 2018. 2, 3

[53] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4

[54] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. DeepHuman: 3D human reconstruction from a single image. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 4

[55] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 4

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5