CLOTH3D: Clothed 3D Humans

Hugo Bertiche[1], Meysam Madadi[2] and Sergio Escalera[1,2]

[1] Dept. Mathematics and Informatics, Universitat de Barcelona, Catalonia, Spain

[2] Computer Vision Center, Edifici O, Campus UAB, 08193 Bellaterra (Barcelona), Catalonia Spain

Figure 1: Left: CLOTH3D is the first big scale dataset of animated clothed humans. It contains thousands of different outfits and subjects, high variability of poses and rich cloth dynamics. Right: generated 3D garments with proposed GCVAE.

Abstract

This work presents CLOTH3D, the first big scale synthetic dataset of 3D clothed human sequences. CLOTH3D contains a large variability on garment type, topology, shape, size, tightness and fabric. Clothes are simulated on top of thousands of different pose sequences and body shapes, generating realistic cloth dynamics. We provide the dataset with a generative model for cloth generation. We propose a Conditional Variational Auto-Encoder (CVAE) based on graph convolutions (GCVAE) to learn garment latent spaces. This allows for realistic generation of 3D garments on top of SMPL model for any pose and shape.

1. Introduction

The modelling, recovery and generation of 3D clothes will allow for enhanced virtual try-ons experience, reducing designers and animators workload, or understanding of physics simulations through deep learning, just to mention a few. However, current literature in the modelling, recovery and generation of clothes is almost focused on 2D data [7, 12, 20, 22]. This is because of two factors. First, deep learning approaches are data-hungry, and nowadays not enough 3D cloth data is available (see 3DPW, BUFF, and Untitled dataset features in Tab. 1). Second, garments present a huge variability in terms of shape, sizes, topologies, fabrics, or textures, among others, increasing the complexity of representative 3D garment generation.

One could define three main strategies in order to pro-

duce data of 3D dressed humans: 3D scans, 3D-from-RGB, and synthetic generation. In the case of 3D scans, they are costly, and at most they can produce a single mesh (human+garments). Alternatively, datasets that infer 3D geometry of clothes from RGB images are inaccurate and cannot properly model cloth dynamics. Finally, synthetic data is easy to generate and is ground truth error free. Synthetic data has proved to be helpful to train deep learning models to be used in real applications [18, 24, 21].

In this work, we present CLOTH3D, the first synthetic dataset composed of thousands of sequences of humans dressed with high resolution 3D clothes, see Fig.1. CLOTH3D is unique in terms of garment, shape, and pose variability, including more than 2 million 3D samples. We developed a generation pipeline that creates a unique outfit for each sequence in terms of garment type, topology, shape, size, tightness and fabric. While other datasets contain just a few different garments, ours has thousands of different ones. On Tab. 1 we summarize features of existing datasets and CLOTH3D.

Additionally, we provide with a baseline model able to generate dressed human models. Similar to [2, 16, 28] we encode garments as offsets connecting skin to cloth, using SMPL[14] as human body model. This yields an homogeneous dimensionality on the data. As in [19], we use a segmentation mask to extract the garment by removing body vertices. In our case, the mask is predicted by the network. We propose a Conditional Variational Auto-Encoder

Dataset	3DPW[25]	BUFF[31]	Untitled[26]	CLOTH3D
Res.	2.5cm	0.4cm	1cm	1cm
Missing	x	✓	x	x
Dyn.	x	✓	x	✓
Garments	18[1]	10∼20	3[2]	∼9.6K
Fabrics	x	x	x	✓
Poses[3]	Low	Low	Very low	High
Subjects	18[1]	6	2K	7.2K
Layered	x	x	✓	✓
#frames	51k	11K	24K	2.1M
Type	Real	Real	Synth.	Synth.
RGB	✓	x	✓	x
GT error	26mm	1.5-3mm	None	None

Table 1: CLOTH3D vs. available 3D cloth datasets. [1]: 3DPW contains 18 clothed models that can be shaped as SMPL. [2]: garments of [26] are shaped to different sizes. [3]: poses are strongly related to number of frames, and in [26] most samples share the same static pose.

(CVAE) based on graph convolutions [5, 6, 16, 17, 27, 30] (GCVAE) to learn garment latent spaces. This later allows for the generation of 3D garments on top of SMPL model for any pose and shape (right on Fig.1).

2. Related work

3D garment datasets. Current literature on 3D garment lacks on large public available datasets. One strategy to capture 3D data is through **3D scans**. The BUFF dataset [31] provides high resolution 3D scans, but few number of subjects, poses and garments. Furthermore, scanning techniques cannot provide layered models (one mesh for the body and one for each garment) and often one can find regions occluded at scanning time, meaning missing vertices or corrupted shapes. The work of [19] proposed a methodology to segment scans to obtain layered models. Authors of [29] combined 3D scans with cloth simulation fitting at each frame to deal with missing vertices. Similarly, [4] provided a dataset from 3D scans. However, the amount of samples is in the order of a few hundreds. The 3DPW dataset [25] is not focused on garments, but rather on pose and shape in-the-wild. The authors proposed a modified SMPL **parameterized model** for each outfit (18 clothed models), which, as SMPL, can be shaped and posed. Nevertheless, resolution is low and posing is through rigid rotations. Therefore, cloth dynamics are not represented. Finally, the dataset of [26] is synthetically created through **physics simulation**, with three different garment types: tshirt, skirt and kimono. They propose an automatic garment resizing based on real patterns, but provide only static samples on few different poses. Our CLOTH3D dataset aims to overcome previous datasets issues. We automatically generate garments to obtain a huge variability on garment type, topology, shape, size, tightness and fabric. Afterwards, we simulate clothes on top of thousands of different pose sequences and body shapes. Tab.1 shows a comparison of features for existing datasets and ours. In CLOTH3D we focus on sample vari-

ability (garments, poses, shapes), containing realistic cloth dynamics. 3DPW sequences are based on rotation on rigged models, dataset of [26] contains static poses only, and BUFF has very few and short sequences. Moreover, none other provides metadata about fabrics, which has a strong influence on cloth behaviour. Similarly, the scarcity of these datasets implies low variability on garments, poses and subjects. Finally, note how only synthetic datasets provide with layered models and have no annotation error.

3D garment generation. Current works in 3D clothing focus on the generation of dressed humans. We split related work into non-deep and deep-learning approaches. Regarding **non-deep learning**, the authors of [9] proposed a data-driven model that learns deformations from template garment to garment fitted to the human body, shaped and posed. They factorize deformations into shape-dependant and pose-dependant by training on rest pose data first, and later on posed bodies. Transformations are learnt per triangle, and thus it yields inconsistent meshes that need to be reconstructed. The data-driven model of [19] is able to recover and retarget garments from 4D scan sequences relying on masks to separate body and cloth. Authors propose an energy optimization process to indetify underlying body shape and garment geometry, later, cloth displacements w.r.t. body are computed and applied to new body shapes. This means information such as wrinkles is "copied" to new bodies, which produces valid samples but cannot properly generate its variability. Regarding **deep learning** strategies, the work of [10] deals with body and garments as different point clouds through different streams of a network, which are later fused. They also use skin-cloth correspondences for computing local-features and losses through nearest neighbour. The works of [2, 16, 28] consider encoding clothes as offsets from SMPL body model with different goals. On [16] authors propose a combination of graph VAE and GAN to model SMPL offsets into clothing. In [26, 28] a PCA decomposition is used to reduce clothing space. Similar to previous approaches, our proposed methodology also encodes clothes as SMPL offsets. Nevertheless, the assumption that garments follow body topology does not hold for skirts and dresses. In this sense, we propose a novel body topology specific for those cases. Additionally, our model predicts garment mask along offsets to generate layered models.

3. Dataset

CLOTH3D is the first big scale dataset of 3D clothed humans. The dataset is composed of 3D sequences of animated human bodies wearing different garments. Fig. 1 depicts a sequence (first row) and randomly sampled frames from different sequences. Samples are layered, meaning each garment and body are represented by different 3D meshes. Garments are automatically generated for each se-
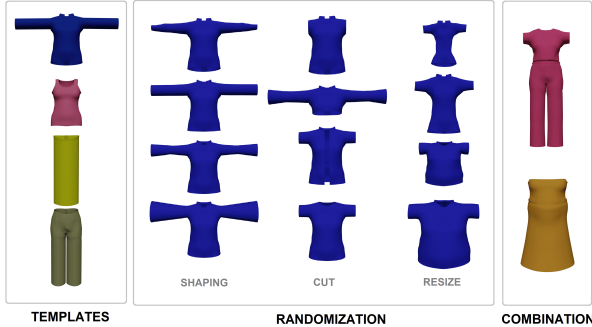
Figure 2: Unique outfit generation pipeline. First, one upper-body and lower-body garment template is selected. Then, garments are individually shaped, cut and resized. Finally, garments might be combined into a single one.

quence with randomized shape, tightness, topology and fabric, and resized to target human shape. This process yields a unique outfit for each sequence. It contains over 7000 non-overlapping sequences of 300 frames each at 30fps, yielding a total of 2.1M samples. As seen in Tab. 1, garment and pose variability is scarce in available datasets, and CLOTH3D aims to fill that gap. To ensure garment type balance, given that females present higher garment variability, we balance gender as 2:1 (female:male). Finally, for validation purposes, we split the data in 80% sequences as training and 20% as test. Splitting by sequences ensures no garment, shape or pose is repeated in training and test.

The data generation pipeline starts with sequences of human bodies in 3D. Human pose data source is [1], later transformed to volumetric bodies through SMPL [14]. This sequences might present body self-collisions which will hinder cloth simulation, not only on affected regions, but also in global garment dynamics. We automatically solve collisions or reject these samples. Human generation process is described in subsec. 3.1. Later, we generate unique outfits for each sequence. We start from a few template meshes which are randomly shaped, cut and resized to generate a unique pair of garments for each sample, with the possibility to be combined into a single full-body garment. Fig. 2 shows the generation process, which is also detailed in subsec. 3.2. Finally, once human sequence and outfit are done, we use a physics based simulation to obtain the garment 3D sequences. Simulation details are described in subsec. 3.3.

Generation algorithm. Brief summary of the steps performed by our generator:

1. HUMAN
   1.1. Pick SMPL parameters
   1.2. Compute SMPL body sequence
   1.3. Solve self-collisions

2. OUTFIT
   2.1. Pick template garments
   2.2. Shape sleeves/legs/skirt
   2.3. Cut
   2.4. Resize
   2.5. Sew garments into jumpsuit/dress (optional)

3. SIMULATION
   3.1. Fabric settings
   3.2. Body shape transition (from $\beta + \gamma$ to β)
   3.3. Pose transition
   3.4. Simulate sequence

3.1. Human 3D sequences

SMPL. It is a parametric human body model which takes as input shape $\beta \in \mathbb{R}^{10}$ and pose $\theta \in \mathbb{R}^{24 \times 3}$ to generate the corresponding mesh with 6890 vertices. We use this model to generate animated human 3D sequences. We refer to [15] for SMPL details. To generate animated bodies, we need a source of valid sequences of SMPL pose parameters $\theta \in \mathbb{R}^{f \times 24 \times 3}$. We take such data from the work of [24], where pose is inferred from CMU MoCap data [1] following the methodology proposed at [13]. These pose data come from around 2600 sequences of 23 different actions (dancing, playing, running, walking, jumping, climbing, etc.) performed by over 100 different subjects. SMPL shape deformations are linearly modeled through PCA. To obtain a balanced dataset we uniformly sample shape within range $[-3, 3]$ for each sequence.

Self-collision. Body collides with itself for certain combinations of pose and shape parameters. Intersection volumes create regions where simulated repel forces are inconsistent, corrupting global cloth dynamics. We classify these collisions in three generic cases. Solvable Fig.3a: small intersection volumes near joints, specially armpits and crotch. Through visual inspection, we identified these problematic body regions on which to detect and solve collisions. Using SMPL segmentation, we separate vertices belonging to different body parts. Collisions appear as intersection of pairs of segments. For each of these pairs, we test edges of a segment vs faces of the other, and vice-versa. Since we identified problematic regions, the number of edge-vs-face tests is significantly reduced. This yields a set of intersection points to which we approximate a plane. Then, each collided body vertex is moved to the corresponding side of this plane based on segment index. Separation space is 4mm so that a folded cloth can fit. Unsolvable Fig.3b: big intersection volumes or incompatible intersections (e.g.: arm vs. leg). We reject or re-simulate with thinner body. Special cases Fig.3c: removing hands, forearms or arms for short-sleeved upper-body and lower-body garments significantly increases the amount of valid samples. This requires manual supervision. Self-collision solution is not stored, hence, if collided vertices change significantly, garments might present interpenetration w.r.t. unsolved body. Only
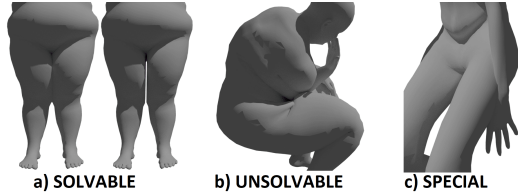
Figure 3: Types of self-collision: a) collided vertices can be linearly separated with the aid of a body part segmentation, b) no trivial solution, we reject this kind of sample, c) correct simulation might be possible if forearm is removed.

small intersected volumes are corrected and the rest are rejected (or simulated with thinner body). The goal of self-collision solving is to avoid invalid cloth dynamics. Accurate, realistic solving of soft-body self-collision is out of the scope of this work.

3.2. Garment generation

Garment templates. Generation starts with a few template garments for each gender. Garments can be classified in upper-body and lower-body. Lower-body can be further split into trousers and skirts. These three categories, and combinations between them, encompass almost any day-to-day garment. Template garments have been manually created by designers from real patterns and are: t-shirt, top, trousers and skirt.

Shaping. On sleeves, legs and skirt, we find a significant shape variability. It is possible to define them as cylinders of variable width around certain axes: along arms for sleeves, legs for trousers and vertical body axis for skirt. For sleeves and legs, width will be constant or decreasing while moving towards wrist/ankle, and beyond a randomly sample point along its axis, it might start increasing (widening). For skirts, width always increases, from waist to bottom. Rate of width decrease/increase is uniformly sampled within ranges empirically set per garment.

Formally:

$$W(x) = \alpha_1 x + \alpha_2 \max(0, x - x_{offset}) + W_0, \quad (1)$$

where x is position along axis (0 at shoulder/hips), $W(x)$ is width at position x, W_0 is width at $x = 0$, x_{offset} is a uniformly sampled point along the axis and α_1 and α_2 are constants empirically defined for each garment. For t-shirts and trousers, $\alpha_1 < 0 < \alpha_2$. For skirts, $\alpha_1 > \alpha_2 = 0$.

Cut. Template garments cover most of the body (long sleeves, legs and skirt). At this generation step, garments are cut to increase variability on length and topology. These cuts are along arms, legs and torso. Plus, upper-body garments have specific cuts to generate different types of garments (e.g., t-shirt, shirt, polo).

Resizing. Garments are resized to random body shapes. It is safe to assume that size variability on garments is similar to body shape variability. Following this reasoning,

SMPL shape displacements are transferred to garments by nearest neighbour. Nevertheless, this process is noisy and human body details are transferred to garment. To address these issues, an iterative laplacian smoothing is applied to shape displacements, removing noise and filtering high frequency body details, while preserving the geometry of the original garment. On SMPL, first and second shape parameters correspond to global human size and overall fatness. Knowing this, garments are resized to a different target shape. This new shape has two offsets at first and second parameters, the garment tightness $\gamma \in \mathbb{R}^2$. This offsets on garment resizing will generate loose or tight variability. As tighter garments present less dynamics and complexity, we bias the generator towards loose clothes by sampling tightness on the range $[-1.5, 0.5]$.

While first and second shape parameters represent overall size and fatness respectively, the sign of the first parameter has opposite meaning for male and female. To take this into account and so that tightness remains semantically consistent, the sign of the first offset is $(-1)^{g+1}$, where g is gender (0 = female and 1 = male). This means a positive tightness shall produce smaller garments.

Jumpsuits and dresses. Full-body garments can be generated by combining upper-body and lower-body garments. After generating the clothes individually, a final step automatically sews them together.

3.3. Simulation

Cloth simulation is performed on Blender, an open source 3D creation suite. Blender's cloth physics as it is in version 2.8 has been implemented with state-of-the-art algorithms on cloth simulation based on mass-spring model. Simulation performs $420 - 600$ steps per second, depending on the complexity of the garment.

Body transition. Outfit generation process yields garments on rest pose resized to SMPL shape plus tightness. We need a body transition from this state to the state of the initial frame of the sequence for a correct simulation. To ensure no body-to-cloth penetration is present due to resizing to a different shape, we generate a few frames of transition where the body shape changes from $\beta + \gamma$ (shape+tightness) to β. Finally, more frames are devoted to a transition from rest pose to the initial pose of the sequence split. Pose transition is computed based on quaternions for a smooth posing.

Fabrics. Changing the parameters of the mass-spring model allows simulation of different fabrics. Blender provides with different presets for *cotton*, *leather*, *silk* and *denim*, among others. These four fabrics have been used for the creation of the dataset. Upper-body garments might be cotton or silk, while the rest of the garment types can be any of those fabrics. Different fabrics produce different dynamics and wrinkles on simulation time.

Elastics. At simulation time, sleeves and legs have a 50% chance each of presenting an elastic behaviour at their ends, also at waist on full-body garments.

3.4. Variability and size

Each template garment is shaped with linear deformations, similar to SMPL shaping, where coefficients are uniformly sampled to yield a balanced distribution. Garments are cut at uniformly sampled lengths along limbs and waist/torso. With this we potentially obtain all possible combinations of sleeve length and t-shirt length, or leg length and waist height. Furthermore, upper-body garments have specifically designed cuts that change shape and topology, also uniformly sampled. Finally, on resizing stage, a tightness two-dimensional factor is uniformly sampled from $[-1.5, 0.5]$ (biased to loose garments for more complex dynamics and wrinkles), further increasing garment variability. All these randomly generated properties, once combined, almost guarantee that an outfit will never appear twice in the dataset. Afterwards, different simulation properties will also ensure cloth shall behave differently. Each pose sequence from source data is downgraded from 60fps to 30fps and split into 300 frames subsequences, each of them simulated once with a different outfit, thus totalling 2.1M frames.

3.5. Additional dataset statistics

Tab. 2 shows the CLOTH3D statistics in terms of action labels by grouping them into generic categories. Note that original data action label is very heterogeneus, specific and incomplete. These labels are gathered from CMU MoCap dataset. We observe a high density on *Walk*, but it is important to note that this gathers many different sub-actions (walk backwards, zombie walk, walk stealthily, etc) as many other action labels do. Additionally, most of these actions were performed by different subjects, which implies an increase in intra-class variability. The label 'others' contains all action labels that cannot be included in any of the categories plus all the missing action labels.

3.6. Data format

Each sample is a 600-frames split of the original sequence on source data, downgraded from 60fps to 30fps, totalling 300 frames for each split. Note that frames refer to time instants, not images. The name of the sample contains the name of the original sequence and the number of the split (e.g.: '01_01_s0', sequence is '01_01' and split is 's0'). As the number of frames of original sequences is not a multiple of 300, some splits will have smaller length. Each sample has static and dynamic garment information. **Static** information is the outfit fitted to rest pose for the corresponding body shape. Static garments are represented by OBJ files, which include vertices on rest position and topology

Walk	27.49%	Exercise	0.84%
Animal	10.79%	Climb	0.71%
Fight	4.38%	Carry	0.67%
Jump	2.78%	Stand	0.66%
Run	2.49%	Wash	0.63%
Sing	2.38%	Balancing	0.54%
Wait	2.31%	Trick	0.51%
Swim	1.97%	Sit	0.28%
Story	1.70%	Interact	0.20%
Sports	1.63%	Drink	0.14%
Dance	1.37%	Pose	0.14%
Yoga	1.01%	Bend	0.12%
Spin	0.90%	Others	33.36%

Table 2: CLOTH3D statistics per action label.

data. The **dynamic** information contains garment animation data. 3D animation data is usually stored as PC2 (Point Cache 2) files. We propose the PC16 format, a PC2 conversion from 32-bit floats to 16-bit, halving dataset space requirements. Precision loss on this conversion is none to minimal within the range $[-1, 1]$ ($\leq 2^{-11}$) and insignificant in $[-2, 2]$ ($\leq 2^{-10}$). By storing garment vertex positions relative to SMPL body root joint, we ensure values will always be in range $[-2, 2]$. Sample metadata contains SMPL parameters, garment names and their fabrics. To ease the fitting process, rest pose is redefined such that legs are slightly open, due to the high geometric complexity of this region. Average number of vertices per outfit is around 20K, which implies an average size of $35 - 40$MB per sequence.

Fig. 12 shows random static samples. Fig. 13 shows random frames of sequences. Finally, Fig. 14 shows random samples of different sequences with different representation modalities that can be obtained from CLOTH3D data: depth maps, surface normals, 3D velocities and segmentation masks.

4. Dressed human generation

This section presents the methodology for deep learning garment generation. A main challenge in garment modeling is its enormous variability in terms of garment types and topology. This may produce a variable input size and structure to the network. By exploiting the garments nature, there is a simplification to this problem which allows to encode the garment as a set of offsets from body surface to the garment surface [2, 16, 28]. Therefore, by fixing the body topology, data has a homogeneous dimensionality to fed the network architecture. Then, given body points, the problem becomes an estimation of the offsets. In addition, by masking body vertices we represent different garment types and separate them from the body, e.g. in a similar fashion to the segmentation in [19]. To compute ground truth offsets, a body-to-garment matching is needed. A ded-
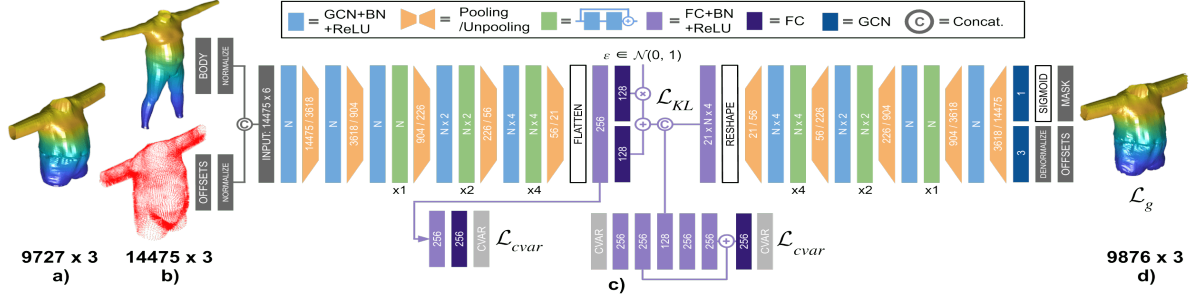
Figure 4: Model pipeline. a) Input garment b) body and offsets w.r.t. body (Sec. 4.1). Model input is the concatenation of body and offsets. c) Network architecture. Conditional variables (CVAR) are processed by an AutoEncoder. To improve latent space factorization, CVAR are also regressed from the first encoder FC layer. Decoder outputs are offsets and mask. d) Reconstruction of the garment by adding offsets to body and removing body vertices according to mask. We set N as 128.
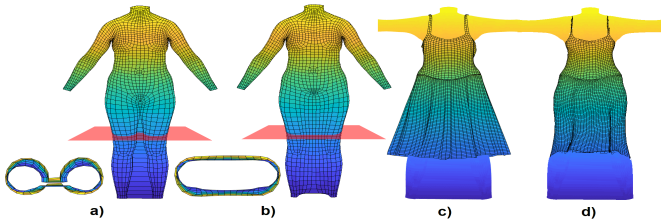

Figure 5: Dual topology and registration. a) New additional proposed topology, where inner legs are connected. This topology is used for graph convolutions as well. b) Result of Laplacian smoothing of inner leg vertices. It is used only for skirt/dress registration. We show top view of meshes around an imaginary red cutting plane. c) Garment in rest pose. d) Garment registered to body model.

icated algorithm for this task should be able to correctly register skirt-like garments which have a different topology than the body. In sec. 4.1 we explain details of our data pre-processing. Our proposed garment modeling is a Graph Conditional Variational Auto-Encoder (GCVAE). Given a known and fixed topology, we learn input offset features by convolutional operations. Besides, by conditioning on available metadata (pose, shape and tightness), we learn a latent space encoding specific information about garment type and its dynamics (details are given in sec. 4.2). Fig. 4 illustrates the proposed model.

4.1. Data pre-processing

We represent garments as a set of offsets connecting skin to cloth [2, 16, 28]. Additionally, a garment mask allows encoding different garments out of all body vertices. This methodology needs a matching between garment ($C \in \mathbb{R}^{V_C \times 3}$ from the dataset) and body ($T \in \mathbb{R}^{V_T \times 3}$ from SMPL model) vertices per sequence. We apply non-rigid ICP [3] to register C on top of T. The accuracy of registration (and consequently garment reconstruction) depends on: 1) body pose: registration is done in the rest pose while C and T are in rigid alignment. 2) resolution of the body vertices (i.e. V_T): default SMPL spatial density is not enough

to accurately encode garments, and fine-grain details can be lost. To solve this, we extend SMPL to SuperSMPL by subdividing the mesh, assigning model parameters to new vertices by linear interpolation w.r.t. their neighbours. Head, hands and feet are not used to find correspondences and removing them halves input dimensionality. This yields a final mesh with $V_T = 14475$ vertices. 3) type of the garment: skirt-like garments do not follow the same topology as SMPL mesh. For this task we introduce a novel topology where inner faces of the legs are removed and new faces that connect both legs are created, as seen in Fig. 5a. To have a better registration of skirts and dresses, we further apply a Laplacian smoothing iteratively to inner leg vertices of the new topology (see Fig. 5b). An example of the registration is shown in Fig. 5d. Finally, correspondences and garment mask are extracted by nearest neighbor matching. Fig. 7 shows an example of a reconstructed garment.

4.2. Network

As shown in Fig. 4, our network is based on a VAE generative model. The goal is to learn a meaningful latent space associated to the garments of any type, shape or with wrinkles which is used to generate realistic draped garments. Garment type and shape are associated to the static state of the garment while wrinkles are belonging to the dynamics of the garments. Here, we disentangle the latent space between statics and dynamics of the garments, and refer to learnt latent codes as garment code ($z_s \in \mathbb{R}^{128}$) and wrinkle code ($z_d \in \mathbb{R}^{128}$), respectively. To do so, we build two separate networks, one trained on static garments (so called SVAE) and one trained on dynamic garments (so called DVAE). To factorize the latent space from irrelevant parameters to the garment type and shape, we condition SVAE on body shape ($\beta \in \mathbb{R}^{11}$)[1] and garment tightness ($\gamma \in \mathbb{R}^2$). Likewise, DVAE is conditioned on β, γ, body pose ($\theta \in \mathbb{R}^{f \times 72}$) and z_s, where f is the number of frames in a temporal sequence. Let $cvar_s$ and $cvar_d$ be the stacking of conditioning variables of SVAE and DVAE in a single

---

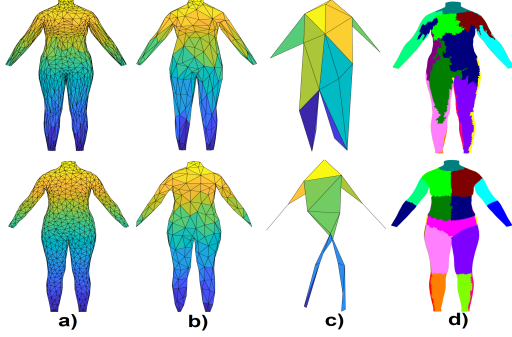[1]We include gender as an additional dimension to the shape parameters.

Figure 6: Mesh hierarchy for pooling. Upper: default [8]. Lower: proposed. Observe difference on spatial distribution at a) and b). c) shows how lowest pooling is more meaningful regarding the segments (one vertex per segment). d) is the visualization of correspondences (receptive field) between highest and lowest hierarchy levels.

vector. It is worth noting that θ is all zero (i.e. SMPL rest pose) in SVAE so that we do not include it in $cvar_s$.

A common network design for 3D data is done by fully connected layers when the order of the vertices is known [28, 23]. However, fully connected networks can stick to local minima when the input dimensionality is huge (14475×3 in this work). A solution is to build geometrical descriptors (e.g. invariant to rotation) and feed computed features to the network [23]. Although this can bypass local minima, output features need a post-processing to reconstruct garments. Instead, we build our network based on graph convolutional networks.

Graph convolution. Our network input data is defined by V_T and topology (Sec.4.1). This allows to use graph convolutional filters to learn features. Following the definition of spectral graph convolution [6, 16, 17, 30], filtering is computed as:

$$y = g_\omega(\mathbf{L})x = \sum_{i=0}^{K} \omega_i \mathbf{T}_i(\hat{\mathbf{L}})x, \qquad (2)$$

where ω_i are the learnable filters, $\mathbf{L}$ is the normalized Laplacian matrix, $\hat{\mathbf{L}} = 2\mathbf{L}/\lambda_{max} - \mathbf{I}$, and $\mathbf{T}_i(\hat{\mathbf{L}})$ is the i-th Chebyshev polynomial order. Given an input graph with F_{in} features for each node, the described convolution will return that same graph with a different set of features F_{out}. Chebysev polynomial order defines the size of the receptive field K, meaning feature filtering aggregates the K-ring neighbourhood for each node. In our case, we keep an small neighborhood with $K = 1$. Therefore to have a high receptive field we build a deep network. Each convolutional and pooling layer further combines node features with higher K-ring neighbours. We also include skip connections throughout the whole network. This leads to an effective information passing helping to learn more details of the garment. We refer to [5, 6, 27] for further details on graph convolutions.

Pooling. We resort to a mesh simplification algorithm [8] to create a hierarchy of meshes with decreasing details in order to implement the pooling operator. We follow [30] to have vertices uniformly distributed in the graph coarsening. However, this approach does not guarantee a uniform or meaningful receptive field on a high resolution mesh. To achieve a homogeneous distribution of correspondences throughout the body between pooling layers, we define a segmentation (Fig. 6(d)) and forbid the algorithm from contracting edges connecting vertices of different segments. Segmentation contains 21 segments and it is designed such that regions of the body with highest offset variability have smaller segments. Thus, more capacity of the network is available to model those parts. See Fig. 6. Our mesh hierarchy is formed by 6 different levels. The dimensionality of those meshes is: $14475 \rightarrow 3618 \rightarrow 904 \rightarrow 226 \rightarrow 56 \rightarrow 21$, leaving a single node for each segment on the last pooling layer. We use max-pooling in the proposed hierarchy. For unpooling, features are copied to all corresponding vertices of the immediate higher mesh.

Architecture. Let $X_s \in \mathbb{R}^{V_T \times 3}$ and $X_d \in \mathbb{R}^{V_T \times 3}$ be offsets computed on static and dynamic samples, respectively. From now on we use subscript s and d for static and dynamic variables and discard them for general cases. We normalize i-th vertex offset as $\bar{X}^i = \Omega(X^i, \mu^i, \sigma^i) = (X^i - \mu^i)/(\sigma^i + \epsilon)$ where μ^i and σ^i are mean and variance of i-th vertex. Likewise, body vertices T are normalized to $\bar{T}$. SVAE and DVAE have a similar structure with three main modules: encoder $\{cvar^z, z\} = \Psi(\bar{X}, \bar{T}; \tau^\Psi)$, conditioning $\{cvar, cvar^z\} = \Gamma(cvar; \tau^\Gamma)$ and decoder $\{\bar{X}, M\} = \Phi(z, cvar^z; \tau^\Phi)$, where $M \in \mathbb{R}^{V_T \times 1}$ is the garment mask and $\tau^\cdot$ are networks weights. Conditioning network Γ is an autoencoder with one skip connection and $cvar^z$ is its middle layer features. The goal of this network is to provide a trade-off between $cvar$ and z. The architecture details are shown in Fig. 4. Note that all GCN layer features (except first and last layers) are doubled in DVAE vs. SVAE.

As a standard operation in CVAE, conditional variables should be fed into the encoder. However, these variables ($cvar$) are not balanced with offsets (X) in terms of size and scale. $cvar$ can be partially decoded to body vertices $\bar{T}$ and offsets $\bar{X}$. Therefore, we concatenate $\bar{T}$ to $\bar{X}$ and feed it to the network Ψ. To better factor out $cvar$ from latent code z we include an additional MLP branch at the end of the encoder and before sampling layer. The goal of this branch is to regress $cvar$ during training. It also helps to have a more stable training. Regularizing the encoder by regressing $cvar$ has a limitation when the dimensionality of $cvar$ is high (e.g. $cvar_d$), that is, optimization can stick to local minima. Therefore, we regress $cvar^z$ instead of $cvar$. Finally, decoder generates normalized offsets $\bar{X}$ and

garment mask M in two branches at the last layer. Note that DVAE does not have branch M in the decoder. At the end, the garment is reconstructed by $\{\Omega^{-1}(\bar{X}, \mu, \sigma) + T | M > 0.5\}$ where Ω^{-1} is the reverse normalization operator. Note that we train our GCN with dual topology (as explained in Sec. 4.1) to handle skirt-like garments vs. the rest which has not been done before in 3D garment reconstruction.

Wrinkle factorization vs. the rest conditional variables. We introduce two versions of DVAE. In one version, we computed offsets X as $C - T$ and concatenated with T as input to the network. In this version, offsets are not invariant to pose. To better factorize z_d to encode garment wrinkles, we also implement another version of DVAE where offsets are partially invariant to pose and garment type. Let $\mathcal{K}(A, \theta, J)$ be forward kinematic function that receives articulated object $A \in \mathbb{R}^{V_A \times 3}$ with joints $J \in \mathbb{R}^{V_J \times 3}$ and transforms it w.r.t. the rotations (or pose) θ. Note that A and J are in rest pose in this definition. Here, A can be replaced by C or T. Since we register garments on top of body, garments can be transformed by body joints. Given this definition, $\mathcal{K}^{-1}$ can be defined as the inverse of the kinematic function which transforms back the object to its rest pose.

We define new offsets as $X^{new} = \mathcal{K}^{-1}(C_d, \theta, J) - C_s$ where C_d is the dynamic dressed garment in pose θ and C_s is the static garment in rest pose. We then concatenate X^{new} with C_s and T, normalize it and feed the network. Finally, at the output of the network, garments are reconstructed by the reverse process. Since we train SVAE to reconstruct C_s, we can rely on it at test time to generate dressed garments. By doing this factorization on the offsets, we are able to feed the network with relevant information of the wrinkles and factorize z_d in a more meaningful way.

Loss. We train conditioning network Γ independently using $L1$ loss and freeze its weights (τ^Γ) while training VAE. S/DVAE loss is a combination of a garment related term, a $cvar$ term and KL-divergence:

$$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_{cvar} + \lambda_{KL} D_{KL}(q(z|X, cvar)||p(z|cvar)), \quad (3)$$

where $q(z|X, cvar)$ is the posterior and $p(z|cvar)$ the prior. Garment related term handles offsets and mask (if available). Additionally, mesh face normals are included into the loss to improve recovered geometry:

$$\mathcal{L}_g = \mathcal{L}_o + \lambda_n \mathcal{L}_n + \lambda_m \mathcal{L}_m, \quad (4)$$

where $\mathcal{L}_o$, $\mathcal{L}_n$, and $\mathcal{L}_m$ are the $L1$ reconstruction losses on offsets, normals and mask, respectively. λ_n and λ_m are balancing coefficients of normals and mask. $\mathcal{L}_{cvar}$ is $L1$ loss on encoder $cvar^z$ regressor.

Implementation. Models have been implemented on TensorFlow, trained with a learning rate of $\lambda = 0.001$ and a batch size of 32. Balancing coefficient λ_{KL} is dynamically increased to obtain a faster stable training. We start training
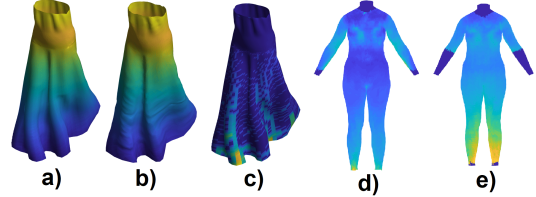

Figure 7: Implicit registration and S/DVAE per vertex error. a) original garment. b) reconstruction from offsets and body vertices. c) original vs. reconstruction error heatmap (scale in cm.). d-e) SVAE vs. DVAE garment generation error heatmaps. Dark blue=0, yellow>10cm.

at $\lambda_{KL} = 0.0001$ for 50 epochs and increase it gradually by 0.005 per epoch. λ_n and λ_m are 70 and 0.1 respectively. We used Adam optimizer and trained for 100 epochs. We save the best model with lowest error on validation set w.r.t. $\mathcal{L}_o + 0.1 D_{KL}$.

5. Experiments

5.1. Metrics

Surface error. Given that input and prediction have the same dimensionality and order, we use standard L2-norm.

Normals error. In 3D domain, normals encode local geometric information that can be used as a measure of surface quality. We compute normals error based on mesh face normals by their angle difference to ground truth normals:

$$\mathcal{E}_n = \frac{1}{\sum M^f} \sum_{i=1}^{\#M^f} M_i^f \arccos\left(\frac{(\mathbf{n}_i)^T \times \mathbf{n}_i^g}{\|\mathbf{n}_i\|\|\mathbf{n}_i^g\|}\right), \quad (5)$$

where M^f is the binary mask of garment faces, $\mathbf{n}$ and $\mathbf{n}^g$ are predicted and groundtruth face normals, respectively.

Mask IoU. Predicted garment mask is evaluated by the intersection over union (IoU).

KL loss. We use KL loss as a measure of quality of latent code factorization and meaningfulness of the latent space.

5.2. Ablation study

We trained SVAE on an additional dataset of static samples (in rest pose) with 30K samples. 20% of the data is kept for evaluation and the rest for training. The results are shown in Tab. 3a and 3b.

Normals. Looking at the second row of Tab.3a we observe that enforcing a reconstruction consistent with normals significantly reduces surface error and, as expected normals error. However, including normals has a negative impact on KL loss comparing to first row.

Mask. To see if predicting mask is benefitial or detrimental for reconstruction, we performed an experiment without it. As seen in third row of Tab. 3a, both, surface and normals error are significantly higher without mask prediction (comparing to first row).

	Surface error	Normals error	Mask (IoU)	KL loss
All	14.7	1.02	0.9522	0.9414
All without normals	22.8	1.07	0.9472	0.5966
All without mask	92.7	1.19	-	0.8799
All without regressing conditional variables	14.8	1.02	0.9520	1.1009
All with default pooling	14.9	1.03	0.9390	0.7623

(a)

	Surface error	Normals error	Mask (IoU)	KL loss
Top	12.4	1.19	0.9038	0.9229
T-shirt	15.9	1.20	0.9569	1.1380
Trousers	11.4	0.82	0.9475	0.8728
Skirt	21.3	0.79	0.9518	0.9952
Jumpsuit	13.8	1.04	0.9638	0.8543
Dress	16.8	1.05	0.9665	0.9737

(b)

Table 3: (a) Ablation results on the static dataset for all clothes. (b) Ablation results (full model) on the static dataset for each cloth category. Surface and normal errors are shown in mm and radians, respectively.

# frames	Top	T-shirt	Trousers	Skirt	Jumpsuit	Dress	Avg.
1	22.3/1.22	29.5/1.29	21.0/0.88	37.6/0.91	29.6/1.10	35.5/1.13	29.3/1.09
4	20.3/1.22	28.1/1.28	18.7/0.87	33.2/0.88	26.3/1.09	32.3/1.11	26.4/1.08

Table 4: Ablation results (full model) on the dynamic dataset conditioning on different number of frames. Left: surface error (mm) / Right: normals error (radians).

CVARs. As explained in Sec.4.2, conditional variables are regressed from the first FC layer of the encoder to improve latent space factorization. On fourth row of Tab. 3a we can see that, while surface or normals error have no significant differences, KL loss improves.

Pooling. On Sec.4.2 we discussed different approaches for tackling the pooling on a graph neural network. To do this, we built a mesh hierarchy. We compared default mesh simplification algorithm versus our proposed modification. Results are shown in the last row of Tab. 3a. While improvement on surface and normals errors is marginal, this new pooling benefits mask prediction.

Per garment category error. Results per garment are shown in Tab. 3b. Skirts present the highest surface error, as its vertices are far away from the body compared to other garments. Following this reasoning, we find trousers having the less surface error. If we look at normals error, we find an opposite behaviour for skirts, as their geometry is the simplest one. On the other hand we see that upper-body garments present more complex geometries, and therefore, higher normals error. Looking at mask error, we see that garments that cover most of the body have the lowest error. This is due to IoU metric nature, the lower the number of points, the more impact shall have each wrong prediction. Finally, looking at KL loss, we observe the model has difficulties to obtain meaningful spaces for T-shirts. As explained on Sec.3.2, T-shirts category includes open shirts as well, which highly increases class variability. We also see that trousers and jumpsuits have the lowest KL loss.

Garment latent space. In Fig. 8, we show distribution of 5K random static samples computed by t-SNE algorithm. As one can see, the proposed GCVAE network can group garments in a meaningful space. Interestingly, dress and jumpsuit that share more vertices also share the same latent space. Additionally, we show garment transitions in this space in Fig. 9. One can see how garments transit between two different topologies (3rd row) or among different
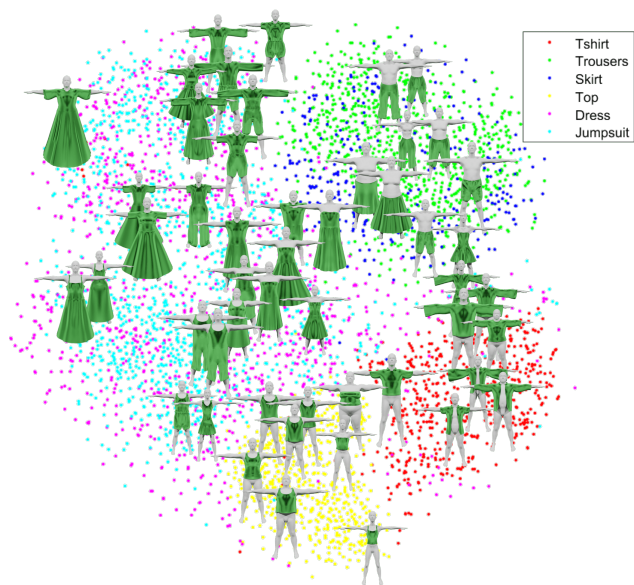

Figure 8: Visualization of the learned latent space for static samples using t-SNE algorithm.

genders and shapes (4th row).

Dynamic garment generation. We study DVAE model in Tab. 4. We condition DVAE on pose for a single frame vs. four frames. Four frames are selected every 3 frames, resulting in a 12-frame clip. Training the model on a sequence of frames leads to better results in all garment categories (3mm improvement in average). This is while we do not include any temporal information in the encoder nor any specific sequence prediction loss. Per vertex reconstruction error is shown in Fig. 7 for static and dynamic samples. It can be seen the error is higher near the feet. This is because of dynamics of skirts and dresses which have the highest error in Tab. 4. Interestingly, trousers has the lowest error among others. Some qualitative images of reconstructed dynamic samples are shown at right of Fig. 1.
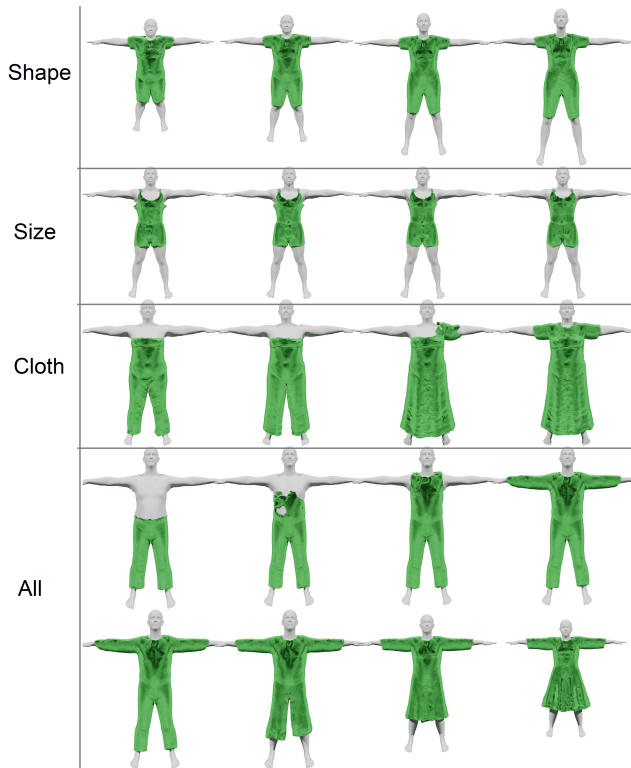
9

Figure 9: Transitions of static samples. First three rows: conditioning on shape, tightness or cloth while the rest are fixed. Last two rows: transition of all variables. Variables are linearly graduated.

Wrinkles latent space In Fig. 10 we show some qualitative results of the learnt latent space and conditioning on different variables, specifically pose (θ), garment code (z_s) and wrinkle code (z_d). One can see the network can learn a meaningful and consistent space. Regarding the learnt wrinkle code (Fig. 10c), a rest pose (upper row) shows less wrinkle variability than a complex action category (lower row). This is while by conditioning on pose (Fig. 10a) or garment code (Fig. 10b), we can accurately retarget fixed wrinkle codes to new scenarios.

5.3. Applications of the dataset

CLOTH3D could be used not just for 3D garment generation but in other application scenarios, such as human pose and action recognition in depth images, garment motion analysis, filling missing vertices of scanned bodies with additional meta data (e.g. garment segments), support designers and animators tasks, or estimating 3D garment from RGB images, just to mention a few. We ran some proof-of-concept applications using our CLOTH3D data, shown in Fig. 11: a rendered depth image, garment motion velocities, and RGB-to-3D cloth estimation. For the later, given our layered garment structure and SMPL segmentation, we rendered 10K samples of t-shirts and trousers with different poses of Human3.6M [11] dataset. These data contains images of body segments and garment silhouette. We then trained ResNet50 to regress available static garment codes. In test time, we assume body shape, pose and garment silhouette are available. These information can be extracted by state-of-the-art SMPL based pose estimation and cloth parsing methods. In our case, we manually segmented garments of two frames of this dataset (shown in Fig. 11c) and used them to estimate garment code. We then copied the wrinkles from the nearest sample in CLOTH3D. Finally 3D garments were reconstructed and rendered as shown.

6. Conclusions

In this paper we presented CLOTH3D, the first large scale synthetic dataset of 3D clothed humans. It includes a large data variability in terms of body shape and pose, garment type, topology, shape, tightness and fabric. Generated garments also show complex dynamics, providing with a challenging corpus for 3D garment generation. We developed a baseline method using a graph convolutional network trained with a variational autoencoder, and proposed a new pooling grid. Evaluation of the proposed GCVAE on CLOTH3D showed realistic garment generation.

7. Acknowledgements

References

[1] Carnegie-Mellon Mocap Database. http://http://mocap.cs.cmu.edu/. 3

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018. 1, 2, 5, 6

[3] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 6

[4] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5420–5430, 2019. 2

[5] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning:

(a) Pose variation.

(b) Garment variation.

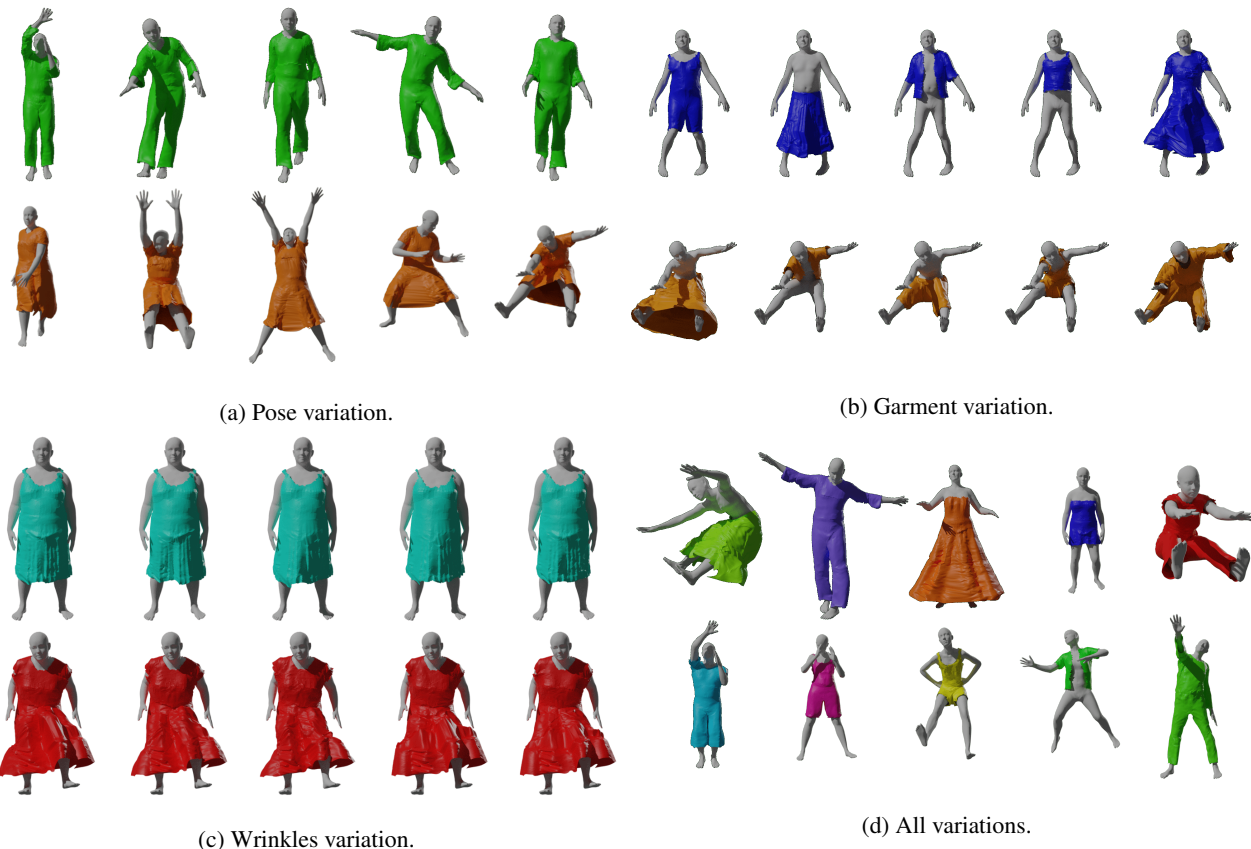(c) Wrinkles variation.

(d) All variations.

Figure 10: Generated samples from learnt latent codes, conditioning on different variables for dynamic garments (DVAE network).
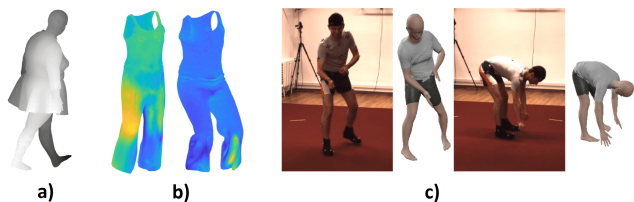

Figure 11: Applications of CLOTH3D. a): A rendered depth image, b): vertex velocities (Dark blue=0, yellow>0.8.), c): Automatic RGB to 3D cloth estimation.

going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. 2, 7

[6] 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, 7

[7] Qi Dong, Shaogang Gong, and Xiatian Zhu. Multi-task curriculum transfer deep learning of clothing attributes. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 520–529. IEEE, 2017. 1

[8] Michael Garland and Paul S Heckbert. Surface simplification using quadric error metrics. In *Proceedings of the 24th an-*

nual conference on Computer graphics and interactive techniques, pages 209–216. ACM Press/Addison-Wesley Publishing Co., 1997. 7

[9] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. *ACM Trans. Graph.*, 31(4):35–1, 2012. 2

[10] 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 *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 2

[11] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 36(7):1325–1339, jul 2014. 10

[12] Kevin Lin, Huei-Fang Yang, Kuan-Hsien Liu, Jen-Hao Hsiao, and Chu-Song Chen. Rapid clothing retrieval via deep learning of binary codes and hierarchical search. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 499–502. ACM, 2015. 1

[13] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):220, 2014. 3
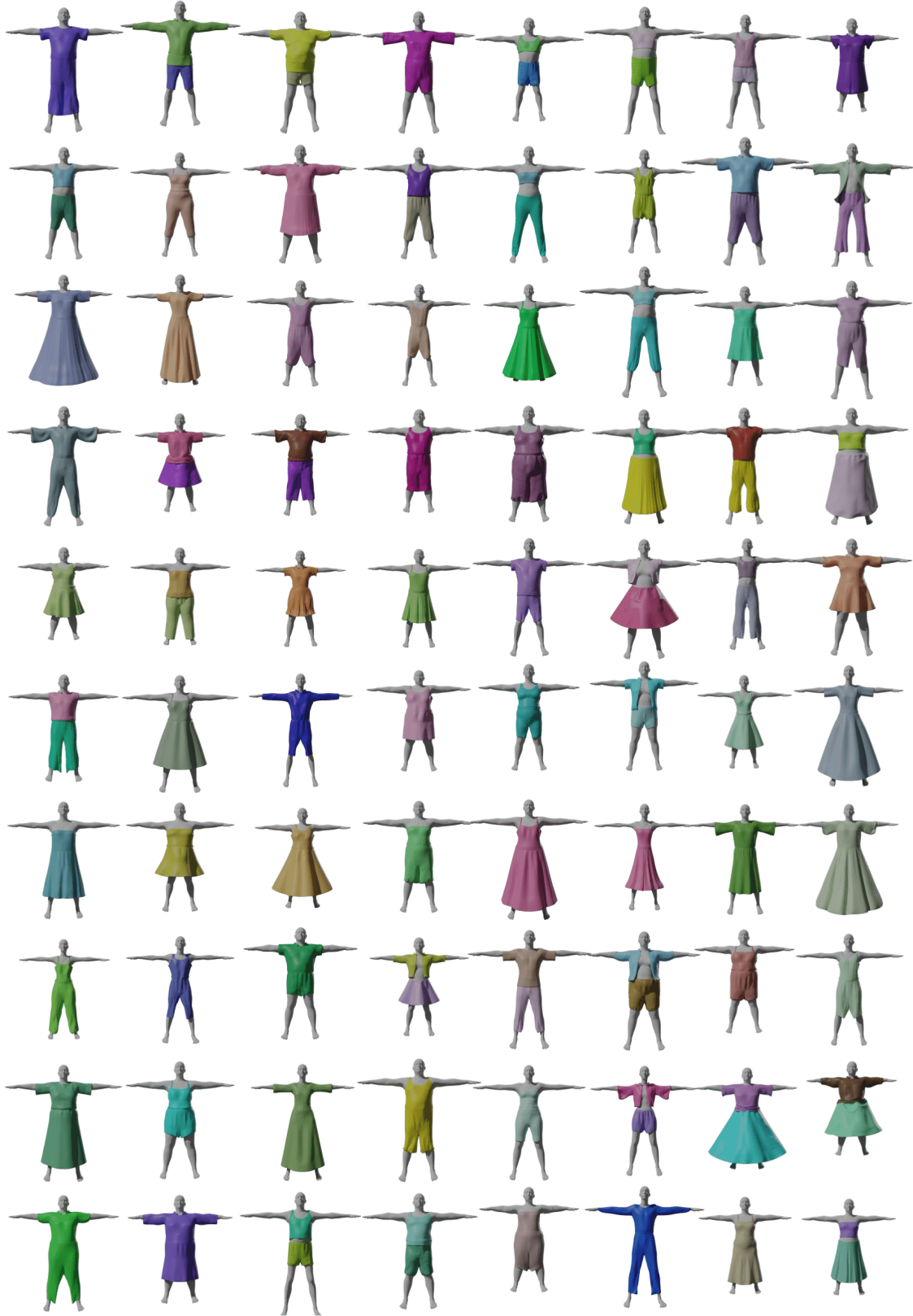
Figure 12: CLOTH3D static samples.

Figure 13: CLOTH3D dynamic samples.

Figure 14: CLOTH3D data representations. From left to right: RGB, depth, normals, velocities, and segmentation masks.

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

[15] 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. 3

[16] Qianli Ma, Siyu Tang, Sergi Pujades, Gerard Pons-Moll, Anurag Ranjan, and Michael J Black. Dressing 3d humans using a conditional mesh-vae-gan. *arXiv preprint arXiv:1907.13615*, 2019. 1, 2, 5, 6, 7

[17] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *International conference on machine learning*, pages 2014–2023, 2016. 2, 7

[18] Sergey I. Nikolenko. Synthetic data for deep learning. *ArXiv*, abs/1909.11512, 2019. 1

[19] 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. 1, 2, 5

[20] Albert Pumarola, Vedanuj Goswami, Francisco Vicente, Fernando De la Torre, and Francesc Moreno-Noguer. Unsupervised image-to-video clothing transfer. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 1

[21] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[22] Dongjoe Shin and Yu Chen. Deep garment image matting for a virtual try-on system. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 1

[23] Qingyang Tan, Lin Gao, Yu-Kun Lai, and Shihong Xia. Variational autoencoders for deforming 3d mesh models. In *Pro-ceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5841–5850, 2018. 7

[24] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017. 1, 3

[25] 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 *European Conference on Computer Vision (ECCV)*, sep 2018. 2

[26] Tuanfeng Y Wang, Duygu Ceylan, Jovan Popovic, and Niloy J Mitra. Learning a shared shape space for multimodal garment design. *arXiv preprint arXiv:1806.11335*, 2018. 2

[27] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596*, 2019. 2, 7

[28] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Analyzing clothing layer deformation statistics of 3d human motions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 237–253, 2018. 1, 2, 5, 6, 7

[29] Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, and Yebin Liu. Simulcap: Single-view human performance capture with cloth simulation. *arXiv preprint arXiv:1903.06323*, 2019. 2

[30] Yu-Jie Yuan, Yu-Kun Lai, Jie Yang, Hongbo Fu, and Lin Gao. Mesh variational autoencoders with edge contraction pooling. *arXiv preprint arXiv:1908.02507*, 2019. 2, 7

[31] Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4191–4200, 2017. 2