Multi-Garment Net: Learning to Dress 3D People from Images

Bharat Lal Bhatnagar  Garvita Tiwari  Christian Theobalt  Gerard Pons-Moll  
Max Planck Institute for Informatics, Saarland Informatics Campus, Germany  
{bbhatnag,gtiwari,theobalt,gpons}@mpi-inf.mpg.de

Abstract

We present Multi-Garment Network (MGN), a method to predict body shape and clothing, layered on top of the SMPL [40] model from a few frames (1-8) of a video. Several experiments demonstrate that this representation allows higher level of control when compared to single mesh or voxel representations of shape. Our model allows to predict garment geometry, relate it to the body shape, and transfer it to new body shapes and poses. To train MGN, we leverage a digital wardrobe containing 712 digital garments in correspondence, obtained with a novel method to register a set of clothing templates to a dataset of real 3D scans of people in different clothing and poses. Garments from the digital wardrobe, or predicted by MGN, can be used to dress any body shape in arbitrary poses. We will make publicly available the digital wardrobe, the MGN model, and code to dress SMPL with the garments at [1].

1. Introduction

The 3D reconstruction and modelling of humans from images is a central problem in computer vision and graphics. Although a few recent methods [5, 3, 4, 25, 41, 51] attempt reconstruction of people with clothing, they lack realism and control. This limitation is in great part due to the fact that they use a single surface (mesh or voxels) to represent both clothing and body. Hence they can not capture the clothing separately from the subject in the image, let alone map it to a novel body shape.

In this paper, we introduce Multi-Garment Network (MGN), the first model capable of inferring human body and layered garments on top as separate meshes from images directly. As illustrated in Fig. 1, this new representation allows full control over body shape, texture and geometry of clothing and opens the door to a range of applications in VR/AR, entertainment, cinematography and virtual try-on.

Compared to previous work, MGN produces reconstructions of higher visual quality, and allows for more control: 1) we can infer the 3D clothing from one subject, and dress a second subject with it, (see Fig. 1) and 2) we can trivially map the garment texture captured from images to any garment geometry of the same category (see Fig. 7).

To achieve such level of control, we address two major challenges: learning per-garment models from 3D scans of people in clothing, and learning to reconstruct them from images. We define a discrete set of garment templates (according to the categories long/short shirt, long/short pants and coat) and register, for every category, a single template to each of the scan instances, which we automatically segmented into clothing parts and skin. Since garment geometry varies significantly within one category (e.g. different shapes, sleeve lengths), we first minimize the distance between template and the scan boundaries, while trying to preserve the Laplacian of the template surface. This initialization step only requires solving a linear system, and nicely stretches and compresses the template globally, which we found crucial to make subsequent non-rigid registration work. Using this, we compile a digital wardrobe of real 3D garments worn by people, (see Fig. 3). From such registrations, we learn a vertex based PCA model per garment. Since garments are naturally associated with the underlying SMPL body model, we can transfer them to different body shapes, and re-pose them using SMPL.
The digital wardrobe, MGN is trained to predict, given one or more images of the person, the body pose and shape parameters, the PCA coefficients of each of the garments, and a displacement field on top of PCA that encodes clothing detail. At test time, we refine this bottom-up estimates with a new top-down objective that forces projected garments and skin to explain the input semantic segmentation. This allows more fine-grained image matching as compared to standard silhouette matching. Our contributions can be summarized as:

- A novel data driven method to infer, for the first time, separate body shape and clothing from just images (few RGB images of a person rotating in front of the camera).
- A robust pipeline for 3D scan segmentation and registration of garments. To the best of our knowledge, there are no existing works capable of automatically registering a single garment template set to multiple scans of real people with clothing.
- A novel top-down objective function that forces the predicted garments and body to fit the input semantic segmentation images.
- We demonstrate several applications that were not previously possible such as dressing avatars with predicted 3D garments from images, and transfer of garment texture and geometry.
- We will make publicly available the MGN to predict 3D clothing from images, the digital wardrobe, as well as code to “dress” SMPL with it.

2. Related Work

In this section we discuss the two branches of work most related to our method, namely capture of clothing and body shape and data-driven clothing models.

**Performance Capture.** The classical approach to bring dynamic sequences into correspondence is to deform meshes non-rigidly [11, 18, 10] or volumetric shape representations [28, 2] to fit multiple image silhouettes. Without a pre-scanned template, fusion [30, 29, 55, 53] trackers incrementally fuse geometry and appearance [66] to build the template on the fly. Although flexible, these require multi-view [57, 37, 14], one or more depth cameras [19, 49], or require the subject to stand still while turning the cameras around them [53, 38, 63, 16]. From RGB video, Habermann et al. [25] introduced a real time tracking system to capture non-rigid clothing dynamics. Very recently, SimulCap [59] allows multi-part tracking of human performances from a depth camera.

**Body and cloth capture from images and depth.** Since current statistical models can not represent clothing, most works [7, 26, 40, 58, 48, 32, 22, 67, 51, 50, 54, 53, 44, 46] are restricted to inferring body shape alone. Model fits have been used to virtually dress and manipulate people’s shape and clothing in images [50, 67, 62, 36]. None of these approaches recover 3D clothing. Estimating body shape and clothing from an image has been attempted in [24, 12], but it does not separate clothing from body and requires manual intervention [65, 49]. Given a depth camera, Chen et al. [13] retrieve similar looking synthetic clothing templates from a database. Daněřek et al. [9] use physics based simulation to train a CNN but do not estimate garment and body jointly, require pre-specified garment type, and the results can only be as good as the synthetic data.

Closer to ours is the work of Alldieck et al. [3, 5, 6] which reconstructs, from a single image or a video, clothing and hair as displacements on top of SMPL, but can not separate garments from body, and can not transfer clothing to new subjects. In stark contrast to [3], we register the scan garments (matching boundaries) and body separately, which allows us to learn the mapping from images to a multi-layer representation of people.

**Data-driven clothing.** A common strategy to learn ef-
ficient data-driven models is to use off-line simulations \cite{17, 34, 21, 32, 23} for generating data. These approaches often lack realism when compared to models trained using real data. Very few approaches have shown models learned from real data. Given a dynamic scan sequence, Neophytou et al. \cite{42} learn a two layer model (body and clothing) and use it to dress novel shapes. A similar model has been recently proposed \cite{61}, where the clothing layer is associated to the body in a fuzzy fashion. Other methods \cite{60, 64} focus explicitly on estimating the body shape under clothing. Like these methods, we treat the underlying body shape as a layer, but unlike them, we segment out the different garments allowing sharp boundaries and more control. For garment registration, we build on the ideas of ClothCap \cite{47}, which can register a subject specific multi-part model to a 4D scan sequence. By contrast, we register a single template set to multiple scan instances—varying in garment geometry, subject identity and pose, which requires a new solution. Most importantly, unlike all previous work \cite{35, 47, 61}, we learn per-garment models and train a CNN to predict body shape and garment geometry directly from images.

3. Method

In order to learn a model to predict body shape and garment geometry directly from images, we process a dataset of 356 scans of people in varied clothing, poses and shapes. Our data pre-processing (Sec. 3.1) consists of the following steps: SMPL registration to the scans, body aware scan segmentation and template registration. We obtain, for every scan, the underlying body shape, and the garments of the person registered to one of the 5 garment template categories: shirt, t-shirt, coat, short-pants, long-pants. The obtained digital wardrobe is illustrated in Fig. 3. The garment templates are defined as regions on the SMPL surface; the original shape follows a human body, but it deforms to fit each of the scan instances after registration. Since garment registrations are naturally associated to the body represented with SMPL, they can be easily reposed to arbitrary poses. With this data, we train our Multi-Garment Network to estimate the body shape and garments from one or more images of a person, see Sec. 3.2.

3.1. Data Pre-Processing: Scan Segmentation and Registration

Unlike ClothCap \cite{47} which registers a template to a 4D scan sequence of a single subject, our task is to register single template across instances of varying styles, geometries, body shapes and poses. Since our registration follows the ideas of \cite{47}, we describe the main differences here.

Body-Aware Scan Segmentation We first automatically segment the scans into three regions: skin, upper-clothes and pants (we annotate the garments present for every scan). Since even SOTA image semantic segmentation \cite{20} is inaccurate, naïve lifting to 3D is not sufficient. Hence, we incorporate body specific garment priors and segment scans by solving an MRF on the UV-map of the SMPL surface after non-rigid alignment.

A garment prior (for garment $g$) derives from a set of labels $I^g_i \in \{0, 1\}$ indicating the vertices $v_i \in S$ of SMPL that are likely to overlap with the garment. The aim is to penalize labelling vertices as $g$ outside this region, see Fig. 3. Since garment geometry varies significantly within one category (e.g. t-shirts of different sleeve lengths), we define a cost increasing with the geodesic distance $\text{dist}_{\text{geo}}(v) : S \rightarrow \mathbb{R}$ from the garment region boundary—efficiently computed based on heat flow \cite{15}. Conversely, we define a similar penalty for labeling vertices in the garment region with a label different than $g$. As data terms, we incorporate CNN based semantic segmentation \cite{20}, and appearance terms based Gaussian Mixture Models in $La$ color space. The influence of each term is illustrated in Fig. 4 for more details we refer to the supp. mat.

After solving the MRF on the SMPL UV map, we can segment the scans into 3 parts by transferring the labels from the SMPL registration to the scan.

Garment Template We build our garment template on top of SMPL+D. $M(\cdot)$, which represents the human body as a parametric function of pose($\theta$), shape($\beta$), global translation($t$) and optional per-vertex displacements ($D$):

$$M(\theta, \beta, D) = W(T(\beta, \theta, D), J(\beta, \theta, W)) \quad (1)$$

$$T(\beta, \theta, D) = T + B_s(\beta) + B_p(\theta) + D. \quad (2)$$

The basic principle of SMPL is to apply a series of linear displacements to a base mesh $T$ with $n$ vertices in a T-pose, and then apply standard skinning $W(\cdot)$. Specifically, $B_s(\cdot)$ models pose-dependent deformations of a skeleton $J$, and $B_p(\cdot)$ models the shape dependent deformations. $W$ represents the blend weights.

For each garment class $g$ we define a template mesh, $G^g$ in T-pose, which we subsequently register to explain the scan garments. We define $\mathbf{I}^g \in \mathbb{Z}^{m_g \times n}$ as an indicator matrix, with $I^g_{ij} = 1$ if garment $g$ vertex $i \in \{1 \ldots m_g\}$ is associated with body shape vertex $j \in \{1 \ldots n\}$. In our experiments, we associate a single body shape vertex to each garment vertex. We compute displacements to the corresponding SMPL body shape $\beta^g$ under the garment as

$$D^g = G^g - \mathbf{I}^g T(\beta^g, 0_\theta, 0_\beta) \quad (3)$$

Consequently, we can obtain the garment shape (unposed), $T^g$ for a new shape $\beta$ and pose $\theta$ as

$$T^g(\beta, \theta, D^g) = \mathbf{I}^g T(\beta, \theta, 0) + D^g \quad (4)$$

To pose the vertices of a garment, each vertex uses the skinning function in Eq. 1 of the associated SMPL body vertex.

$$G(\beta, \theta, D^g) = W(T^g(\beta, \theta, D^g), J(\beta, \theta, W)) \quad (5)$$
Garment Registration

Given the segmented scans, we non-rigidly register the body and garment templates (upper-clothes, lower-clothes) to scans using the multi-part alignment proposed in [47]. The challenging part is that garment geometries vary significantly across instances, which makes the multi-part registration fail (see supplementary). Hence, we first initialize by deforming the vertices of each garment template with the shape and pose of SMPL registrations, obtaining deformed vertices $G_{g}^{\text{init}}$. Note that since the vertices defining each garment template are fixed, the clothing boundaries of the initially deformed garment template will not match the scan boundaries. In order to globally deform the template to match the clothing boundaries in a single shot, we define an objective function based on Laplacian deformation [56].

Let $L^{g} \in \mathbb{R}^{m_{g} \times m_{g}}$ be the graph Laplacian of the garment mesh, and $\Delta_{\text{init}}^{g} \in \mathbb{R}^{m_{g} \times 3}$ the differential coordinates of the initially deformed garment template $\Delta_{\text{init}}^{g} = L^{g} G_{g}^{\text{init}}$. For every vertex $s_{i} \in S_{g}$ in a scan boundary $S_{g}$, we find its closest vertex in the corresponding template garment boundary, obtaining a matrix of scan points $q_{1:C} = \{q_{1}, \ldots, q_{C}\}$ with corresponding template vertex indices $j_{1:C}$. Let $I_{C \times m_{g}}$ be a selector matrix indicating the indices in the template corresponding to each $q_{k}$. With this, we minimize the following least squares problem:

$$
\begin{bmatrix}
L^{g} \\
I_{C \times m_{g}}
\end{bmatrix} G^{g} =
\begin{bmatrix}
\Delta_{\text{init}} \\
I_{C \times m_{g}} q_{1:C}
\end{bmatrix}
$$

(6)

with respect to the template garment vertices $G^{g}$, where the first block $L^{g} G^{g} = \Delta_{\text{init}}^{g}$ forces the solution to keep the local surface structure, while the second block $I_{C \times m_{g}} G^{g} = I_{C \times m_{g}} q_{1:C}$ makes the boundaries match. The nice property of the linear system solve is that the garment template globally stretches or compresses to match the scan garment boundaries, which would take many iterations of non-linear non-rigid registration [47] with the risk of converging to bad local minima. After this initialization, we non-linearly register each garment $G^{g}$ to fit the scan surface. We build on top of the proposed multi-part registration in [47] and propose additional loss terms on garment vertices, $v_{k} \in G^{g}$, to facilitate better garment unposing, $E_{\text{unpose}}$, and minimize interpenetration, $E_{\text{interp}}$, with the underlying SMPL body surface, $S$.

$$
E_{\text{interp}} = \sum_{g} \sum_{v_{k} \in G^{g}} d(v_{k}, S)
$$

(7)

$$
d(x, S) = \begin{cases} 
0, & \text{if } x \text{ outside } S \\
 w \times |x - y|_2, & \text{if } x \text{ inside } S
\end{cases}
$$

(8)

where $w$ is a constant ($w = 25$ in our experiments), $v_{k}$ is the $k^{th}$ vertex of $G^{g}$ and $y$ is the point closest to $x$ on $S$.

Our garment formulation allows us to freely repose the garment vertices. We can use this to our advantage for applications such as animating clothed virtual avatars, garment re-targeting etc. However, posing is highly non-linear and can lead to undesired artefacts, specially when re-targeting garments across subjects with very different poses.
Since we re-target the garments in unposed space, we reduce distortion by forcing distances from garment vertices to the body to be preserved after unposing:

$$E_{\text{unpose}} = \sum_{g} \sum_{v_k \in G^g} (d(x, S) - d(v_k, S))^2$$

(9)

where \(d(x, S)\) is the \(L_2\) distance between point \(x\) and surface \(S\). \(v_k^0\) and \(S^0\) denote garment vertex and body surface in unposed space, using Eq. 5 and 1 respectively.

**Dressing SMPL** The SMPL model has proven very useful for modelling unclothed shapes. Our idea is to build a wardrobe of digital clothing compatible with SMPL to facilitate semantic garment segmentation [20] to reduce the risk of over-fitting, albeit at the cost of disregarding useful shading signal. For simplicity, let now \(\theta\) denote both the joint angles \(\theta\) and translation \(t\).

The base network, \(f_w\), maps the 2D poses \(J\), and image segmentations \(I\), to per frame latent code \((I_P)\) corresponding to 3D poses

$$I_P = f_w^d(I, J),$$

(10)

and to a common latent code corresponding to body shape \((I_{\beta})\) and garments \((I_G)\) by averaging the per frame codes

$$I_{\beta}, I_G = \frac{1}{F} \sum_{f=0}^{F-1} f_w^{\beta,G}(I_f, J_f).$$

(11)

For each garment class, we train separate branches, \(M_w^g(\cdot)\), to map the latent code \(I_G\) to the unposed garment \(G^g\), which itself is reconstructed from low-frequency PCA coefficients \(z^g\), plus \(D^{hl,g}\) encoding high-frequency displacements

$$M_w^g(I_G, B^g) = G^g = B^g z^g + D^{hl,g}.\quad (12)$$

From the shape and pose latent codes \(I_{\beta}, I_{\theta}\), we predict body shape parameters \(\beta\) and pose \(\theta\) respectively, using a fully connected layer. Using the predicted body shape \(\beta\) and geometry \(M_w^g(I_G, B^g)\) we compute displacements as in Eq. 3

$$D^g = M_w^g(I_G, B^g) - I^g T(\beta, 0_\theta, 0_D).\quad (13)$$

Consequently, the final predicted 3D vertices posed for the \(f^{th}\) frame are obtained with \(C(\beta, \theta_f, D)\), from which we render 2D segmentation masks

$$R_f = R(C(\beta, \theta_f, D), c),\quad (14)$$

where \(R(\cdot)\) is a differentiable renderer [27]. \(R_f\) the rendered semantic segmentation image for frame \(f\), and \(c\) denotes the camera parameters that are assumed fixed while the person moves. The rendering layer in Eq. (14) allows us to compare predictions against the input images. Since MGN predicts body and garments separately, we can predict a semantic segmentation image, leading to a more fine-grained 2D loss, which is not possible using a single mesh surface representation [3]. Note that Eq. (14) allows to train with self-supervision.

**3.3. Loss functions**

The proposed approach can be trained with 3D supervision on vertex coordinates, and with self supervision in the form of 2D segmented images. We use upper-hat for variables that are known and used for supervision during training. We use the following losses to train the network in an end to end fashion:

- **3D vertex loss in the canonical T-pose (\(\theta = 0_\theta\))**:

$$\mathcal{L}_{\theta=0}^{3D} = \|C(\beta, 0_\theta, D) - C(\hat{\beta}, 0_\theta, \hat{D})\|^2,$$

(15)

where, \(0_\theta\) represents zero-vector corresponding to zero pose.
Figure 5: Dressing SMPL with just images. We use MGN to extract garments from the images of a source subject (middle) and use the inferred 3D garments to dress arbitrary human bodies in various poses from SMPL shape subjects. The two sets correspond to male (left) and female (right) body shapes respectively.

- **3D vertex loss in posed space:**
  \[
  L_{3D}^P = \sum_{f=0}^{F-1} ||C(\beta, \theta_f, D) - C(\hat{\beta}, \hat{\theta}_f, \hat{D})||^2
  \]  
  (16)

- **2D segmentation loss:** Unlike [3] we do not optimize silhouette overlap, instead we jointly optimize the projected per-garment segmentation against the input segmentation mask. This ensures that each garment explains its corresponding mask in the image:
  \[
  L_{2D}^{seg} = \sum_{f=0}^{F-1} ||R_f - I_f||^2,
  \]  
  (17)

- **Intermediate losses:** We further impose losses on intermediate pose, shape and garment parameter predictions:
  \[
  L_\theta = \sum_{f=0}^{F-1} ||\theta_f - \hat{\theta}_f||^2, \quad L_\beta = ||\beta - \hat{\beta}||^2, \quad L_z = \sum_{g=0}^{L-1} ||\hat{z}^g - z^g||^2
  \]  
  where \(F\) and \(L\) are the number of images and garments respectively. \(\hat{z}\) are the ground truth PCA garment parameters. While such losses are a bit redundant, they stabilize learning.

### 3.4. Implementation details

**Base Network** \(f_w^b\): We use a CNN to map the input set \(\{I, J\}\) to the body shape, pose and garment latent spaces. It consists of five, 2D convolutions followed by max-pooling layers. Translation invariance, unfortunately, renders CNNs unable to capture the location information of the features. In order to reproduce garment details in 3D, it is important to leverage 2D features as well as their location in the 2D image. To this end, we adopt a strategy similar to [39], where we append the joint coordinates to the output of every CNN layer. We split the last convolutional feature maps into three parts to individuate the body shape, pose and garment information. The three branches are flattened out and we append 2D joint estimates to the pose branch. Three fully connected layers and average pooling on garment and shape latent codes, generate \(l_\beta, l_\theta\) and \(l_G\) respectively. See supplementary for more details.

**Garment Network** \(M_g^w\): We train separate garment networks for each of the garment classes. The garment network consists of two branches. The first predicts the overall mesh shape, and second one adds high frequency details. From the garment latent code \(l_G\), the first branch, consisting of two fully connected layers (sizes=1024, 128), regresses the PCA coefficients. Dot product of these coefficients with the PCA basis generates the base garment mesh. We use the second fully connected branch \(size = m_g\) to regress displacements on top of the mesh predicted in the first branch. We restrict these displacements to \(\leq 1\, \text{cm}\) to ensure that overall shape is explained by the PCA mesh and not these displacements.

### 4. Dataset and Experiments

**Dataset** We use 356 3D scans of people with various body shapes, poses and in diverse clothing. We held out 70 scans for testing and use the rest for training. Similar to [3, 5], we also restrict our setting to the scenario where the person is turning around in front of the camera. We register the scans using multi-mesh registration, SMPL+G. This enables further data augmentation since the registered scans can now be re-posed and re-shaped.

We adopt the data pre-processing steps from [3] including the rendering and segmentation. We also acknowledge the scale ambiguity primarily present between the object size and the distance to the camera. Hence we assume that the subjects in 3D have a fixed height and regress their distance from the camera. Same as [3], we also ignore the effect of camera intrinsics.

**4.1. Experiments**

In this section we discuss the merits of our approach both qualitatively and quantitatively. We also show real world applications in the form of texture transfer (Fig. 7), where
Figure 6: Qualitative comparison with Alldieck et al. [3]. In each set we visualize 3D predictions from [3] (left) and our method (right) for five test subjects. Since our approach explicitly models garment geometry, it preserves more garment details, as is evident from minimal distortions across all the subjects. For more results see supplementary.

Figure 7: Texture transfer. We model each garment class as a mesh with fixed topology and surface parameterization. This enables us to transfer texture from any garment to any other registered instance of the same class. The first column shows the source garment mesh, while the subsequent images show original and transferred garment texture registrations.

We maintain the original geometry of the source garment but map novel texture. We also show garment re-targeting from images using MGN in Fig. 8.

**Qualitative comparisons:** We compare our method against [3] on our scan dataset. For fair comparison we re-train the models proposed by Alldieck et al. [3] on our dataset and compare against our approach (Dataset used by [3] is not publicly available). Figure 6 indicates the advantage of incorporating the garment model in structured prediction over simply modelling free form displacements. Explicit garment modelling allows us to predict sharper garment boundaries and minimize distortions (see Fig. 6). More examples are shown in the supplementary material.

**Quantitative Comparison:** In this experiment we do a quantitative analysis of our approach against the state of the art 3D prediction method, [3]. We compute a symmetric error between the predicted and GT garment surfaces similar to [3]. We report per-garment error, $E^g$ (supplementary), and overall error, i.e. mean of $E^g$ over all the garments

$$E^g = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{1}{|S_i^g|} \sum_{v_k \in S_i^g} d(v_k, S_i^g) + \frac{1}{|S_i|} \sum_{v_k \in S_i} d(v_k, \hat{S}_i^g) \right),$$

where $N$ is the number of meshes with garment $g$. $S_i^g$ and $S_i$ denote the set of vertices and the surface of the $i^{th}$ predicted mesh respectively, belonging to garment $g$. Operator $(\cdot)$ denotes GT values. $d(v_k, S)$ computes the $L_2$ distance between the vertex $v_k$ and surface $S$.

This criterion is slightly different than [3] because we do not evaluate error on the skin parts. We reconstruct the 3D garments with mean vertex-to-surface error of 5.78 mm with 8 frames as input. We re-train octopus [3] on our dataset and the resulting error is 5.72 mm.

We acknowledge the slightly better performance of [3] and attribute it to the fact that the single mesh based approaches do not bind vertices to semantic roles, i.e these approaches can pull vertices from any part of the mesh to explain 3D deformations where as our approach ensures that only semantically correct vertices explain the 3D shape.

It is also worth noting that MGN predicts garments as linear function (PCA coefficients) of latent code, whereas [3] deploys GraphCNN. PCA based formulation though easily tractable is inherently biased towards smooth results. Our work paves the way for further exploration into building garment models for modelling the variations in garment geometry over a fixed topology.

We report the results for using varying number of frames in the supplementary.
GT vs Predicted pose: The 3D vertex predictions are a function of pose and shape. In this experiment we do an ablation study to isolate the effect of errors in pose estimation on vertex predictions. This experiment is important to better understand the strengths and weaknesses of the proposed approach in shape estimation by marginalizing over the errors due to pose fitting. We study two scenarios, first where we predict the 3D pose and second, where we have access to GT pose. We report mean vertex-to-surface error of 5.78mm with GT poses and 11.90mm with our predicted poses.

4.2. Re-targeting

Our multi-mesh representation essentially decouples the underlying body and the garments. This opens up an interesting possibility to take garments from source subject and virtually dress a novel subject. Since the source and the target subjects could be in different poses, we first unpose the source body and garments along with the target body. We drop the (.) notation for the unposed space in the following section for clarity. Below we propose and compare two garment re-targeting approaches. After re-targeting the target body and re-targeted garments are re-posed to their original poses.

Naive re-targeting: The simplest approach to re-target clothes from source to target is to extract the garment offsets, \( \mathbf{D}_{s,g} \) from the source subject using Eq. 13 and dress a target subject using Eq. 5.

Body aware re-targeting: The naive approach is problematic because it relies on non-local pre-set vertex association between the garment and the body \( (I^g) \). This results in inaccurate association between the body blend shapes, \( B_{p,s} \) and the garment vertices. This eventually leads to incorrect estimation of source offsets, \( \mathbf{D}_{s,g} \) and in turn leads to higher inter-penetration between the re-targeted garment and the body (see supplementary). In order to mitigate this issue, we compute the new \( k^{th} \) target garment vertex location, \( v^t_k \) as follows

\[
v^t_k = v^s_k - S^s_{I_k} + S^t_{I_k}
\]

where \( v^s_k \) is the source garment vertex, \( S^s_{I_k} \) is the vertex (indexed by \( I_k \)) among the source body vertices, \( S^t \), closest to \( v^s_k \) and \( S^t_{I_k} \) is the corresponding vertex among the target body vertices.

MGN allows us to predict separable body shape and garments in 3D, allowing us to do garment re-targeting (as described above) using just images. To the best of our knowledge this is the first method to do so. See Fig. 8 for results on garment re-targeting by MGN. See supplementary for more results.

5. Conclusion and Future Works

We introduce MGN, the first model capable of jointly reconstructing from few images, body shape and garment geometry as layered meshes. Experiments demonstrate that this representation has several benefits: it is closer to how clothing layers on top of the body in the real world, which allows control such as re-dressing novel shapes with the reconstructed clothing. Additionally, we introduce for the first time, a dataset of registered real garments from real scans obtained with a robust registration pipeline. When compared to more classical single mesh representations, it allows more control and qualitatively the results are very similar. In summary, we think that MGN provides a first step in a promising research direction. We will release the MGN model and the digital wardrobe to stimulate research in this direction. Further discussion on limitations and future works in supplementary.

Acknowledgements This work is partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans) and Google Faculty Research Award. We thank twindom (https://web.twindom.com/) for providing scan data, Thiemo Alldieck for providing code for texture/segmentation stitching, and Verica Lazova for discussions.
References

[1] https://virtualhumans.mpi-inf.mpg.de/mgn/. 1
[27] Paul Henderson and Vittorio Ferrari. Learning to generate and reconstruct 3d meshes with only 2d supervision. In British Machine Vision Conference (BMVC), 2018. 2


