Dynamic Surface Function Networks for Clothed Human Bodies

Andrei Burov
Matthias Nießner
Justus Thies

1Technical University of Munich
2Max Planck Institute for Intelligent Systems, Tübingen

Abstract

We present a novel method for temporal coherent reconstruction and tracking of clothed humans. Given a monocular RGB-D sequence, we learn a person-specific body model which is based on a dynamic surface function network. To this end, we explicitly model the surface of the person using a multi-layer perceptron (MLP) which is embedded into the canonical space of the SMPL body model. With classical forward rendering, the represented surface can be rasterized using the topology of a template mesh. For each surface point of the template mesh, the MLP is evaluated to predict the actual surface location. To handle pose-dependent deformations, the MLP is conditioned on the SMPL pose parameters. We show that this surface representation as well as the pose parameters can be learned in a self-supervised fashion using the principle of analysis-by-synthesis and differentiable rasterization. As a result, we are able to reconstruct a temporally coherent mesh sequence from the input data. The underlying surface representation can be used to synthesize new animations of the reconstructed person including pose-dependent deformations.

1. Introduction

Digital capture of human bodies is a rapidly growing research area in computer vision and computer graphics. There are many high impact applications, in particular, in the field of telepresencing and man-machine interaction that rely on the reconstruction of the surface as well as the motion of a person. For instance, for telepresencing in virtual reality (VR) or augmented reality (AR), the ultimate goal is to photo-realistically re-render the people who want to interact with each other. A key challenge is to capture and reproduce the natural motions, including the dynamically changing surface (especially clothing). In our work, we present a novel representation for the surface of a clothed human called dynamic surface function networks which captures pose-dependent surface deformations such as wrinkles of the clothing. In contrast to recent works on implicit representations [11, 6, 23], our representation explicitly models the surface of the human. Both representations have their advantages and disadvantages. We use an explicit surface representation to leverage fast forward rendering (exploiting the rasterization units of a GPU), and global surface correspondences between different frames. Specifically, our dynamic surface function network is attached to the surface of the SMPL [28] body model which gives us access to the
kinematic chain, as well as to a topology for rendering. In particular, our model represents a continuous offset surface that can be evaluated at arbitrary points of the SMPL surface (also within triangles). However, note that our representation is agnostic to the underlying parametric model and can also be used for other dynamically changing surfaces, e.g., human faces. The dynamic surface function network is trained in a person-specific fashion. To this end, we assume a short sequence (few seconds) of the person moving in front of a single consumer-level RGB-D camera. Note that we do not assume any specific motion sequences (e.g., standing in T-pose or alike). We jointly optimize the surface representation network as well as the pose parameters of the underlying SMPL body model. This global optimization strategy allows us to fuse all captured data into a consistent surface representation. This person-specific surface model can be animated explicitly using the joint control handles of the SMPL model, thus, allowing pose transfer (see Fig. 1).

To summarize, we present a method that allows reconstruction of a person-specific controllable body model based on a monocular input sequence captured by a commodity RGB-D sensor. Our key contributions are:

- an explicit surface representation network which is able to model the pose-dependent deformations of the surface, such as wrinkles of the clothing.
- a global analysis-by-synthesis formulation that allows for the joint optimization of the pose and the surface over the entire sequence, leading to a temporally consistent tracking of the person in the input sequence.

2. Related Work

Our work is based on an explicit surface representation. In the literature, explicit surface representations (especially, triangle meshes) are the most prominent representations for human faces and bodies [28, 44, 69]. However, implicit representations are also used to represent the surface of a body including the clothing [6, 11, 13, 23, 49, 50]. Implicit surface representations have the advantage of not needing special care for handling topological changes (as would be required for an explicit representation like a mesh). On the other hand, implicit representations do not provide explicit correspondences over a time series. For instance, methods such as Occupancy Flow [40] predict scene flow, but are limited by the sequence length. Methods like [16, 15, 58, 8] reconstruct implicit function of an object’s geometry in a patch-based manner and empirically show consistency of patches across a sequence. PIFu [49] and PIFuHD [50] are able to reconstruct a pixel-aligned implicit function only based on RGB input images. They leverage a learned prior trained on a synthetic dataset of humans. Similarly, IF-Nets [11] learn a prior to reconstruct an implicit function of a human, based on pointcloud or depth-map inputs. IF-Nets have been extended in a follow-up work called IP-Nets [6] to fit an explicit surface to the reconstructed implicit surface (SMPL model + displacements).

Aside from these per-frame reconstruction methods, there exist methods that incrementally fuse observations into a discretized implicit function (volumetric SDF grid) [12]. The seminal work of DynamicFusion by Newcombe et al. [39] is able to reconstruct dynamically changing objects only based on a depth sequence. Follow-up works [24, 18, 14, 19] added additional color constraints or dense SDF alignment [53, 54]. BodyFusion [64] and DoubleFusion [65] added a deformation prior using a human skeleton. In contrast to these methods, our approach reconstructs a controllable mesh including pose-dependent deformations, allowing us to animate the reconstructed mesh. Note that these fusion methods work on a frame-to-frame tracking scheme.

Our method is an optimization-based approach that jointly optimizes the surface over the entire input sequence and does not require a learned prior as [6, 11, 49, 50]. The global optimization scheme of our method is closely related to MonoClothCap [63]. MonoClothCap gets an RGB sequence as input to track, and reconstructs a clothed human using an explicit surface representation. In order to handle clothing, they rely on a PCA model and shading-based refinement of the mesh. Another technique to model the surface of clothing explicitly has been shown in CAPE [35]. They learn a variational auto-encoder to represent offsets for clothing relative to the SMPL model based on high quality multi-view reconstructions using ClothCap [46]. This learned prior allows them to reconstruct clothed humans from single RGB inputs only, including pose-dependent deformations. Alldieck et al. [5] reconstruct a detailed human body shape based on a single RGB input leveraging a learned aligned image-to-image translation technique to regress texture maps for the geometry and color of the model. In contrast, our approach is not based on a learned prior and optimizes the surface based on the observations from an RGB-D camera. Our dynamic surface function network represents the pose-dependent geometry and is not restricted to a static surface [7, 61, 68, 22]. Methods that estimate the body shape under cloth [66, 62, 55, 38, 47] can be used as an initialization of our method. Our approach works with a consumer-grade RGB-D camera, and does not require a multi-camera setup to learn or reconstruct a person-specific template in advance [20, 21]. We use RGB-D data to explore the representation power of our surface network, but we see the potential of our representation to be used for video-based reconstruction similar to [4, 3, 2]. Note that our method is not designed for real-time use such as LiveCap [20], since our focus lies on the global reconstruction of a controllable mesh with a pose-dependent sur-
face. Concurrent work analyzes other representations of a human surface that are based on points [32], local surface patches [31], or implicit representations for shape and motions [42, 36, 59, 51, 60, 27, 43], including [10, 45, 56, 41] which are based on Neural Radiance Fields [37].

3. Method

We formulate the reconstruction of the human body as an energy optimization problem, globally, across the entire input sequence. To model deformning clothing, we use an explicit offset surface that is embedded on the SMPL surface. This offset surface is represented as a multi-layer perceptron (MLP). For an overview of our method refer to Fig. 2.

3.1. Model Definition

Our method is built upon the SMPL body model [28] and extends it to handle non-rigid offsets represented as a multi-layer perceptron (MLP).

Constrained SMPL model The SMPL model is parametrized by the shape parameters $\beta \in \mathbb{R}^{10}$ and the joint parameters $\delta \in \mathbb{R}^{72}$. We use a hard constraint to enforce the parameters to stay in a predefined range. That is, we use a differentiable mapping function $\delta(x)$ that maps our unconstrained joint parameters $x \in \mathbb{R}^{72}$, in an element-wise fashion, to a fixed range of the SMPL parameters:

$$\delta(x) = \delta_{\min} + \frac{\tanh(x) + 1}{2} \cdot (\delta_{\max} - \delta_{\min}).$$

Note that we use axis-angle representation. Thus, we constrain the angle-scaled axis to be in a min-max bounding box. We compute $\delta_{\min}$ and $\delta_{\max}$ from the MOSH dataset [29]. In our experiments, this hard constraint was effective in stabilizing the tracking without the need of any additional regularization, e.g., using VPoser [44] (especially, for scenarios with more occlusions) or pose-conditioned joint angle limits [1].

Dynamic Offset Surface To model the dynamic offset surface on top of the SMPL model, we employ an MLP which is conditioned on the reparameterized SMPL parameters $x$. Specifically, we embed the offsets $O_v(v, x)$ in the canonical pose of the SMPL template surface $T$. The offset of the surface point $v$, given the pose parameters $x$ is computed using a 8-layer ReLU-MLP with 256 feature channels per intermediate layer ($\Theta$ being the learnable parameters of the MLP). We apply positional encoding [37] to the input point $v$ before feeding it into the MLP (using 10 frequencies). The pose conditioning is concatenated to the point input in matrix form. Note that we do not use the pose of the root joint into consideration. This representation of the pose conditioning is similar to the linear pose-dependent corrections of the SMPL model. Note that the 3D space allows us to have a continuous input to the MLP without the need of handling texture seams, distortions or alike. To rasterize the surface represented by the MLP, we sample the surface at point using a subdivided SMPL topology.

Model Given the constrained SMPL model and the dynamic offset model, we represent the vertices of a body as:

$$V(\beta, x, \Theta) = LBS(V_{SMPL}(\beta, \delta(x)) + O_v(T, x), \delta(x))$$

$LBS(V', \delta)$ is the linear blend skinning function that applies the rotations defined by $\delta$ to the 'unposed' vertices $V' = V_{SMPL} + O$. $V_{SMPL}(\beta, \delta(x))$ computes the unposed SMPL model surface based on the shape PCA and the pose-dependent corrective shape (see SMPL [28]). In our experiments, we exclude the hands and feet from the optimization (see supplemental material).

3.2. Fitting Energy Formulation

Given an RGB-D video sequence with $N + 1$ frames, we minimize the following global energy:

$$E_{Seq}(P) = \sum_{i=0}^{N} E_i(P) + \sum_{i=1}^{N-1} E_{\Pi}(P) + \sum_{i=0}^{N} \sum_{j=0}^{N} E_C^{ij}(P)$$

This energy formulation considers per-frame energies $E_i$, temporal energies $E_{\Pi}$, and pair-wise surface consistency constraints $E_C^{ij}$. $P$ is the set of unknowns; namely the MLP weights $\Theta$, the shape $\beta$ and the per frame pose parameters $x_i$.

Per-frame Energy The per-frame energy is based on data-terms using the color frame $C_i$ and depth frame. Given the intrinsics of the depth camera, we back-project the depth maps into the camera space using the pinhole camera model $\Pi$, resulting in 3D locations per pixel which we call $D_i$. Based on the color frame $C_i$, we estimate the 2D joint positions $J_i^{OP}$ using OpenPose [9] and dense correspondences $M_i^{DP}$ using DensePose [48]. In addition, we use Graphonomy [17] to estimate the silhouette $S_i^{ij}$ of the person in the input image. With these inputs, we define the per-frame energy as:

$$E_i(P) = w_{OP} \cdot E_{OpenPose}(P) + w_{DP} \cdot E_{DensePose}(P) + w_{Proj} \cdot E_{Projective}(P) + w_{Sil} \cdot E_{Silhouette}(P) + w_{Reg} \cdot E_{Reg}(P)$$

The detected 2D joint locations are used as a sparse energy term:

$$E_{OpenPose}^{ij}(P) = 1/K_J \cdot |J_i^{OP} - \Pi(J_i^{ij})|$$

where $K_J = 25$ is the number of joints and $J_i^{ij}$ are the corresponding regresssed joints of the SMPL model. The per-pixel DensePose energy term $E_{DensePose}^{ij}(P)$ is defined as:

$$E_{DensePose}^{ij}(P) = \sum_{(p,c) \in M_i^{DP}} \frac{P^2 P(V_c, D_i(p))}{K_{DP}}$$
$K_{DP}$ is the number of valid correspondences $M_{DP}$ estimated by DensePose ($p$ is the pixel and $c$ the correspondence on the SMPL surface). $P2P(V_c, xyz)$ measures the $\ell_1$ point-to-point distance from the observation point $V_c$ to the surface point of the model $V_c = \text{Sample}(c, V(\beta, x, \Theta))$. The function $\text{Sample}(c, V(\beta, x, \Theta))$ computes the surface point based on the vertex indices $(i_0, i_1, i_2)$ and barycentric coordinates $(b_0, b_1, b_2)$ provided by the correspondence $c = (i_0, i_1, i_2, b_0, b_1, b_2)$. In addition to the DensePose term, we use a dense data term using projective correspondences $M_{Proj}$. The projective energy term is defined as:

$$E_{\text{proj}}(P) = \sum_{(p,c)\in M_{Proj}} P2P(V_c, D_i(p)) + N2N(V_c, D_i(p))$$

$N2N(V_c, D_i(p))$ measures the cosine similarity of the normals from the observation point $D_i(p)$ to the surface point $V_c$. $P2N(V_c, D_i(p))$ measures point to plane distance between the observation point $D_i(p)$ and the tangent plane at the source point $V_c$.

The per-frame silhouette energy term $E_{\text{silhouette}}(P)$ computes the difference between the mask predictions from [17] and the rendered subject’s silhouette $S_i(P)$. It is defined as:

$$E_{\text{silhouette}}(P) = |S^C_i - S_i(P)|$$

Note that our differentiable rasterizer explicitly applies edge sampling at the boundary of the silhouette of the surface to compute gradients.

To optimize for a smooth surface, we use the regularizer $E_{\text{Reg}}(P)$. Based on the topology of the template surface $T$, we apply a Laplacian regularizer:

$$E_{\text{Reg}}(P) = \sum_{m \in T} \left| V_m - \frac{\sum_{n \in \text{neigh}(m)} V_n}{|\text{neigh}(m)|} \right|^2$$

Here, $\text{neigh}(m)$ denotes the 1-ring neighborhood of the $m$-th vertex of the template topology.

**Temporal Energy** The temporal regularizer $E_{t}^T(P)$ is defined as:

$$E_{t}^T(P) = w_{t}^{rot} \cdot E_{\text{rot}}^T(P) + w_{t}^{surf} \cdot E_{\text{surf}}^T(P)$$

We use a Laplacian regularizer on the surface of the model in the temporal domain:

$$E_{\text{surf}}^T(P) = |V_i - (V_{i-1} + V_{i+1})/2|^2$$

In addition, we apply a temporal regularizer on the joint rotation matrices $R_i = R(\delta(x_i))$:

$$E_{\text{rot}}^T(P) = |R_i - R_{i-1}|^2 + |R_i - R_{i+1}|^2$$

**Pair-wise Surface Consistency** In each optimization step for each frame $i$, we randomly select another frame $j$ from the sequence to measure surface consistencies. Specifically, we measure the difference of the offset surfaces:

$$E_{c}^{ij}(P) = w_{c} \cdot \omega(x_i, x_j) |V_j \cdot (O_\Theta(T, x_i) - O_\Theta(T, x_j))|$$

$$\omega(x_i, x_j) = \exp(-|R(\delta(x_i)) - R(\delta(x_j))|^2)$$

measures the similarity of the poses in the two frames (excluding the root joint pose) and $V_j$ denotes the surface visibility in frame $j$. 

Figure 2: Given an RGB-D input sequence, we jointly optimize for the SMPL body parameters and a pose-conditioned offset surface function (Dynamic Surface Function Network) which is represented by an MLP. Specifically, the MLP gets a positionally encoded surface point of the canonical template surface and the pose parameters as input and predicts the offset surface location. The pose parameters and the weights of the network are optimized based on the optimization of a global analysis-by-synthesis energy formulation (green arrow).
3.3. Optimization Scheme

The energy formulation in Eq. 3.2 is optimized in two stages. Specifically, we first minimize the energy only considering the SMPL body parameters for shape and pose to get a good initial estimate. We use L-BFGS [26] to optimize for these shape and pose parameters in sequential order (we initialize the parameters with the parameters of the previous frame). Using this initial fitting, we optimize for all parameters $P$ in a joint optimization using ADAM [25] with random sampling of the input images. We refer to the supplemental material for the used hyperparameters.

4. Results

In this section, we evaluate our method on synthetic as well as on real-world data. For quantitative evaluations, we use synthetic sequences based on the BUFF [67] dataset. Real data inputs are captured with a Microsoft Kinect Azure at a depth resolution of 640x576 and color image resolution of 1080p at 15 fps (each recording is 200 frames long). We align color images with depth maps using OpenCV and apply distortion correction. These sequences are used to demonstrate the applicability of our method to real-world data, as well as to show qualitative comparisons. Specifically in Fig. 8, we show reconstructions of different persons in varying clothing. As can be seen, we faithfully reproduce the input data including the pose-dependent deformations of the surface. The temporal coherence of our reconstructions can be seen in the supplemental video. In Fig. 1, we show results for novel poses using our representation based on poses from the MOSH dataset [29].

4.1. Ablation Studies

The key components of our approach are the temporally consistent tracking of the SMPL body as well as the reconstruction of the surface using a pose-dependent MLP. In the following, we will analyze the effects of our optimization scheme as well as the dynamic surface MLP.

Static vs. Dynamic Surface MLP To handle pose dependent deformation of the geometry such as wrinkles of the clothing, we use a dynamic surface function network. This network is conditioned on the pose parameters of the underlying SMPL model. In Fig. 3, we show a comparison of using a dynamic surface function network and a static one (i.e., without providing the pose conditioning to the network). As can be seen, only the pose conditioned network is able to represent the dynamically deforming surface.

Pair-wise Surface Consistency The pair-wise consistency regularizes the surface network to predict similar surfaces if the pose is similar. In Fig. 4, we show the effect of this regularizer. As can be seen, the pair-wise surface consistency loss leads to more details as well as to a 3D consistent surface reconstruction. Without the pair-wise loss, the surface tends to be closer to the SMPL surface.

Joint Optimization Our method is based on a joint global optimization, i.e., the MLP is trained in conjunction to the SMPL model parameters. In Tab. 1, we list the quantitative evaluation for an ablation study that uses sequential optimization (first optimization of the SMPL parameters, and then training of the MLP). As can be seen, joint optimization is crucial for the reconstruction quality.
4.2. Comparisons

Implicit Surface Representations  Our dynamic surface function networks estimate the surface points explicitly. A classical rasterizer can be used to render these surfaces. In contrast, implicit surface representations (like an occupancy function) need to be ray-casted or converted into an explicit surface representation [30]. Since implicit functions like IF-Nets [11] do not provide explicit correspondences over a time series, regularizing the surface to deform in a smooth and temporally consistent manner is non-trivial (regions can disappear and appear at a different location). In Fig. 5, we show a comparison to IF-Nets and IP-Nets on an RGB-D sequence. IP-Net [6] is a hybrid, leveraging the reconstruction abilities of IF-Nets and the controllability of the SMPL body model. Both IF-Nets and IP-Nets take a point cloud of a single frame as input and predict the implicit surface. We use the masks predicted by Graphonomy to remove the background points from the input. This input is fed into the pretrained networks for single views provided by the authors as shown in their publications. Note that the authors explicitly state that the networks generalize to continuous articulations of temporal data from new data sources. In Tab. 1, we show the corresponding errors w.r.t. IoU, chamfer-distance and normal consistency. We observe that our method leads to better reconstructions, with a temporally consistent output mesh (see supplemental video).

BodyFusion [64] is a fusion method that integrates the depth measurements into a volumetric SDF representation. It leverages the skeleton of a body as a kinematic prior, thus, outperforming general non-rigid fusion methods like DynamicFusion [39] or VolumeDeform [24] in the scenario of body reconstruction. In Tab. 2, we quantitatively compare our method to BodyFusion as well as to IP-Nets. Note that both BodyFusion as well as our method are using temporal information, while IP-Net is applied frame-by-frame, thus, leading to the highest tracking error. The tracking error is measured by a mean $\ell_1$ end-point error on the dataset provided by the authors of BodyFusion [64].

<table>
<thead>
<tr>
<th>Method</th>
<th>IoU ↑</th>
<th>C-$\ell_2$ ↓</th>
<th>NC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF-Nets [11]</td>
<td>0.818</td>
<td>1.8cm</td>
<td>0.903</td>
</tr>
<tr>
<td>IP-Nets [6]</td>
<td>0.783</td>
<td>2.1cm</td>
<td>0.861</td>
</tr>
<tr>
<td>CAPE [34]</td>
<td>0.648</td>
<td>2.5cm</td>
<td>0.844</td>
</tr>
<tr>
<td>Ours w/o joint optimization</td>
<td>0.687</td>
<td>2.4cm</td>
<td>0.877</td>
</tr>
<tr>
<td>Ours</td>
<td>0.832</td>
<td>1.6cm</td>
<td>0.916</td>
</tr>
</tbody>
</table>

Table 1: Quantitative comparisons based on a sequence of the BUFF dataset [67] (see Fig. 5). For all methods, we provide synthetically rendered monocular RGB-D data inputs. In the table, we report the numbers for IoU, chamfer distance (using $\ell_2$-norm) and normal consistency w.r.t. the complete meshes from the dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>EPE ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>BodyFusion [64]</td>
<td>2.77cm</td>
</tr>
<tr>
<td>IP-Nets [6]</td>
<td>14.17cm*</td>
</tr>
<tr>
<td>CAPE [34]</td>
<td>5.51cm</td>
</tr>
<tr>
<td>SMPL</td>
<td>4.33cm</td>
</tr>
<tr>
<td>Ours</td>
<td>2.63cm</td>
</tr>
</tbody>
</table>

Table 2: Quantitative comparisons based on the BodyFusion dataset [64] which provides tracked VICON marker locations for each RGB-D frame. The mean end-point-error (EPE) is measured in 3D based on an $\ell_1$ distance between the tracked VICON markers and the corresponding points on the reconstructed mesh. Note that both the CAPE baseline and our method use the SMPL tracking as initialization. * not using temporal information.
low-quality depth and color images, the dataset provides the coordinates of tracked VICON markers used for evaluation (see Fig. 6). Our method results in the lowest error and qualitatively results in sharper results.

**Explicit Surface Representations** In the results discussed above, we already mention IP-Nets [6] which model the surface explicitly with per-vertex displacements on top of the SMPL model, also known as SMPL-D. Our approach with an MLP without pose conditioning is closely related to SMPL-D (see Fig. 3). Instead of a discrete number of displacement vectors, we represent the surface using a continuous function that can be evaluated on any surface point of the SMPL model. In Tab. 1, we also include a comparison to the graph neural network-based GAN for clothing CAPE [34]. We used our optimization framework to estimate the latent codes, based on the RGB-D inputs (see supplemental material for more details). While the generative approach has several advantages for reconstructing humans from partial views or RGB images, it does not capture the detail that our reconstruction method recovers.

5. Discussion

Our Dynamic Surface Function Networks are able to reconstruct and track a variety of sequences of different people. The explicit representation of the surface allows us to use differentiable rasterization for rendering, and to generate a temporally consistent mesh as well as global correspondences, but it comes with the limitation of a fixed topology, and, thus, is not able to represent topologically changing surfaces.

Our approach is an optimization approach and belongs to the class of methods that must be trained for each new input sequence [37, 57, 52]. In particular, for the reconstruction of a controllable representation this retraining of the dynamic surface function network is practical and crucial to get fine scale details such as wrinkles. If regions of the body are not visible in the input sequence (i.e., no data terms), the surface in these regions is only regularized to be smooth (see Fig. 7). The assumption of the pair-wise consistency loss (similar clothing geometry for similar poses) might not always hold for loose garment. Extending our approach to such apparel is an interesting direction for future work. Further improvements on wrinkle expressiveness might be achieved by leveraging the dense color input.

6. Conclusion

Our approach reconstructs and tracks a clothed human using a single commodity RGB-D sensor, obtaining a temporally-consistent surface reconstruction. The underlying dynamic surface function network is able to represent pose-dependent deformations of the surface and allows to re-animate the body. In our experiments, we demonstrate the effectiveness of this representation and show state-of-the-art reconstruction performance. The proposed representation offers a variety of advantages such as a consistent mesh structure, dynamically changing surface and the possibility to extend it to other attributes (see supplemental material).

Acknowledgements

The work is funded by Huawei, a TUM-IAS Rudolf Mößbauer Fellowship, the ERC Starting Grant Scan2CAD (804724), the German Research Foundation (DFG) Grant Making Machine Learning on Static and Dynamic 3D Data Practical, and the BMBF-funded Munich Center for Machine Learning.
Figure 8: Temporally coherent reconstructions of clothed humans using our proposed dynamic surface function networks. The input data has been captured by a Microsoft Kinect Azure. Each sequence is 200 frames long.
References


10762


[32] Qianli Ma, Jinlong Yang, , Siyu Tang, and Michael J Black. The power of points for modeling humans in clothing. In International Conference on Computer Vision (ICCV), 2021. 3


[34] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to Dress 3D People in Generative Clothing. In Computer Vision and Pattern Recognition (CVPR), 2020. 6, 7


