ETH_zürich_

Neural Point-based Shape Modeling of Humans in Challenging Clothing

Neural Point-based Shape Modeling of Humans in Challenging Clothing

Qianli Ma[1,2]   Jinlong Yang[2]   Michael J. Black[2]   Siyu Tang[1]

[1]ETH Zürich   [2]Max Planck Institute for Intelligent Systems, Tübingen, Germany

{qianli.ma, siyu.tang}@inf.ethz.ch, {qma,jyang,black}@tuebingen.mpg.de

(a) SCANimate [38]   (b) POP [24]   (c) Ours   (d) Predicted LBS Weights   (e) Ours, animated
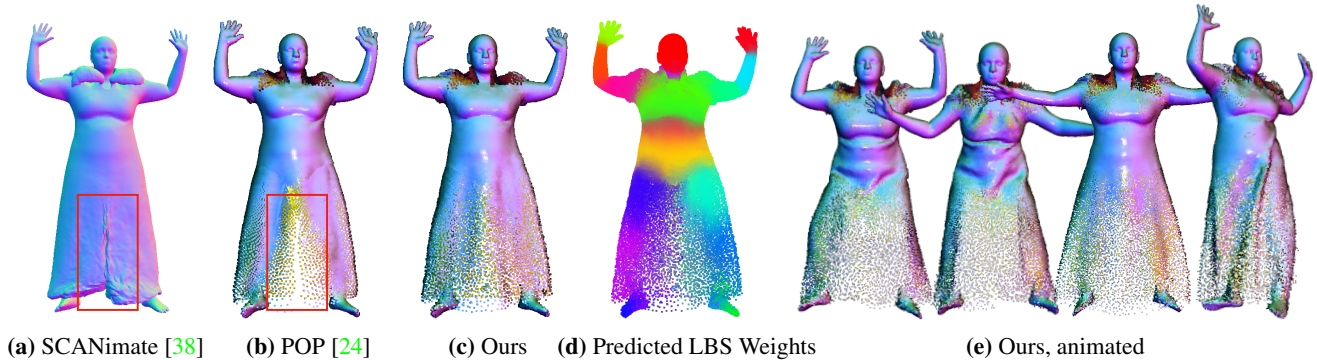
Figure 1: **Learning clothed avatars with SkiRT (Ours)**. When applied to challenging clothing types such as skirts and dresses, existing methods for learning clothed human shape often **(a)** erroneously generate pant-like structures [38], or **(b)** suffer from inhomogeneous point density and split-like artifacts in the predicted shape [24]. Our point-based approach **(c)** addresses these issues by predicting the LBS weights relating the clothed surface to the body **(d)** and using a novel coarse shape representation in the posed space. SkiRT achieves state-of-the-art modeling accuracy for pose-dependent shapes of humans in diverse types of clothing. Results from the point-based methods here are rendered using a plain point renderer to highlight the difference between each method's "raw" outputs.

Abstract

Parametric 3D body models like SMPL only represent minimally-clothed people and are hard to extend to clothing because they have a fixed mesh topology and resolution. To address these limitations, recent work uses implicit surfaces or point clouds to model clothed bodies. While not limited by topology, such methods still struggle to model clothing that deviates significantly from the body, such as skirts and dresses. This is because they rely on the body to canonicalize the clothed surface by reposing it to a reference shape. Unfortunately, this process is poorly defined when clothing is far from the body. Additionally, they use linear blend skinning to pose the body and the skinning weights are tied to the underlying body parts. In contrast, we model the clothing deformation in a local coordinate space without canonicalization. We also relax the skinning weights to let multiple body parts influence the surface. Specifically, we extend point-based methods with a coarse stage, that replaces canonicalization with a learned pose-independent "coarse shape" that can capture the rough surface geometry of clothing like skirts. We then refine this using a network that infers the linear blend skinning weights and pose dependent displacements from the coarse representation. The approach works well for garments that both conform to, and deviate from, the body. We demonstrate the usefulness of our approach by learning person-specific avatars from examples and then show how they can be animated in new poses and motions. We also show that the method can learn directly from raw scans with missing data, greatly simplifying the process of creating realistic avatars. Code is available for research purposes at `https://qianlim.github.io/SkiRT`.

1. Introduction

A wide range of applications are driving recent interest in creating animatable 3D human avatars from data. Given a set of 3D scans or meshes of a person performing various poses, the goal is to create a model of the clothed body that can be controlled by the body pose, with the clothing deforming naturally as the body moves. Such technology will greatly simplify the creation of avatars for immersive experiences in teleconferences, gaming, and virtual try-on, to name a few. Traditional solutions to this task typically require an artist-designed clothing template, a step that registers the template to data, and a process to associate the

registered data to a controllable bone structure. These all involve expert knowledge and manual effort.

While registering a mesh template to a clothed body scan is challenging and error-prone, there are tractable solutions to estimate the minimally-clothed body pose and shape that lies under clothing [3, 4, 32, 46, 54], i.e., to fit a parametric human body model [17, 21, 33, 51] to clothed scans. This gives rise to a promising alternative to the complex traditional pipeline. Recent methods build shape models of *clothed* humans using clothed scans and the corresponding minimally-clothed bodies, thereby side-stepping the need for per-garment mesh templates [2, 31, 39].

To go beyond minimally-clothed body models such as SMPL [21] to capture the complex shape of clothed bodies, recent attempts either displace the vertices of the body template mesh [5, 23, 26, 41] or learn a neural implicit field (without relying on templates) that captures the clothing shape conditioned on the body pose parameters [7, 38, 42, 47]. While these methods show promising results on T-shirts, pants, or even closed jackets, they exhibit limitations for clothing types that are topologically different from the body, like dresses and skirts. To date, these remain a major challenge for both tempate-based and template-free models of clothed humans.

One of the key problems of these existing formulations is that they require explicit *canonicalization*. Some work assumes the canonicalized data is given [23, 27], whereas the recent trend is to learn a backward transformation field to "unpose" the data to a consistent canonical pose [7, 38, 47]. On a high-level, canonicalization is a process that transforms all points on the posed, clothed body surface to a pre-defined pose, using the bone transformations derived from the underlying body. Here, a hidden assumption is that the unposing of clothing can be driven simply by the unposing of the minimally-clothed body. While this is true for many clothing types that lie close to the body (such as T-shirts and pants), the assumption often fails for garments that differ greatly from the body shape and topology, such as skirts and dresses. The canonicalization step for a skirt is ill-defined, and canonicalizing such surfaces with existing body-based approaches often leads to artifacts, as we demonstrate in Sec. 3; see Fig. 2.

In contrast to canonicalization methods, shape deformation can be modeled in local, posed, coordinates; e.g., recent point-based clothed human models [22, 24] directly model shape in the *posed* space. For a query point on the body surface, a clothing displacement vector is predicted in a pre-defined local coordinate system anchored to that point. These methods represent the pose-dependent shapes of skirts by simple offsets from the unclothed body model. However, as the definition of the local coordinate frame relies on the linear blend skinning (LBS) weights of SMPL (Sec. 3), these methods are tied to body part transforma-

tions, resulting in clear surface discontinuities on the skirts and dresses; see Fig. 1.

Here, we follow the point-based approach but further equip it with a few key innovations that combine the traditional pipeline with implicit surface-based body modeling techniques. Specifically, we extend the point-based method, POP [24], with a learned, pose-*in*dependent "coarse shape" that captures the rough surface geometry of clothing. The coarse shape serves a similar role as the clothing template in the traditional pipeline. Instead of using a manually crafted mesh, however, our point-based coarse geometry is learned from data, is not tied to specific mesh topology, and is hence flexible in representing various clothing types. We refine this coarse shape using a neural network that predicts pose-dependent displacements from the coarse representation, in their respective local coordinates. Differing from prior work [24], the transformations from the local to the world coordinates are not taken from the underlying body model, but *predicted* by a neural network. Additionally, we introduce an adaptive regularization technique to train the pipeline, which encourages a more uniform spatial distribution of the generated point clouds. As we show in the experiments, these innovations lead to improved modeling accuracy and, more importantly, are essential to solving the typical surface discontinuity problems found in previous work.

These key components enable our new method, SkiRT (*Skinned Refined Template-free*), which creates pose-dependent shape models of clothed humans from 3D scans. We evaluate SkiRT on diverse clothing types, with a special focus on skirts and dresses of different lengths, tightness and styles; of course, SkiRT also models simpler clothing types. SkiRT tackles the long-standing difficulty of modeling challenging clothing types such as skirts and dresses, outperforming state-of-the-art methods from both the point-based and implicit surface-based families on the pose-dependent clothing modeling task. Furthermore, with an extensive ablation study, we carefully analyze the key components in our pipeline, providing insights for future point-based models of clothed human body geometry.

2. Related Work

In this section we review previous work that captures and models 3D clothing with data-driven learning approaches.

Template-based clothed human modeling. Existing graphics software and, in particular, 3D clothing simulation, uses 3D meshes to represent human bodies and clothing. Physics-based cloth simulation has been widely used to train data-driven clothing models [2, 11, 16, 31, 39, 40, 44, 52], and mesh-based models of minimally-clothed humans are common [1, 21, 33, 51]. Mesh-based body models have been a foundation for capturing and representing clothing deviation from the the body [14, 18, 23, 26, 34, 45, 53, 55].

While deep learning methods can be used to model clothing deviation on meshes [13, 35, 43], these methods remain limited when it comes to skirts and dresses. Unlike many pants and tops, skirts and dresses differ greatly from the body surface, making it infeasible to deform a common body template mesh to fit the data. To capture skirts and dresses, some methods exploit customized template meshes [14, 15, 34, 45, 55, 56]. With such an approach, one must manually design templates for a wide variety of clothing styles, making this approach impractical. In contrast, our method is "template-free" in that it replaces the per-garment mesh templates with data-driven coarse shapes that serve a similar role. This obviates the expensive manual labor and can, in principle, model arbitrary skirt styles.

Template-free clothed human modeling. To deal with the problems of template-based approaches, recent work goes beyond meshes and exploits new representations that do not require a per-garment template. A popular and promising approach uses deep implicit shapes [8, 10, 25, 30] to capture [3, 12, 36, 37, 48–50] and model clothed humans [6, 7, 9, 27, 28, 38, 42, 47]. Although such representations adapt to varied topology, these methods [7, 27, 38, 47] do not perform well when modeling skirts, especially wide and long skirts. The problem originates from the need to canonicalize the training data as detailed in Sec. 3. In contrast, point clouds provide an alternative representation. Ma et al. [22] successfully model clothing or clothed humans with point clouds, and further demonstrate in [24] that their approach can be extended to train a single model of various clothing styles. While promising, dense point clouds can have trouble with skirts, producing a discontinuous representation between the legs. The reason is that these methods are still tied to the topology of the unclothed body, as we analyze in Sec. 3. Our work uses a "coarse stage" with a local coordinate space to avoid the issues in previous work. It also uses a network to predict more reasonable skinning weights for the skirt points, minimizing the discontinuities seen with previous work. In a similar spirit, concurrent work [19] also enhances the point-based clothing representation with a coarse template; but [19] uses diffusion to acquire clothing skinning weights, whereas our method learns the skinning weights from data.

3. Analyzing the Challenges of Skirts

Given a clothed scan with known body pose, many methods transform the clothed scan to a common canonical pose. This helps factor out the articulation-related deformations in the clothed body shape, reducing data variation and facilitating model learning. A premise for canonicalization to work is that the clothing, when dressed on the body, behaves similarly to the unclothed body under different poses. Many tops and pants mostly satisfy this assumption, as evidenced
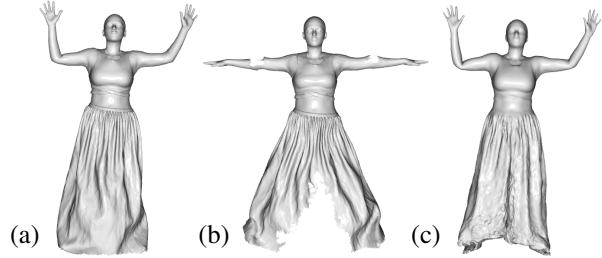

Figure 2: Typical failure case due to canonicalization of skirts using the learned inverse-LBS methods. Result produced by [38] on a training example. Given a raw scan (a), canonicalization often tears or creases the skirt (b), leading to areas in canonical space without good supervision. A neural implicit function trained with such canonicalized shapes yields visible artifacts (c).

by the success of recent canonicalization-based models on these garment types [7, 23, 34, 38, 47]. However, the surface deformation of skirts and dresses does not directly follow the body articulation. Consequently, extra care needs to be taken to canonicalize and animate such clothing [34], e.g., by introducing extra/virtual bones [20, 29].

Figure 2 shows a typical failure mode of a state-of-the-art canonicalization method, SCANimate [38], on a dress. Here, the clothing-body association is predicted by a neural network, which is learned from data. Theoretically, the network should be able to learn an optimal canonical skirt shape such that it matches the data when transformed back to the posed space. However, in practice, it is difficult to avoid over-stretching and splitting the raw scans. When the neural network is trained to fill the gap without supervision, reposing the canonical representation produces creases. As a result, even on training data (as in Fig. 2), the learned reconstructed skirt shape typically has a clear crease, with the resulting garment looking more like wide pants.

The recent point-based clothed human model, POP [24], side-steps explicit canonicalization and models the clothed body shapes in the posed space using local coordinates. While showing the potential of capturing clothing that differs from the body topology, POP typically exhibits a failure mode when representing skirts: the skirt surface is often torn apart in the region between the legs, as shown in Figs. 1 and 3. One of the key reasons is that the local coordinates are defined based on the unclothed body shape and topology, as we analyze below.

In a nutshell, the POP model densely queries locations on the body surface and generates a clothing displacement vector at each query. For a query point $b_i \in \mathbb{R}^3$ on the *posed* body, the clothing displacement $\hat{r}_i \in \mathbb{R}^3$ is predicted in its associated local coordinate system. The local coordinate frame is a Cartesian coordinate system with b_i being the origin. The clothing point's location in world coordinates is
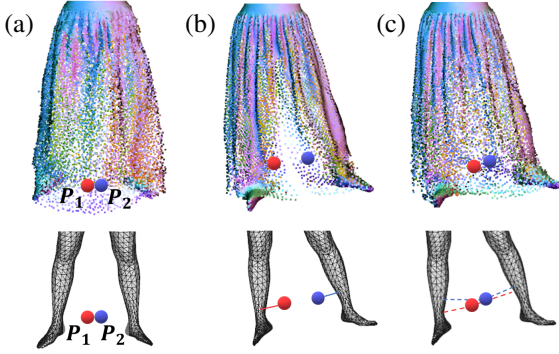
Figure 3: **Failure mode** of the SoTA point-based method on a skirt. (a) Points P_1, P_2 denote two neighboring points on the skirt surface, displaced from the left and right leg respectively. When the pose changes, (b), previous work [24] skins these according to the body point skinning weights, resulting in a "hard" association to only the respective body part (the solid colored lines). This leads to a discontinuous skirt surface in the new pose. (c) Our approach predicts blended skinning weights for each point (the dashed colored lines) leading to a more even distribution of points.

then given by:

$$x_i = \hat{T}_i \cdot \hat{r}_i + b_i. \tag{1}$$

In POP, the transformations $\hat{T}_i$ between the local and world coordinate systems are *pre-computed* using the joint transformations of the SMPL model:

$$\hat{T}_i = \sum_{j=1}^{J} w_{j,\text{SMPL}}^{(i)} T_j, \tag{2}$$

where each T_j is the 4×4 rotation-translation transformation matrix of the body's j-th joint from the canonical-pose space to the posed space, J is the number of joints, and $w_{j,\text{SMPL}}^{(i)}$ is SMPL's linear blend skinning (LBS) weight that associates b_i to the j-th joint, s.t. $\sum_{j=1}^{J} w_{j,\text{SMPL}}^{(i)} = 1$.

Notice that $\hat{T}_i$ in fact also yields b_i (on the posed body) from its corresponding location $\hat{b}_i$ in the canonical pose[1]: $b_i = \hat{T}_i \cdot \hat{b}_i$. It then becomes clear that the POP pipeline is equivalent to building the model in the canonical space:

$$x_i = \hat{T}_i \cdot \hat{x}_i, \tag{3}$$

with $\hat{x}_i = \hat{r}_i + \hat{b}_i$ being the clothed body in the canonical pose. While the predictions are made in the canonical space, POP does not explicitly canonicalize the scan data, and uses the original posed data as supervision during training. This bypasses the drawbacks of the explicit canonicalization as earlier discussed.

However, the pre-computed local coordinate transformations are problematic for skirts. The points on the skirt surface are displaced from *either* the left or right leg. Consider

---

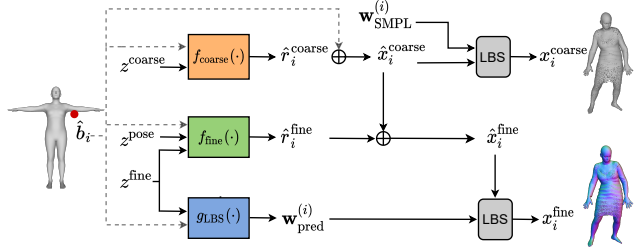[1]With a slight abuse of notation, b_i can represent the homogeneous coordinates when necessary.

Figure 4: **Method Overview.** Given a point $\hat{b}_i$ on the body surface in canonical pose, we use it to query the three MLP networks that predict the pose-independent coarse clothing shape $\hat{r}^{\text{coarse}}$, the pose-dependent detailed clothing offsets $\hat{r}^{\text{fine}}$ (Sec. 4.1), and the LBS weights on the clothing $w_{\text{pred}}^{(i)}$ (Sec. 4.2), respectively. The learned LBS weights are used to transform the canonical-space prediction $\hat{x}_i^{\text{fine}}$ to the posed space x_i^{fine}.

two neighboring points, P_1 and P_2, from the between-leg region on the skirt surface in Fig. 3. P_1 and P_2 are displaced from the right and left leg, respectively. With POP, the LBS weights for P_1, P_2 are effectively drawn from one leg or the other; i.e. $w_{right\ leg}^{P_1} \approx 1$ and $w_{left\ leg}^{P_2} \approx 1$, while other LBS weights from other bones are approximately 0. When the leg pose changes, P_1 and P_2 thus follow the rigid transformation of *only* their respective leg. When the legs move apart, so do the points on the skirt, Fig. 3(b). Although POP's predicted non-rigid pose-dependent displacements can, in theory, compensate for the discrepancy of the rigid transforms, in our experiments, we find this insufficient. Specifically, the proximity of points on the skirt varies with pose. This results in a gap in the 3D representation that splits the skirt into two parts. In fact, the local transformations, as defined in Eq. (2), result in a discontinuity of the displacement field on the skirt surface. We show this with an experiment in Sec. 5.1.

4. Method

Overview. To cope with the challenging clothing types such as skirts and dresses, we push the limit of the local coordinate-based point representation by introducing a series of new components. The training of our method takes a set of scans (point clouds) that capture the shape of a person wearing the same outfit in varying poses. For each scan, we fit a SMPL body that captures the body shape and pose under clothing. We first train a coarse-stage neural network that takes in the SMPL bodies and predicts a pose-*in*dependent shape of the clothed body (Sec. 4.1). We then refine this coarse shape with another neural network that predicts a pose-dependent, detailed clothing displacement field in the local coordinates. The fine stage also predicts a field of transformations of these local coordinates in the form of LBS weights (Sec. 4.2). The displacements are

then added to the coarse shape and posed using the predicted transformations, yielding the final clothed body prediction. We additionally introduce an adaptive regularization algorithm in Sec. 4.3 to facilitate training the pipeline. An overview of our method is illustrated in Fig. 4.

4.1. Coarse-to-fine Prediction

Based on the analysis in Sec. 3, an immediate remedy is to associate the points on a skirt to multiple body parts, instead of a "hard" association to only one part. Taking Fig. 3 again as an intuitive example: if both points P_1, P_2 are associated with both legs, they will undergo similar rigid transformations as the legs move, hence maintaining their proximity, Fig. 3(c). This amounts to "spreading" the LBS weights to all the leg joints instead of tightly associating them with one, such that the LBS weights on the skirt surface vary smoothly in space. The remaining question is, how to get such smoothly transitioning LBS weights on the clothed body surface? Our answer is to first learn a pose-independent coarse shape and then use it to query a pre-diffused smooth LBS weight field derived from SMPL.

Pose-independent coarse shape. We first introduce a neural network that predicts the coarse shape of clothing, represented as residual vectors $\hat{r}_i^{\text{coarse}}$ from the body:

$$\hat{r}_i^{\text{coarse}} = f_{\text{coarse}}(\hat{b}_i, z^{\text{coarse}}) : \mathbb{R}^3 \times \mathbb{R}^{\mathbb{Z}^{\text{coarse}}} \to \mathbb{R}^3, \quad (4)$$

where $f_{\text{coarse}}(\cdot)$ is a multi-layer-preceptron (MLP), $\hat{b}_i$ is a query point on the neutrally-posed body, and z^{coarse} is a *global* geometry code. For the coarse stage, we use the LBS from the SMPL model to bring the predictions back to the posed space, Eq. (1). By querying over the entire body surface, a coarse shape of the clothed body can be constructed.

Since the coarse MLP does not take in any body pose information, the output is pose-independent. Training it on all data with varied poses results in a coarse clothed body shape that minimizes the discrepancy to all examples. Intuitively this facilitates the subsequent fine stage, where the network then only needs to predict a small pose-dependent residual on the coarse base shape. The learned coarse shape takes the place of an artist-designed garment template in a traditional graphics pipeline, but the flexible point-based representation enables our method to be applied to different clothing types without manually defining garment templates.

Pose-dependent fine shape. For the fine stage, we follow the design of POP [24] and train another MLP, $f_{\text{fine}}(\cdot)$, that takes in a local geometry descriptor z_i^{fine}, a local pose descriptor z_i^{pose}, and outputs a clothing residual in the local coordinates:

$$\hat{r}_i^{\text{fine}} = f_{\text{fine}}(\hat{b}_i, z_i^{\text{fine}}, z_i^{\text{pose}}) : \mathbb{R}^3 \times \mathbb{R}^{\mathbb{Z}^{\text{fine}}} \times \mathbb{R}^{\mathbb{Z}^{\text{pose}}} \to \mathbb{R}^3, \ (5)$$

where the subscript i of the descriptors denotes that they are local to each query point. Following POP, the geome-

try features z_i^{coarse} and z_i^{fine} are learned in an auto-decoding fashion. But, unlike POP, the predicted residuals are now added to the coarse clothed body shape, instead of the unclothed body model. The $f_{\text{fine}}(\cdot)$ MLP branches out another head at the final layer to predict the normal vector in the local coordindates. More details of the local descriptors are provided in the Sup. Mat.

Pre-diffused LBS weight field. With the learned clothed coarse body shape, we are now able to define smoothly-varying LBS weights for the coarse shape. To do so, we first diffuse SMPL's LBS weights to $\mathbb{R}^3$ using nearest neighbor assignment as in LoopReg [4]. Such a pre-diffused LBS weight field has many undesirable discontinuities in space (see Sup. Mat.). We then run an optimization to smooth the field. To do so, we sample a regular grid in the diffused space, and minimize the discrepancy between each grid point's LBS weight and the average weight of its 1-ring neighbors; this is analogous to applying Laplacian smoothing to the LBS weights. The optimization results in a LBS weight field that has a smoother spatial variation. Finally, we use the points from the coarse shape to query this field, obtaining smoothly-varing LBS weights on the clothed body "template". More details about our LBS smoothing algorithm can be found in the Sup. Mat.

4.2. Predicting Local Transformations

Although the optimized LBS weight field in theory addresses the single association problem, the optimization is based on heuristics and it is unclear if these pre-processed LBS weights are the optimal choice for every clothing type. Therefore, we train another neural network, $g_{\text{LBS}}(\cdot)$, to predict the LBS weights, and use the pre-diffused field for regularization:

$$\mathbf{w}_{\text{pred}}^{(i)} = g_{\text{LBS}}(\hat{b}_i, z_i^{\text{fine}}) : \mathbb{R}^3 \times \mathbb{R}^{\mathbb{Z}^{\text{fine}}} \to \mathbb{R}^J, \quad (6)$$

with $\mathbf{w}_{\text{pred}}^{(i)} = \{w_{j,\text{pred}}^{(i)}\}_{j=1}^J$ being the set of predicted LBS weights at the query location $\hat{b}_i$. Note that the LBS prediction network is conditioned on the clothing geometry features but not the pose features, following the common design principle that the LBS weights are outfit-specific but pose-independent.

Using the predicted LBS weights, we modify Eq. (2) as:

$$\hat{T}_i' = \sum_{j=1}^J w_{j,\text{pred}}^{(i)} T_j. \quad (7)$$

The clothing residuals and surface normals predicted from the fine stage, $\hat{r}_i^{\text{fine}}$, are now transformed with the predicted local transformations $\hat{T}_i'$.

4.3. Point-adaptive Upsampling and Regularization

In POP, when the skirt "splits", the points becomes exceptionally sparse on the region between the legs, Fig. 3(b).

We propose an algorithm to increase the point density for such regions, aiming to eliminate such artifacts in a simple post-processing step. To identify the points in the low-density region, we compute the average distance of each point to its k-nearest neighbors (kNN, we use k=5). This measures how isolated each point is from its neighbors, giving a measure of local point density. We then again query the underlying body surface using the kNN radius as guidance, and assign a higher sampling probability to the regions with low point density.

As we show in the experiments, this simple post-process can make the point distribution more uniform and visually reduce the "split" artifact, but unfortunately does not eliminate it entirely. This indicates that the learned neural field for the skirt surface is still discontinuous.

To address this, we introduce the point-adaptive idea in training. For points that have a large kNN radius, we set a smaller weight λ_i^{adapt} for the regularization on the norm of their displacements (see Eq. (8)). Intuitively, with a uniform sampling on the body surface, the displaced clothing points with a large kNN radius are more likely to be far from the body. Consequently, the clothing deformations there receive less influence from the body articulations, hence the non-rigid deformations (displacements) can have more freedom. As we show in the experiments, the point-adaptive regularization effectively improves the uniformity of the generated point clouds. Further details of the point-adaptive sampling are provided in the Sup. Mat.

4.4. Training and Inference

Our pipeline is trained with a two-stage regime. The coarse shape network is trained first such that it provides a base shape for the fine stage training. It also serves as a medium to acquire the pre-diffused LBS weights for regularizing the LBS prediction. The pose-dependent fine shape network and the LBS weight prediction networks are trained together end-to-end.

Since the coarse base shape does not vary with pose, when dealing with unseen poses at test-time, we only need to run the fine-stage inference given the query pose.

Loss Functions. Both the coarse and fine stages are trained with a weighted sum of the Chamfer Distance $\mathcal{L}_{\text{CD}}$, a normal consistency loss $\mathcal{L}_{\text{n}}$, and a regularization term $\mathcal{L}_{\text{rgl}}$ on the norm of the predicted displacements. The standard Chamfer and normal losses follow the formulation in [24] and we defer their description in the Sup. Mat.

The predicted displacements $\hat{r}^{\text{coarse}}$ and $\hat{r}^{\text{fine}}$ are regularized by their L2-norm, weighted by the point-adaptive regularization strength λ_i^{adapt} as discussed in Sec. 4.3:

$$\mathcal{L}_{\text{rgl}} = \frac{1}{|\mathbf{X}|} \sum_{i=1}^{|\mathbf{X}|} \lambda_i^{\text{adapt}} \big\| \hat{r}_i \big\|_2^2. \qquad (8)$$

For the fine stage specifically, we use two regularization terms to facilitate the learning of the LBS weight prediction. The first one is a direct L1 regularization using the ground truth LBS weights from the underlying body:

$$\mathcal{L}_{\text{LBS}} = \frac{1}{|\mathbf{X}|} \sum_{i=1}^{|\mathbf{X}|} \big\| \mathbf{w}_{\text{pred}}^{(i)} - \mathbf{w}_{\text{SMPL}}^{(i)\prime} \big\|_1, \qquad (9)$$

where $\mathbf{w}_{\text{SMPL}}^{(i)\prime}$ is the pre-diffused and smoothed SMPL LBS weights as discribed in Sec. 4.1.

Additionally we introduce a reprojection loss as follows:

$$\mathcal{L}_{\text{reproj}} = \frac{1}{|\mathbf{X}|} \sum_{i=1}^{|\mathbf{X}|} \big\| T_i' \hat{b}_i - b_i^{(gt)} \big\|_2, \qquad (10)$$

where $\hat{b}_i$ is a sampled point on the body in canonical pose, T_i' is the local transform derived from the predicted LBS weights (Eq. (7)), and $b_i^{(gt)}$ is the corresponding point on the ground truth body surface. That is, we use the learned transformations to reproject the canonical-posed body to the pose space, such that it matches the ground truth posed body. Similar to $\mathcal{L}_{\text{LBS}}$, this term also penalizes the learned LBS weights for deviating too far from those of the body.

Further details on the model architecture and training hyperparameters are provided in the Sup. Mat.

5. Experimental Evaluation

Baselines. We compare SkiRT with two state-of-the-art clothed human modeling methods: SCANimate [38] and POP [24]. SCANimate uses a neural implicit shape represenaiton and performs explicit data canonicalization. POP, like our method, represents clothed bodies using dense point clouds predicted in local coordinates.

We also compare with various ablated versions to our method: **(a)** we train POP with our adaptive regularization technique as described in Sec. 4.3; **(b)** we replace SMPL LBS weights in POP with our pre-diffused and smoothed LBS weights as discussed in Sec. 4.1; and **(c)** we enhance POP with predicted LBS weights as described in Sec. 4.2, but without the coarse-to-fine strategy.

Datasets and Metrics. We primarily evaluate on the ReSynth dataset [24], a synthetic dataset of clothed humans featuring rich geometric details and salient pose-dependent clothing deformations. We pay special attention to subjects wearing skirts and dresses of varied styles, lengths and tightness, but also evaluate on several non-skirt outfits, to holistically characterize each method. Following [24], we evaluate all the methods using the Chamfer Distance (m^2) and the L1 normal discrepancy, averaged over all sampled points from each method's prediction. See Sup. Mat. for more details on the evaluation setup.

Table 1: Quantitative comparison with baselines and the ablated versions of our method on the ReSynth [24] dataset. CD: Chamfer Distance ($\times 10^{-4} m^2$); NML: L1 discrepancy between the predicted and ground truth unit normals ($\times 10^{-1}$). Both the lower the better. The best results are shown in bold. The upper section contains subjects wearing skirt/dress outfits; the lower section contains subjects wearing non-skirt loose clothing such as jackets (carla-004, eric-035) and wide T-shirts with pants (alexandra-006).

Subject ID	SCANimate [38]		POP [24]		Ablation (a)		Ablation (b)		Ablation (c)		Ours, Full	
	CD	NML	CD	NML	CD	NML	CD	NML	CD	NML	CD	NML
anna-001	1.34	1.35	0.62	0.82	0.59	0.82	0.59	0.82	0.60	**0.81**	**0.58**	**0.81**
beatrice-025	0.74	1.33	0.34	**0.75**	0.32	**0.75**	0.33	**0.75**	0.33	**0.75**	**0.31**	0.77
christine-027	3.21	1.66	1.72	**0.97**	1.74	1.00	1.68	**0.97**	1.68	**0.97**	**1.54**	0.99
janett-025	2.81	1.59	1.24	0.89	1.23	0.85	1.24	0.82	1.19	**0.81**	**1.10**	0.82
felice-004	20.79	2.94	7.34	1.24	7.43	1.25	6.96	**1.22**	6.95	1.22	**6.45**	1.25
debra-014	20.38	2.61	7.40	1.26	7.59	1.30	7.05	1.28	7.13	**1.26**	**6.29**	1.26
carla-004	0.90	1.52	0.51	1.02	0.51	1.07	0.53	1.06	0.52	**1.05**	**0.48**	1.06
alexandra-006	2.28	1.84	1.71	1.29	1.68	**1.28**	1.67	**1.28**	1.74	**1.28**	**1.51**	1.29
eric-035	2.54	1.94	1.34	1.16	1.33	1.16	1.33	1.16	1.33	**1.13**	**1.30**	1.17

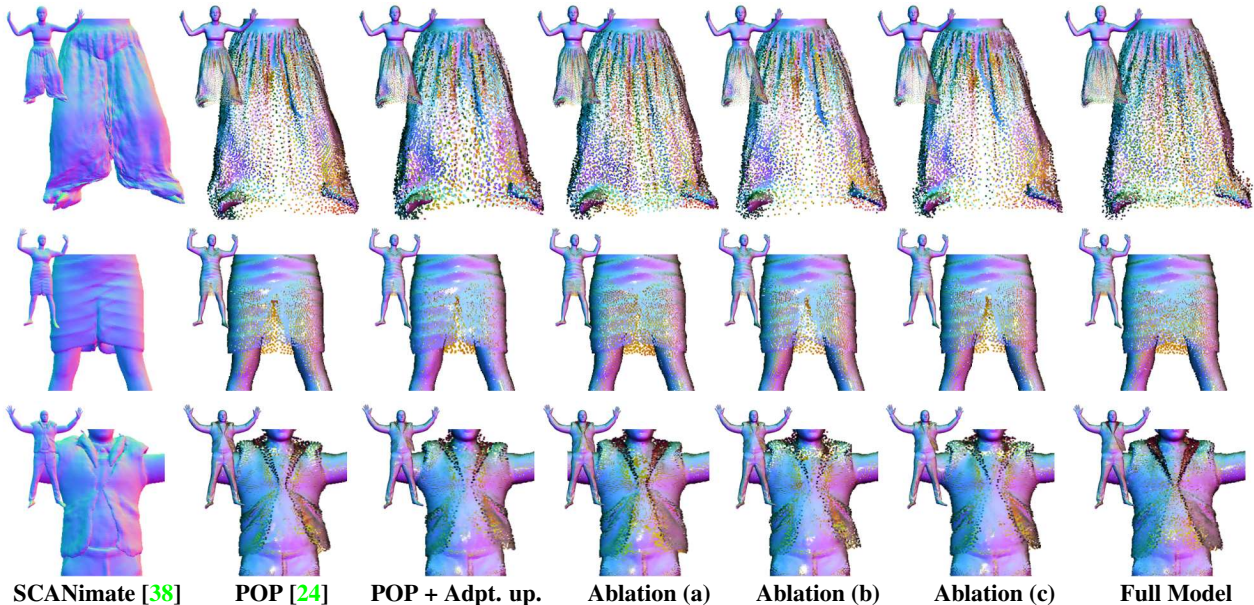| **SCANimate [38]** | **POP [24]** | **POP + Adpt. up.** | **Ablation (a)** | **Ablation (b)** | **Ablation (c)** | **Full Model** |

Figure 5: Qualitative comparison with baselines and ablated versions of our model. "Adpt. up." means using adaptive upsampling as a post-process. Subject IDs from top to bottom: "felice-004", "anna-001", "eric-035". Best viewed zoomed-in on a color screen.

5.1. Comparison to Prior Methods

Table 1 summarizes the quantitative errors on the 9 subject-outfit types from the ReSynth dataset. On most clothing types, our method shows a clear performance advantage over the baselines. As analyzed in Sec. 3, SCANimate often suffers from the pant-like artifacts for skirts, erroneously generating extra surfaces under the skirt and leading to high Chamfer errors especially for skirts and dresses. This shows the limitation of explicit canonicalization on these challenging clothing types. In contrast, POP performs reasonably well on most subjects, showing the advantage of modeling using local coordinates. However, the results from POP often suffer from an uneven distribution of points,

especially for the long skirts and dresses, as shown in Fig. 5 (also see the teaser figure). Using our adaptive upsampling as post-process, the "split" artifact in the skirts can be partially reduced, as shown in the third column in Fig. 5. In contrast, our full model achieves a more uniform point distribution for both skirts and "plain" clothing as shown in Fig. 5: for skirts, it effectively eliminates the "split" artifact present in prior methods; for the blazer jacket it achieves the highest sharpness representing the collar and the lapel.

5.2. Ablation Study

Comparing our full model with its ablated versions (a)-(c), we see the effectiveness of our introduced compo-
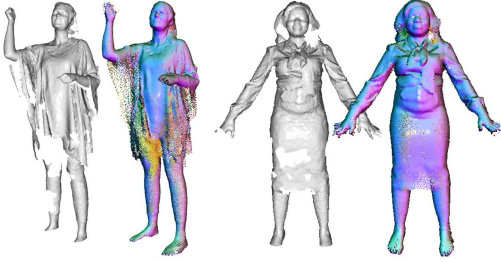
Figure 6: Examples of using SkiRT to complete raw scans on two challenging clothing types. See Sup. Mat. for animated results.
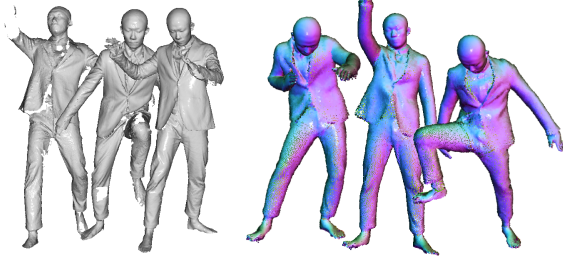

Figure 7: Avatar creation from raw scans. Left: raw training scan data with missing information (holes). Right: animation of learned avatar with pose-dependent clothing deformation.

nents. Ablation **(a)** adds the point-adaptive regularization to POP in training, resulting in a more uniform distribution of points. However, this mechanism alone does not ensure consistently lower errors compared to the original POP model. In ablation **(b)**, the pre-diffused LBS weight field indeed improves the numerical accuracy for all dress/skirt outfits, which, to an extent, verifies our analysis in Sec. 3. However, as shown in Fig. 5, the point density still remains highly uneven for skirts. This motivates us to learn the LBS weights instead of using the fixed, pre-diffused version. Ablation **(c)** differs from POP in that it predicts the LBS weights. But without the coarse stage, here the predicted displacements are added directly to the body surface. Despite a higher accuracy than POP for all skirt outfits, qualitatively this version still suffers from an inhomogeneous point distribution especially for the long skirt. This illustrates the value of the coarse stage. Finally, with the coarse stage, the full model achieves the lowest Chamfer error on all clothing types (both skirts and non-skirts), with an especially clear improvement on the two long skirt subjects, "felice-004" and "debra-014", as can also be seen in the visual results. In the Sup. Mat. we further discuss the influence of the point cloud quality on the mesh reconstruction.

5.3. Raw Scan Completion and Avatar Creation

SkiRT learns a continuous clothing displacement field from the body, and is able to deal with raw scans too. Here we demonstrate two applications of SkiRT on raw scans. Details of these experiments are provided in the Sup. Mat.

Scan completion. The raw scans from real world scanners typically contain holes, and our pipeline can serve as a solution to hole-filling and scan completion. Consider a sequence (typically 200-300 frames) of raw, incomplete scans of a subject. We first fit the SMPL body to each frame, and then train our pipeline with the scan-body pairs. Our model can leverage the shared information between different data frames (i.e. the subject wears the same garment), and predict a completed shape for each frame, as shown in Fig. 6.

Avatar creation from raw scans. As shown above, SkiRT is robust to missing regions in raw scans. With sufficient data, it is possible to create a clothed human avatar with pose-dependent clothing deformation directly from raw scans. Here we show such an application in Fig. 7. Trained directly on raw scans of the subject "03375-blazerlong" in the CAPE dataset [23], SkiRT is capable of animating the subject in novel poses with natural pose-dependent clothing deformation. This automated process significantly reduces the manual effort required for realistic avatar creation.

6. Conclusion

We have introduced a novel point-based representation for modeling clothed humans that, without loss of generality, addresses challenging clothing types such as skirts and dresses. It consists of three key innovations: a coarse-to-fine scheme, predicted local transformations, and point-adaptive up-sampling and regularization. We have systematically analyzed and evaluated our model on diverse clothing types with different skirt lengths, tightness and styles, outperforming the state-of-the-art methods and providing insights for future work on point-based representations for modeling clothed humans.

Limitations and future work. Learning effective skinning weights for clothed body surfaces requires diverse training poses. When training data has very limited pose variations, our learned LBS weights (hence the local-to-world transformations) may fail to generalize to unseen poses. The generalization in pose space with limited training data remains a challenging topic for future research. The point-based coarse stage can sometimes lead to slightly noisier point sets than POP; see Sup. Mat. for detailed discussions. In this paper, we have demonstrated subject-specific modeling, but our formulation is also compatible with multi-subject training by, for example, auto-decoding a clothing geometry code for each subject as in POP [24]. Given sufficient subject and clothing type variation, it should be possible to learn a shape space of clothed bodies using our approach.

References

[1] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 408–416, 2005. 2

[2] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. CLOTH3D: clothed 3d humans. In *European Conference on Computer Vision*, pages 344–359. Springer, 2020. 2

[3] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3D human reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 311–329, 2020. 2, 3

[4] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. LoopReg: Self-supervised learning of implicit surface correspondences, pose and shape for 3D human mesh registration. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12909–12922, 2020. 2, 5

[5] Andrei Burov, Matthias Nießner, and Justus Thies. Dynamic surface function networks for clothed human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[6] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J. Black, Andreas Geiger, and Otmar Hilliges. gDNA: Towards generative detailed neural avatars. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 20427–20437, June 2022. 3

[7] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[8] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5939–5948, 2019. 3

[9] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3D shape reconstruction and completion. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6970–6981, 2020. 3

[10] Julian Chibane, Aymen Mir, and Gerard Pons-Moll. Neural unsigned distance fields for implicit function learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21638–21652, 2020. 3

[11] Edilson De Aguiar, Leonid Sigal, Adrien Treuille, and Jessica K Hodgins. Stable spaces for real-time clothing. In *ACM Transactions on Graphics (TOG)*, volume 29, page 106. ACM, 2010. 2

[12] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. PINA: Learning a personalized implicit neural avatar from a single RGB-D video sequence. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 20470–20480, 2022. 3

[13] Shunwang Gong, Lei Chen, Michael Bronstein, and Stefanos Zafeiriou. Spiralnet++: A fast and highly efficient mesh convolution operator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

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

[15] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5052–5063, 2020. 3

[16] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. BCNet: Learning body and cloth shape from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 18–35. Springer, 2020. 2

[17] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total Capture: A 3D deformation model for tracking faces, hands, and bodies. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8320–8329, 2018. 2

[18] Zorah Lähner, Daniel Cremers, and Tony Tung. DeepWrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 698–715, 2018. 2

[19] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[20] Lijuan Liu, Youyi Zheng, Di Tang, Yi Yuan, Changjie Fan, and Kun Zhou. Neuroskinning: Automatic skin binding for production characters with deep graph networks. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3

[21] 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. 2

[22] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J. Black. SCALE: Modeling clothed humans with a surface codec of articulated local elements. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 16082–16093, June 2021. 2, 3

[23] 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 *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6468–6477, 2020. 2, 3, 8

[24] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10974–10984, Oct. 2021. 1, 2, 3, 4, 5, 6, 7, 8

[25] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4460–4470, 2019. 3

[26] Alexandros Neophytou and Adrian Hilton. A layered model of human body and garment deformation. In *International*

Conference on 3D Vision (3DV), pages 171–178, 2014. 2

[27] Pablo Palafox, Aljaz Bozic, Justus Thies, Matthias Nießner, and Angela Dai. Neural parametric models for 3D deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[28] Pablo Palafox, Nikolaos Sarafianos, Tony Tung, and Angela Dai. Spams: Structured implicit parametric models. *CVPR*, 2022. 3

[29] Xiaoyu Pan, Jiaming Mai, Xinwei Jiang, Dongxue Tang, Jingxiang Li, Tianjia Shao, Kun Zhou, Xiaogang Jin, and Dinesh Manocha. Predicting loose-fitting garment deformations using bone-driven motion networks. *SIGGRAPH*, 2022. 3

[30] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019. 3

[31] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7363–7373, 2020. 2

[32] Priyanka Patel, Chun-Hao Huang Paul, Joachim Tesch, David Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, June 2021. 2

[33] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2

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

[35] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3D faces using convolutional mesh autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 725–741, 2018. 3

[36] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019. 3

[37] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 84–93, 2020. 3

[38] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings IEEE/CVF Conf. on*

Computer Vision and Pattern Recognition (CVPR), June 2021. 1, 2, 3, 6, 7

[39] Igor Santesteban, Miguel A. Otaduy, and Dan Casas. Learning-Based Animation of Clothing for Virtual Try-On. *Computer Graphics Forum*, 38(2):355–366, 2019. 2

[40] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Snug: Self-supervised neural dynamic garments. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8140–8150, 2022. 2

[41] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. SIZER: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 12348, pages 1–18, 2020. 2

[42] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *International Conference on Computer Vision (ICCV)*, October 2021. 2, 3

[43] Nitika Verma, Edmond Boyer, and Jakob Verbeek. Feastnet: Feature-steered graph convolutions for 3d shape analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2598–2606, 2018. 3

[44] Raquel Vidaurre, Igor Santesteban, Elena Garces, and Dan Casas. Fully convolutional graph neural networks for parametric virtual try-on. In *Computer Graphics Forum*, volume 39, pages 145–156. Wiley Online Library, 2020. 2

[45] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *ACM SIGGRAPH 2008 papers*, pages 1–9. 2008. 2, 3

[46] Shaofei Wang, Andreas Geiger, and Siyu Tang. Locally aware piecewise transformation fields for 3D human mesh registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7639–7648, June 2021. 2

[47] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. MetaAvatar: Learning animatable clothed human models from few depth images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3

[48] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[49] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, June 2022.

[50] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-NeRF: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:14955–14966, 2021. 3

[51] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages

6184–6193, 2020. 2

[52] Weiwei Xu, Nobuyuki Umetani, Qianwen Chao, Jie Mao, Xiaogang Jin, and Xin Tong. Sensitivity-optimized rigging for example-based real-time clothing synthesis. *ACM Trans. Graph.*, 33(4):107–1, 2014. 2

[53] 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. 2

[54] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[55] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep Fashion3D: A dataset and benchmark for 3D garment reconstruction from single images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 12346, pages 512–530, 2020. 2, 3

[56] Heming Zhu, Lingteng Qiu, Yuda Qiu, and Xiaoguang Han. Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3845–3854. 3