Learning Implicit Templates for Point-Based Clothed Human Modeling

Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu

Tsinghua University, Beijing, China

Abstract. We present FITE, a First-Implicit-Then-Explicit framework for modeling human avatars in clothing. Our framework first learns implicit surface templates representing the coarse clothing topology, and then employs the templates to guide the generation of point sets which further capture pose-dependent clothing deformations such as wrinkles. Our pipeline incorporates the merits of both implicit and explicit representations, namely, the ability to handle varying topology and the ability to efficiently capture fine details. We also propose diffused skinning to facilitate template training especially for loose clothing, and projection-based pose-encoding to extract pose information from mesh templates without predefined UV map or connectivity. Our code is publicly available at https://github.com/jsnln/fite.

Keywords: 3D modeling; clothed humans; implicit surfaces; point set surfaces.

1 Introduction

The modeling of clothed human avatars is an important topic in many graphics-related fields, such as animation, video games, virtual reality, etc. Traditional solutions [5,21,35,36] are mostly based on rigging and skinning artist-designed avatars, and thus lack realism in representing clothed humans. Another possibility is to apply physics-based simulation [15,25,26,46] which is in general computationally heavy and requires manually designed outfits. Recent work explores learning realistic pose-dependent deformations (e.g. wrinkles) of clothes from posed scan data. From the data-driven perspective, we identify two major challenges in this task: (i) learning the clothing topology; (ii) learning fine details and pose-dependent clothing deformations. In fact, most methods are limited by the representation they choose and cannot fully tackle the challenges above. In particular, mesh-based methods [2,3,8,9,14,13,25,26,33,28,39,44,46,56,59,60,65] are essentially limited by the fixed topology and typically require registered scans for training. On the other hand, implicit surfaces [11,12,17,23,42,41,45,54,53,55,61] can represent varying topology, but are computationally heavy and struggle to represent details. Point sets, on the other hand, enjoys efficiency as well as flexibility. However, generating point sets with details is difficult. Most methods generate sparse points [1,20,34] or points grouped into patches [6,18,19,24,38].

Although some methods achieve higher quality, they typically require dozens of iterations for a single output [32,37,63].

Recently, the state-of-the-art (SOTA) point-based method, POP [40] demonstrates the power of points for cross-outfit modeling and for capturing pose-dependent clothing details. The success of POP lies in its robust body-template-plus-offsets formulation, the flexibility of point sets, and its fine-grained UV-space features. However, POP applies the same minimal body [36,47] for all outfits, which suffers from artifacts such as overly sparse points and discontinuity in clothing, negatively affecting the overall visual quality.

The analysis above suggests implicit surfaces and point sets are complementary in a way that: (i) implicit surfaces can handle varying topology but do not efficiently converge to details; (ii) point sets are efficient and can represent fine details, but even the SOTA point-based scheme [40] is limited to an underlying template with fixed topology, leading to topology-related artifacts. This makes us wonder: Can we incorporate the merits of both types of representations to simultaneously capture the overall topology and final details? With this as motivation, we propose a **First-Implicit-Then-Explicit** framework, abbreviated **FITE**, where the implicit representation and the point set representation are tasked to do what they excel at. Our proposal is a two-stage pipeline: In stage one we train implicit templates that capture the coarse clothing topology for each outfit, and in stage two we predict pose-dependent offsets from the template to generate fine details and pose-dependent clothing deformations. To avoid any conceptual confusion with related work [22,68], we define an *implicit template* as a canonically posed clothed body associated with linear blend skinning weights. Since the templates already capture the coarse topology of given outfits, the second stage can focus on pose-dependent deformations. Compared with POP [40] which directly employs a fixed body template for all outfits, our divide-and-conquer scheme leads to better topology as well as better details. Note that our templates resemble the canonical-space shapes in [11,55], but pose-dependent deformation for the templates is not required in our setting.

Two problems naturally arise with the formulation above. First, in stage one, the training of implicit templates from posed scans requires known correspondences between the canonical space and the posed spaces. Existing approaches [11,55] learn such correspondences by predicting 3D skinning fields. However, they are less accurate in regions far away from the skeleton. To tackle this problem, we precompute a 3D skinning field by smoothly diffusing the skinning weights of SMPL [36] into the whole space, and fix it for subsequent training. This approach effectively reduces the number of learnable parameters and induces more stable correspondences, especially in the case of limited data. Additionally, since the skinning is diffused smoothly, it can handle loose clothing as well. Another problem lies in stage two: How do we encode pose information for learned templates, which do not come with predefined UV or mesh connectivity? We propose to render the canonically posed template to multi-view images whose pixels are the coordinates of the corresponding posed vertices (following [40], we refer to them as *position maps*), and feed them to U-Nets [52] to encode

pose information. Compared with UV-space position maps [38,40], our solution introduces a more continuous feature space for the templates and exhibits less topological artifacts.

We summarize our contributions as follows.

- We propose a **First-Implicit-Then-Explicit** framework for clothed human modeling which incorporates the merits of both implicit and explicit representations, and exhibits better topology properties than current methods.
- For coarse template training, we propose **diffused skinning** which induces stable correspondences from the canonical space to posed spaces, even for limited data or loose clothing.
- For extracting pose information, we propose **projection-based pose encoding** which introduces a continuous feature space on trained templates without predefined UV map or connectivity.

2 Related Work

Modeling clothed human avatars is a task involving various techniques. An ideal approach should at least (i) adopt a suitable 3D representation; (ii) be compatible with existing animation pipelines; (iii) model fine details and pose-dependent clothing deformations. We mainly focus on these three aspects in this review.

2.1 Representations for Clothed Humans

Mesh surfaces are to-date the predominant choice for representing 3D shapes for their compactness and efficiency. For human modeling, most methods represent clothing as deformations [3,8,9,39,44,57,59,65] from minimal bodies [4,27,36,47], or as separate layers [25,26,33,46,56]. Meshes are essentially limited their fixed topology. On the other hand, neural implicit human representations are not subject to a fixed topology [11,12,17,23,42,41,45,54,53,55,61]. Despite the obvious advantages, the low computational efficiency often forbids high-fidelity outputs by implicit methods. Articulating implicitly represented humans is also challenging. Recent methods [11,16,42,55] learn volumetric linear blend skinning to achieve articulation. However, extending linear blend skinning to 3D can be tricky, especially with limited data or loose clothing. Point sets enjoy efficiency as well as flexibility. Nonetheless, producing points with fine details is difficult. Pioneer work [1,20,34] only generates sparse points. Later approaches [6,18,19,24,38] achieve denser generation by grouping points into patches, with notable inter-patch discontinuity as a side-effect. POP [40] conditions the generation on fine-grained UV features to produce dense structured points, but its performance it still limited by the underlying topology of SMPL/SMPL-X [36,47]. Recently, based on neural radiance fields (NeRF) [43], attempts have been made to bypass the underlying geometry and synthesize rendered images of clothed humans directly [49,50,62,67]. However, the lack of an explicit geometry limits their application in downstream tasks such as editing and animation.

2.2 Animating Humans with Linear Blend Skinning

Linear blend skinning (LBS) is a widely used technique for animating human avatars among other articulatable objects. With LBS, a surface is associated with an underlying skeleton, and when the skeleton is articulated, each point on the surface is also transformed by linearly combining the transformations of the bones in the skeleton, according to a set of predefined skinning weights.

LBS is traditionally only applied to 2D mesh surfaces [5,36,47]. Recently, motivated by the need to articulate implicit representations, LBS has also been extended to 3D (called skinning fields) [7,11,16,42,55]. However, there is no direct supervision on the skinning weights for locations far away from the skeleton. LoopReg [7] uses the nearest point on SMPL to extend skinning to 3D, leading to clearly observable spatial discontinuity. SNARF [11] trains a forward skinning field jointly with a canonical-space implicit occupancy field. SCANimate [55] predicts both forward and backward skinning fields and enforce cycle consistency. Despite these efforts, such skinning fields are still limited to tight clothing.

2.3 Modeling Pose-Dependent Clothing Deformations

Clothing deformations are in general non-rigid and pose-dependent, e.g., wrinkles, sliding motions and bulging. Generating realistic pose-dependent deformations requires effectively encoding pose information. Going beyond prior methods that apply a single global feature [16,33,39,46,64], recent work has demonstrated utilizing local information leads to better details and better generalization [38,40,55,57], e.g. UV maps [38,40,57] and attention mechanisms [55].

In our framework, we propose projection-based pose encoding to introduce a continuous pose feature space on templates without predefined UV map or mesh connectivity. Although similar ideas have also appeared elsewhere [10,51,53,54], to the best of our knowledge, we are the first to apply such architectures to encode pose information for animating clothed humans.

3 Method

3.1 Task Formulation and Notations

Our task is to learn animatable clothed human avatars with realistic pose-dependent clothing deformations from a set of posed scans, under a multi-outfit setting. Fig. 1 shows our overall pipeline, where implicit templates are trained in stage one and pose-dependent offsets are predicted in stage two. For simplicity in notations let us for now assume a single outfit worn by the same person. We will introduce how the formulation can be easily extended to the multi-outfit setting at the end of Section 3.3.

We assume input scans are presented in the form of point sets with normals that cover most of the body so that watertight meshes can be extracted for obtaining ground truth occupancy labels (0 for outside and 1 for inside). We denote the point set of the i-th posed scan as $\{p_k^i\}_{k=1}^{N_i} \subset \mathbb{R}^3$, where N_i is the number

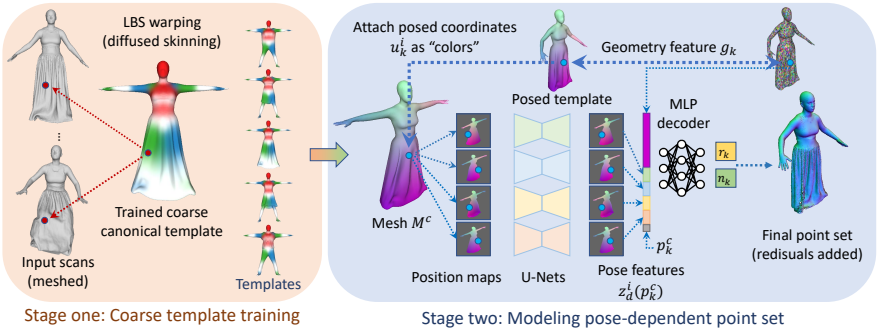

Fig. 1. Overall pipeline of our first-implicit-then-explicit framework. Left: In stage one we learn implicit templates of different outfits with diffused skinning. Right: In stage two we predict pose-dependent offset from features extracted by projection-based encoders.

of points in the i-th scan. The normal at p_k^i is denoted n_k^i. We also assume each scan has a fitted SMPL model [36]. This assumption is reasonable since there are a number of existing techniques that can tractably obtain parametric body models [29,48,58,66]. Note that SMPL is needed only for its skeletal structure and skinning weights for reposing, and that for point set generation our learned implicit templates will be used instead. Let T denote the canonical-pose SMPL body template of the subject and let $\theta^i \in \mathbb{R}^{72}$ denote the SMPL pose parameter corresponding to the i-th scan. T, together with θ^i, determines a set of rigid transformations R_j^i associated to each joint of the SMPL skeleton ($j = 1, \cdots, 24$ is the index for different joints).

Given T and θ^i, if a point $p \in \mathbb{R}^3$ in the canonical space has skinning weights $w(p) = (w_1(p), \cdots, w_{24}(p)) \in \mathbb{R}^{24}$ associated to each joint, we can warp p to its corresponding position q^i in the i-th scan via LBS. We denote this warping as $W(\,\cdot\,,\,\cdot\,; T, \theta^i) : \mathbb{R}^3 \times \mathbb{R}^{24} \to \mathbb{R}^3$. More specifically:

$$q^i = W(p, w(p), T, \theta^i) = \sum_{j=1}^{24} w_j(p) R_j^i(p). \tag{1}$$

Note that $w(p)$ is for now only defined on the SMPL surface T. We introduce how to extend w to the 3D space in Section 3.2.

3.2 Stage One: Coarse Template Training with Diffused Skinning

In this stage, we seek to obtain a template T^c representing the coarse clothing topology. We follow SNARF [11] to learn the template as the 1/2-level-set of a 0-1 occupancy field $F^c : \mathbb{R}^3 \to [0, 1]$ in the canonical space:

$$T^c = \{p \in \mathbb{R}^3 : F^c(p) = 1/2\}. \tag{2}$$

As a quick recap, SNARF [11] jointly optimizes the pose-dependent canonical-space occupancy field $f_{\sigma_f}(\,\cdot\,,\theta^i) : \mathbb{R}^3 \to [0,1]$ and the forward skinning field $w_{\sigma_w}(\,\cdot\,) : \mathbb{R}^3 \to \mathbb{R}^{24}$, both represented by neural networks with σ_f and σ_w as the parameters, by minimizing the binary cross entropy (BCE) loss on the predicted occupancy $f_{\sigma_f}(p,\theta_i)$ and the ground truth occupancy label $o(q^i)$ at the warped location $q^i = W(p, w_{\sigma_w}(p), T, \theta^i)$, i.e.,

$$\min_{\sigma_f,\sigma_w} \mathcal{L}_{\text{BCE}}(f_{\sigma_f}(p,\theta^i), o(q^i)). \tag{3}$$

However, we empirically found that jointly optimizing σ_f and σ_w leads to local minima due to the non-uniqueness of solution of (3). More specifically, an incorrect canonical shape with an incorrect skinning field can accidentally be warped to a correct posed shape (see Fig. 2).

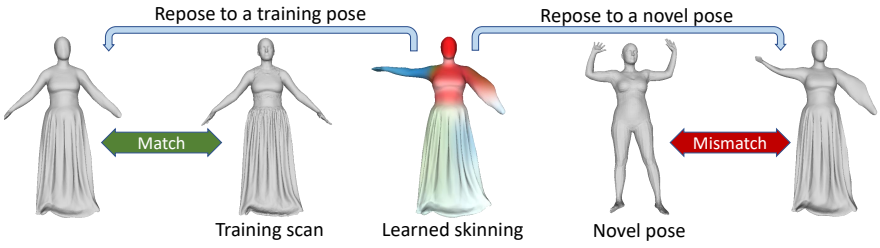

Fig. 2. Illustration of the ill-posedness of jointly optimizing the canonical shape and the skinning fields. An incorrect canonical shape with incorrect skinning can be accidentally warped to a correct pose, but generalization to new poses can be problematic.

To address this ambiguity, we propose to fix the skinning weights in (3) and let the optimization focus on the occupancy only. This effectively reduces learnable parameters and generates more stable results, which is essential for stage two. A good forward skinning weight field $w : \mathbb{R}^3 \to \mathbb{R}^{24}$ should satisfy the following constraints: (i) $w(p)$ should be identical to the SMPL skinning weights $w^s(p)$ for p lying on the SMPL body surface T; (ii) w should naturally diffuse from the SMPL surface, in the sense that its rate of change along the normal direction should be zero. These lead to the following constraints (equations below should be regarded as component-wise):

$$w(p) = w^s(p), \quad \nabla_p w(p) \cdot n^s(p) = 0, \quad \text{for } p \in T, \tag{4}$$

where $w^s(p)$ denotes the SMPL skinning weights at p, and $n^s(p)$ denotes the normal direction of T at p. We remark that the gradient of a scalar function on a curved surface is a defined concept and is in fact a tangent field of the surface. Considering each component of w^s as a scalar function on T, we compute their gradients along T as $\nabla_T w^s$. Note that Eq. (4) says $\nabla_p w$ is tangential to T, and that $w = w^s$ on T. Hence, we can equivalently rewrite Eq. (4) as

$$w(p) = w^s(p), \quad \nabla_p w(p) = \nabla_T w^s(p), \quad \text{for } p \in T, \tag{5}$$

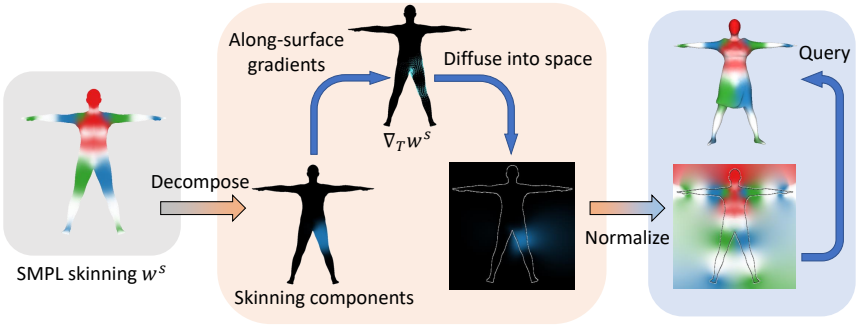

Fig. 3. Diffused skinning visualized. Each component of the skinning weights on SMPL [36] is diffused independently and re-normalized to form a skinning field.

which can be reformulated as minimizing the following energy (with smoothness regularization term $\left\|\nabla^2 w\right\|^2$):

$$\lambda_{\mathrm{p}}^s \int_{p\in T} \|w(p) - w^s(p)\|^2 + \lambda_{\mathrm{g}}^s \int_{p\in T} \|\nabla_p w(p) - \nabla_T w^s(p)\|^2 + \lambda_{\mathrm{reg}}^s \int_{\mathbb{R}^3} \left\|\nabla^2 w\right\|^2, \tag{6}$$

where the λ's are weights to ensure numerical stability. We apply an off-the-shelf solver [30] to obtain w. Note that each component of w is solved separately, clamped to the range of $[0, 1]$ and finally re-normalized to sum up to 1 (Fig. 3). Please refer to the supplementary material for more details.

Having solved w from Eq. (6), we fix it in SNARF [11] and train F_{σ_f} in (3). Techniques such as multiple correspondences in are still employed (see the supplementary material for details). Stage one terminates as soon as a coarse shape is available, which is much shorter than the original setup which strives for fine details. After training, we set F^c as:

$$F^c(p) = F_{\sigma_f}(p, \theta^{i_0}), \quad \text{where } \theta^{i_0} = \arg\min\{\left\|\theta^i\right\|_1 : \theta_i \text{ in the training poses}\}. \tag{7}$$

In other words, we pick the canonical shape closest to the zero-pose in L_1-norm out of all training poses. Note that although F^c is not pose-dependent, it is enough for our purpose. Finally, we extract the 1/2-level-set T^c of $F^c(p)$ as our canonical template, which can be posed via LBS with skinning weights queried from w. This concludes stage one.

3.3 Stage Two: Modeling Pose-Dependent Clothing Deformations

After obtaining the canonical template T^c representing the coarse clothing topology in stage one, we further predict pose-dependent offsets from its surface. First, we uniformly sample a point set $\{p_k^c\}_{k=1}^{N_c}$ on the learned canonical template surface T^c and query their skinning weights in the diffused skinning field: $w(p_k^c)$. Moreover, we follow [40] to assign a geometric feature vector $g_k \in \mathbb{R}^{C_{\mathrm{geom}}}$ to

each p_k^c, learned in an auto-decoding fashion [45]. For a specific pose θ, the final output point set $\{q_k\}_{k=1}^{N_c}$ representing pose-dependent deformations is obtained by offsetting the canonical point set after applying LBS warping. In addition, we take into consideration the possible inaccuracy in T^c and add a template correction offset c_k, leading to the formulation

$$q_k = W(p_k^c + c_k, w(p_k^c), T, \theta) + r_k. \tag{8}$$

Since c_k are corrections made to the template itself, they are designed to be pose-agnostic. In the rest of this section, we introduce how to obtain c_k and r_k.

Pose-Agnostic Template Correction Since T^c is only coarsely trained, it may have not fully converged to align with training scans. For example, T^c may lack facial details, which is generally pose-*in*dependent, but requiring only r_k in Eq. (8) to account for both pose-dependent deformation and pose-independent correction leads to sub-optimal performance. We propose a template correction offset c_k that is pose-agnostic, obtained by feeding the geometric feature g_k to a 4-layer MLP $C(\,\cdot\,)$. The offset c_k is added to p_k^c before LBS warping.

Projection-Based Pose Encoding To generate pose-dependent offsets r_k in Eq. (8), we need to encode pose-dependent features for r_k to condition on. POP [40], which greatly inspired our work, render the coordinates of posed vertices to the UV-space, and encode them with U-Nets [52]. This scheme provides a continuous feature space over the SMPL body surface. However, the continuity of UV-space features only extends up to the boundaries of the UV islands. Moreover, in our case there is no predefined UV mapping for our templates.

To adapt such pose encoding to a more general setting where templates do not have predefined UV, and to make the feature space more continuous, we propose to directly render the posed coordinates to images instead of UV maps. First, we extract the template surface T^c as a triangle mesh M^c, with vertices $\{v_k^c\}$ and associated skinning weights $\{w(v_k^c)\}$. We then warp v_k^c to the i-th pose: $u_k^i = W(v_k^c, w(v_k^c), T, \theta_i)$. Next, we color the mesh M^c by attaching the coordinates of u_k^i to the vertex v_k^c as its "color". Finally we render the "colored" mesh M^c to images with orthographic projections. Each pixel of the rendered images contains the coordinates of the corresponding *posed* vertices. We also adopt a multi-view setup for better coverage of the surface. We choose $N_\text{v} = 4$ views, looking at the template from its left-front side, left-back side, right-front side and right-back side. Moreover, each view is slightly tilted to provide coverage for the top of the head and the bottom of the feet. Following [40], we refer to these rendered images as position maps.

Let the position maps for the i-th pose be denoted as $I_d^i \in \mathbb{R}^{H\times W\times 3}$, where $d = 1, 2, 3, 4$ is the index for 4 viewing directions. We feed them to U-Net [52] encoders U_d (one for each view, but shared across all poses) to extract pose-dependent features $z_d^i = U_d(I_d^i) \in \mathbb{R}^{H\times W\times C_\text{pose}}$, where C_pose is the number of channels of feature maps. With encoded position maps, we are able to extract

pixel-aligned features for an arbitrary point p on T^c by first projecting it to each image and then querying the pixel feature via bilinear interpolation. The sampled pixel-aligned feature is denoted as $z_d^i(p) \in \mathbb{R}^{C_{\text{pose}}}$. We concatenate the sampled features from all views as the final pose feature for p, denoted as $z^i(p) = [z_1^i(p), z_2^i(p), z_3^i(p), z_4^i(p)] \in \mathbb{R}^{N_\text{v} \cdot C_{\text{pose}}}$, where $[\,\cdots\,]$ denotes concatenation.

Decoding Pose-Dependent Deformations The final step is to generate r_k in Eq. (8) and associated normals n_k conditioned on projection-based pose features to represent fine details and pose-dependent clothing deformations. Following [40], we decode the r_k and n_k with an 8-layer MLP $D(\,\cdot\,)$:

$$[r_k^c, n_k^c] = D([z^i(p_k^c), g_k]), \quad r_k^c, n_k^c \in \mathbb{R}^3. \tag{9}$$

In addition, we also employ local transformations as in [40]. Recall that R_j^i in Eq. (1) are rigid transformations determined by θ^i. Let $\hat{R}_j^i$ be the rotation part of R_j^i. Then we apply the weighted combination of $\hat{R}_j^i$ to the output of the decoder $D(\,\cdot\,)$, i.e., $r_k = \sum_{j=1}^{24} w_j(p_k^c)\hat{R}_j^i(r_k^c)$ and $n_k = n_k'/\left\|n_k'\right\|$, where $n_k' = \sum_{j=1}^{24} w_j(p_k^c)\hat{R}_j^i(n_k^c)$. Plugging r_k into (8), together with normals n_k, gives the final point cloud (with normals) $\{(q_k, n_k)\}_{k=1}^{N_c}$ of our method.

When multiple outfits are present in the input data, the templates for different outfits are trained separately in stage one, but share the template corrector C, the pose encoders U_d, and the deformation decoder D. By sharing the neural networks in stage two for all outfits, they can learn clothing deformation patterns in common for different outfit styles.

3.4 Training Losses

For stage one, we did not modify the training loss and training procedure of SNARF [11]. Interested readers are referred to the original paper for more details. For stage two, we define the following loss terms in the spirit of [40]:

$$\mathcal{L}_{\text{total}} = \lambda_\text{p}\mathcal{L}_\text{p} + \lambda_\text{n}\mathcal{L}_\text{n} + \lambda_{\text{c,reg}}\mathcal{L}_{\text{c,reg}} + \lambda_{\text{r,reg}}\mathcal{L}_{\text{r,reg}} + \lambda_{\text{g,reg}}\mathcal{L}_{\text{g,reg}}, \tag{10}$$

where the λ's are the weighting coefficient for different loss terms. The first two terms $\mathcal{L}_\text{p}$ and $\mathcal{L}_\text{n}$ are losses on the point cloud and the normals, respectively. More specifically, let $P^{\text{gt}} = \{(p_k^{\text{gt}}, n_k^{\text{gt}})\}_{k=1}^{N_{\text{gt}}}$ be the ground truth point cloud and let $P^{\text{pd}} = \{(q_k^{\text{pd}}, n_k^{\text{pd}})\}_{k=1}^{N_{\text{pd}}}$ be the predicted point cloud (with normals). Then

$$\mathcal{L}_\text{p} = \frac{1}{N_{\text{pd}}}\sum_{k=1}^{N_{\text{pd}}} \min_{k'} \left((q_k^{\text{pd}} - p_{k'}^{\text{gt}}) \cdot n_{k'}^{\text{gt}}\right)^2 + \frac{1}{N_{\text{gt}}}\sum_{k'=1}^{N_{\text{gt}}} \min_{k} \left\|p_{k'}^{\text{gt}} - q_k^{\text{pd}}\right\|_2^2, \tag{11}$$

$$\mathcal{L}_\text{n} = \frac{1}{N_{\text{pd}}}\sum_{k=1}^{N_{\text{pd}}} \left\|n_k^{\text{pd}} - n_{k'}^{\text{gt}}\right\|_1, \text{ where } k' = \text{argmin}_{k'} \left\|q_k^{\text{pd}} - p_{k'}^{\text{gt}}\right\|. \tag{12}$$

The other three terms are regularization terms on the template correction offsets, the pose-dependent offsets, and the per-point geometric features, respectively:

$$\mathcal{L}_{\text{c,reg}} = \frac{1}{N_{\text{pd}}} \sum_{k=1}^{N_{\text{pd}}} \|c_k\|_2^2 \,,\ \mathcal{L}_{\text{r,reg}} = \frac{1}{N_{\text{pd}}} \sum_{k=1}^{N_{\text{pd}}} \|r_k\|_2^2 \,,\ \mathcal{L}_{\text{g,reg}} = \frac{1}{N_{\text{pd}}} \sum_{k=1}^{N_{\text{pd}}} \|g_k\|_2^2 \,. \quad (13)$$

Please refer to the supplementary material for details of model architectures, hyper-parameter settings and the training procedure.

4 Experiments

In this section we evaluate the representation ability of our method. Due to space limits, we report the results for interpolation, extrapolation and novel scan animation here, and provide extended evaluations, ablation studies and failure cases in the supplementary material.

4.1 Evaluation Details

Baselines We compare our method with the SOTA clothed human modeling methods: POP [40], SNARF [11], and SCANimate [55]. POP [40] also adopts a point cloud representation which is conditioned on SMPL/SMPL-X [36,47] templates. SNARF [11] and SCANimate [55] are currently the SOTA implicit methods with learned volumetric skinning fields.

Datasets We evaluate our method on two large-scale datasets with multiple outfits: ReSynth [40] and CAPE [39]. We follow POP [40] to use 12 outfits and 14 outfits from ReSynth and CAPE, respectively, for cross-outfit training. Note that ReSynth is much more diverse in outfit types than CAPE, and serves as a more convincible test for modeling outfits with different topologies. Moreover, since the implicit baselines do not output point clouds, and the generated point sets from FITE and POP have different densities, we use Screened Poisson Reconstruction [31] to obtain closed meshes for evaluation. Please refer to the supplementary material for more details on dataset preprocessing.

Metrics Following the common evaluation pipeline [40], we use the Chamfer-L_2 distance $d_{\text{cham.}}$ (lower is better) and cosine similarity S_{cos} [45] (higher is better) to measure the error of generated clothed humans. Due to the stochastic nature of clothing, error measurement with the ground truth scans in extrapolated poses does not faithfully reflect the modeling quality. Thus, following previous work [40,55], we conduct a large-scale user study to evaluate the visual quality of different methods for extrapolation experiments. During the user study, each viewer is given either a pair of point clouds or a pair of meshes placed side-by-side, and is asked to vote on the one with higher **overall** visual quality after considering factors such as realism, details and artifacts. The left-right order is randomly shuffled to prevent the preference for a certain side. The choice of presented outfit and pose is also random with equal probability.

4.2 Interpolation Experiments

We evaluate the representation ability of our method with interpolation experiments on the ReSynth dataset and the CAPE dataset. Considering dataset sizes, for training, we choose every 2nd frame for ReSynth and choose every 4th frame for CAPE, both from their official training splits. The rest of the training sequences are used for evaluation.

Table 1 shows the quantitative results of the interpolation evaluation. Note that the modeling difficulty, as well as the error distribution, varies drastically from outfit to outfit. We thus report the quantitative results for each outfit separately. Due to page limits, we report three outfits from CAPE and three from ReSynth and present more in the supplementary material. The results in Table 1 shows that both point-based methods, FITE and POP, outperform implicit methods by a large margin. Between FITE and POP, our method performs notably better for outfits that greatly differ from the minimal body (long dress). Note that the benefit of cross-outfit training for FITE can be more clearly observed from the long dress example. This is due to the fact that projection-based encoding is harder to train than UV encoding and can thus benefit more from the regularization effect brought by cross-outfit training. We will discuss this more closely in the supplementary material.

Table 1. Quantitative results of interpolation experiments. Since SNARF [11] and SCANimate [55] do not support cross-outfit modeling, we evaluate the outfit-specific versions of POP and FITE for fairness (denoted as POP-OS and FITE-OS). Note that d_{cham} reported below have been multiplied by 10^5.

Method	CAPE Data						ReSynth Data					
	00096 jerseyshort		00215 poloshort		03375 blazerlong		Carla 004 long pants		Christine 027 short dress		Felice 004 long dress	
	d_{cham}	S_{cos}	d_{cham}	S_{cos}	d_{cham}	S_{cos}	d_{cham}	S_{cos}	d_{cham}	S_{cos}	d_{cham}	S_{cos}
SCANimate [55]	0.632	0.942	0.730	0.927	0.957	0.914	0.721	0.943	1.750	0.940	17.578	0.803
SNARF [11]	0.155	0.964	0.191	0.941	0.624	0.929	0.340	0.949	0.621	0.953	2.426	0.906
POP-OS [40]	**0.036**	**0.987**	0.084	0.980	**0.249**	**0.967**	0.507	0.940	0.437	**0.964**	1.591	0.925
POP [40]	0.044	0.986	0.091	0.978	0.252	0.966	0.485	0.939	**0.421**	0.960	1.718	0.920
FITE-OS	0.041	**0.987**	**0.074**	**0.981**	0.271	0.966	0.300	**0.957**	0.462	**0.964**	1.805	0.918
FITE	0.042	**0.987**	0.076	0.980	0.274	0.964	**0.299**	0.956	0.455	0.963	**1.355**	**0.933**
	Tight clothing					$\Longrightarrow$					Loose clothing	

The improvement can be more obviously observed in Fig. 4. For a long dress, FITE generates densely distributed points for the loose part and is able to represent more details, while the output of POP becomes sparse and can only model a coarse shape. For implicit methods, the modeling of details is less faithful to the ground truth and perceptually less realistic.

4.3 Extrapolation Experiments

For extrapolation experiments, we use the official training sequences (full data, multi-outfit) and test sequences. Fig. 5 shows qualitative comparisons of our

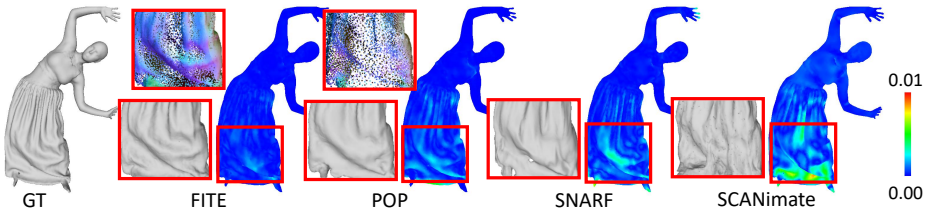

Fig. 4. Qualitative results of the interpolation experiment. Error maps are visualized w.r.t. the largest error in this comparison.

method and POP, with seen outfits in unseen poses. For long dresses, POP produces overly sparse points and fail to represent the surface. For short dresses, POP must deform points from the legs of the SMPL-X template [47] to form the dresses. This incoherency leads to the discontinuity on the dresses. Even for tight clothing, the discontinuity in the UV space also leaves visible seams on the clothing. On the other hand, FITE utilizes templates that already capture the clothing topology and encode pose information in a multi-view projection scheme, producing outputs topologically coherent with given data.

We also conduct a large-scale user study to evaluate quantitatively the extrapolation performance (421 participants, each with 20 votes; 8420 votes in total). Among all votes we received, in terms of generated point clouds (4690 votes), 75.42% prefer FITE over POP (24.58%); in terms of reconstructed meshes (3730 votes), 59.37% prefer FITE over POP (40.63%). Although surface reconstruction partly compensates the drawbacks of POP, the perceptual advantage of our method is still clearly observable. As a final remark, the artifacts of POP in Fig. 5 are not clearly observable in the training set, but they appear frequently in the test set. We believe this reveals that the generalizability of POP is essentially limited by the fixed underlying body template. Please refer to the supplementary material for more details on how the user study is conducted.

4.4 Novel Scan Animation

In stage two, the networks C, U_d and D are shared across outfits and learn the common deformation pattern for different outfits. Thus, they can be used to fit novel scans by optimizing the geometric features only w.r.t. a new scan. Fig. 6 shows the generalization to novel scans. Our method adds pose-dependent to LBS warped templates and produces less noise than POP [40]. Please refer to the supplementary material for more details.

5 Conclusion

We present FITE, a first-implicit-then-explicit framework for modeling clothed humans with realistic pose-dependent deformations. Evaluated on outfits with different topologies, our method is shown to outperform previous methods by

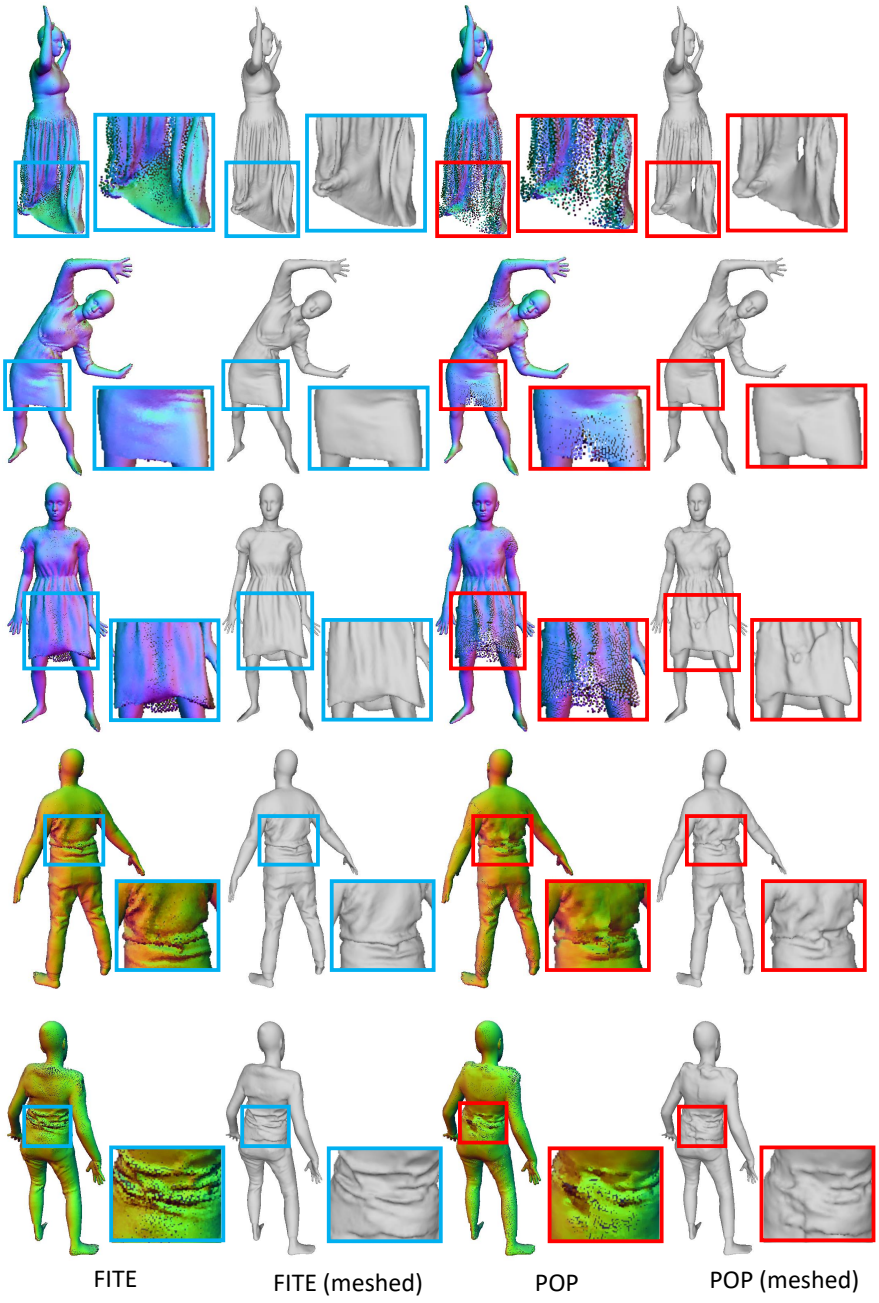

Fig. 5. Qualitative results of pose extrapolation.

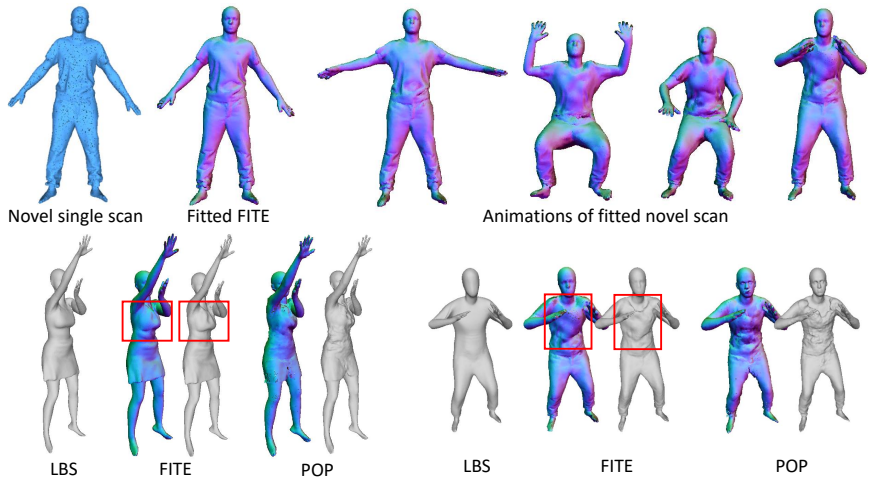

Fig. 6. Novel scan animation results (pose-dependent offsets highlighted).

incorporating the merits of both the implicit representation and the point set representation. Moreover, we believe several individual modules in this framework can also inspire related search, namely, diffused skinning as a smooth interpolation of the learned SMPL skinning weights, and projection-based pose encoding for introducing a continuous feature space on arbitrary mesh surfaces. However, as is currently formulated, several aspects still require further exploration.

Unifying canonical templates In the current framework, stage one learns the coarse templates for each outfit separately. It is worthwhile to explore unifying different outfits with a single shape network, i.e., learning not only a prior for deformations, but also a prior for the outfits, which can lead to faster and more stable outfit generalization.

Driving the underlying templates After obtaining coarse templates in stage one, LBS is applied for reposing. However, LBS does not always reflect true clothing motions, especially for loose clothing in extreme poses. Replacing LBS in stage one with coarse-level physics-based simulation can possibly improve the performance for certain outfits.

Disentanglement of clothing and pose In the second stage of FITE, the projection-based encoders are used to extract pose information from rendered position maps. However, these position maps already contain the clothing information, and thus clothing and pose are not fully disentangled. Future work should explore a representation that further disentangles these factors.

Acknowledgements. This paper is supported by National Key R&D Program of China (2021ZD0113501) and the NSFC project No.62125107 and No.61827805.

References

1. Achlioptas, P., Diamanti, O., Mitliagkas, I., Guibas, L.: Learning representations and generative models for 3D point clouds. pp. 40–49. PMLR (2018)
2. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single rgb camera. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1175–1186 (2019)
3. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2293–2303 (2019)
4. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: ACM Transactions on Graphics. vol. 24, pp. 408–416. ACM (2005)
5. Baran, I., Popović, J.: Automatic rigging and animation of 3D characters. ACM Trans. Graphic. **26**(3), 72–es (2007)
6. Bednařík, J., Parashar, S., Gundogdu, E., Salzmann, M., Fua, P.: Shape reconstruction by learning differentiable surface representations. pp. 4715–4724 (2020)
7. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In: Advances in Neural Information Processing Systems (NeurIPS) (December 2020)
8. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-Garment Net: Learning to dress 3D people from images. pp. 5420–5430 (2019)
9. Burov, A., Nießner, M., Thies, J.: Dynamic surface function networks for clothed human bodies. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10754–10764 (October 2021)
10. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient geometry-aware 3D generative adversarial networks. In: arXiv (2021)
11. Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.: Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 11594–11604 (October 2021)
12. Chibane, J., Mir, A., Pons-Moll, G.: Neural unsigned distance fields for implicit function learning. pp. 21638–21652 (2020)
13. Corona, E., Pumarola, A., Alenya, G., Pons-Moll, G., Moreno-Noguer, F.: Smplicit: Topology-aware generative model for clothed people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11875–11885 (2021)
14. De Aguiar, E., Sigal, L., Treuille, A., Hodgins, J.K.: Stable spaces for real-time clothing. In: ACM Trans. Graphic. vol. 29, p. 106. ACM (2010)
15. Deform Dynamics: `https://deformdynamics.com/`
16. Deng, B., Lewis, J.P., Jeruzalski, T., Pons-Moll, G., Hinton, G., Norouzi, M., Tagliasacchi, A.: Nasa neural articulated shape approximation. pp. 612–628. Springer (2020)
17. Deng, B., Lewis, J., Jeruzalski, T., Pons-Moll, G., Hinton, G., Norouzi, M., Tagliasacchi, A.: Neural articulated shape approximation. pp. 612–628 (2020)
18. Deng, Z., Bednařík, J., Salzmann, M., Fua, P.: Better patch stitching for parametric surface reconstruction. pp. 593–602 (2020)

19. Deprelle, T., Groueix, T., Fisher, M., Kim, V., Russell, B., Aubry, M.: Learning elementary structures for 3D shape generation and matching. pp. 7433–7443 (2019)
20. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3D object reconstruction from a single image. pp. 2463–2471 (2017)
21. Feng, A., Casas, D., Shapiro, A.: Avatar reshaping and automatic rigging using a deformable model. In: Proceedings of the ACM SIGGRAPH Conference on Motion in Games. pp. 57–64 (2015)
22. Genova, K., Cole, F., Vlasic, D., Sarna, A., Freeman, W.T., Funkhouser, T.: Learning shape templates with structured implicit functions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
23. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. pp. 3569–3579 (2020)
24. Groueix, T., Fisher, M., Kim, V.G., Russell, B., Aubry, M.: AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2018)
25. Guan, P., Reiss, L., Hirshberg, D.A., Weiss, A., Black, M.J.: DRAPE: DRessing Any PErson. ACM Trans. Graphic. **31**(4), 35–1 (2012)
26. Gundogdu, E., Constantin, V., Seifoddini, A., Dang, M., Salzmann, M., Fua, P.: GarNet: A two-stream network for fast and accurate 3D cloth draping. pp. 8739–8748 (2019)
27. Hirshberg, D.A., Loper, M., Rachlin, E., Black, M.J.: Coregistration: Simultaneous alignment and modeling of articulated 3d shape. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision – ECCV 2012. pp. 242–255. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
28. Jiang, B., Zhang, J., Hong, Y., Luo, J., Liu, L., Bao, H.: BCNet: Learning body and cloth shape from a single image. pp. 18–35. Springer (2020)
29. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Regognition (CVPR) (2018)
30. Kazhdan, M.: Pointinterpolant. `https://github.com/mkazhdan/PoissonRecon` (2021)
31. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. ACM Trans. Graph. **32**(3) (Jul 2013). https://doi.org/10.1145/2487228.2487237, `https://doi.org/10.1145/2487228.2487237`
32. Klokov, R., Boyer, E., Verbeek, J.: Discrete point flow networks for efficient point cloud generation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 694–710. Springer International Publishing, Cham (2020)
33. Lahner, Z., Cremers, D., Tung, T.: Deepwrinkles: Accurate and realistic clothing modeling. In: Proceedings of the European Conference on Computer Vision. pp. 667–684 (2018)
34. Lin, C.H., Kong, C., Lucey, S.: Learning efficient point cloud generation for dense 3D object reconstruction. pp. 7114–7121 (2018)
35. Liu, L., Zheng, Y., Tang, D., Yuan, Y., Fan, C., Zhou, K.: NeuroSkinning: Automatic skin binding for production characters with deep graph networks. ACM Trans. Graphic. **38**(4), 1–12 (2019)
36. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Transactions on Graphics **34**(6), 248 (2015)
37. Luo, S., Hu, W.: Diffusion probabilistic models for 3d point cloud generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2021)

38. Ma, Q., Saito, S., Yang, J., Tang, S., Black, M.J.: Scale: Modeling clothed humans with a surface codec of articulated local elements. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 16082–16093 (2021)
39. Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M.J.: Learning to dress 3d people in generative clothing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6469–6478 (2020)
40. Ma, Q., Yang, J., Tang, S., Black, M.J.: The power of points for modeling humans in clothing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (Oct 2021)
41. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3D reconstruction in function space. pp. 4460–4470 (2019)
42. Mihajlovic, M., Zhang, Y., Black, M.J., Tang, S.: Leap: Learning articulated occupancy of people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10461–10471 (2021)
43. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Commun. ACM **65**(1), 99–106 (dec 2021). https://doi.org/10.1145/3503250, `https://doi.org/10.1145/3503250`
44. Neophytou, A., Hilton, A.: A layered model of human body and garment deformation. pp. 171–178 (2014)
45. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: DeepSDF: Learning continuous signed distance functions for shape representation. pp. 165–174 (2019)
46. Patel, C., Liao, Z., Pons-Moll, G.: TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style. pp. 7363–7373 (2020)
47. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10975–10985 (2019)
48. Pavlakos, G., Malik, J., Kanazawa, A.: Human mesh recovery from multiple shots. In: CVPR (2022)
49. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: ICCV (2021)
50. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR (2021)
51. Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., Geiger, A.: Convolutional occupancy networks. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 523–540. Springer International Publishing, Cham (2020)
52. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. pp. 234–241. Springer (2015)
53. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2304–2314 (2019)
54. Saito, S., Simon, T., Saragih, J., Joo, H.: PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. pp. 84–93 (2020)

55. Saito, S., Yang, J., Ma, Q., Black, M.J.: Scanimate: Weakly supervised learning of skinned clothed avatar networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2886–2897 (2021)
56. Santesteban, I., Otaduy, M.A., Casas, D.: Learning-Based Animation of Clothing for Virtual Try-On. Computer Graphics Forum **38**(2), 355–366 (2019)
57. Su, Z., Yu, T., Wang, Y., Liu, Y.: Deepcloth: Neural garment representation for shape and style editing. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2022). https://doi.org/10.1109/TPAMI.2022.3168569
58. Tian, Y., Zhang, H., Liu, Y., Wang, L.: Recovering 3d human mesh from monocular images: A survey. arXiv preprint arXiv:2203.01923 (2022)
59. Tiwari, G., Bhatnagar, B.L., Tung, T., Pons-Moll, G.: SIZER: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing. vol. 12348, pp. 1–18 (2020)
60. Vidaurre, R., Santesteban, I., Garces, E., Casas, D.: Fully convolutional graph neural networks for parametric virtual try-on. In: Computer Graphics Forum. vol. 39, pp. 145–156. Wiley Online Library (2020)
61. Wang, S., Mihajlovic, M., Ma, Q., Geiger, A., Tang, S.: Metaavatar: Learning animatable clothed human models from few depth images. In: Advances in Neural Information Processing Systems (2021)
62. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: HumanNeRF: Free-viewpoint rendering of moving people from monocular video. CVPR (2022)
63. Yang, G., Huang, X., Hao, Z., Liu, M.Y., Belongie, S., Hariharan, B.: Pointflow: 3d point cloud generation with continuous normalizing flows. In: Proceedings of the IEEE International Conference on Computer Vision (October 2019)
64. Yang, J., Franco, J.S., Hetroy-Wheeler, F., Wuhrer, S.: Analyzing clothing layer deformation statistics of 3d human motions. In: Proceedings of the European Conference on Computer Vision (September 2018)
65. Yang, S., Pan, Z., Amert, T., Wang, K., Yu, L., Berg, T., Lin, M.C.: Physics-inspired garment recovery from a single-view image. ACM Trans. Graphic. **37**(5), 1–14 (2018)
66. Zhang, Y., Li, Z., An, L., Li, M., Yu, T., Liu, Y.: Lightweight multi-person total motion capture using sparse multi-view cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5560–5569 (2021)
67. Zheng, Z., Huang, H., Yu, T., Zhang, H., Guo, Y., Liu, Y.: Structured local radiance fields for human avatar modeling. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (Jun 2022)
68. Zheng, Z., Yu, T., Dai, Q., Liu, Y.: Deep implicit templates for 3d shape representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1429–1439 (June 2021)