CloSET: Modeling Clothed Humans on Continuous Surface with Explicit Template Decomposition

Hongwen Zhang¹ Siyou Lin¹ Ruizhi Shao¹ Yuxiang Zhang¹ Zerong Zheng¹
Han Huang² Yandong Guo² Yebin Liu³
¹Tsinghua University ²OPPO Research Institute

Abstract

Creating animatable avatars from static scans requires the modeling of clothing deformations in different poses. Existing learning-based methods typically add pose-dependent deformations upon a minimally-clothed mesh template or a learned implicit template, which have limitations in capturing details or hinder end-to-end learning. In this paper, we revisit point-based solutions and propose to decompose explicit garment-related templates and then add pose-dependent wrinkles to them. In this way, the clothing deformations are disentangled such that the pose-dependent wrinkles can be better learned and applied to unseen poses. Additionally, to tackle the seam artifact issues in recent state-of-the-art point-based methods, we propose to learn point features on a body surface, which establishes a continuous and compact feature space to capture the fine-grained and pose-dependent clothing geometry. To facilitate the research in this field, we also introduce a high-quality scan dataset of humans in real-world clothing. Our approach is validated on two existing datasets and our newly introduced dataset, showing better clothing deformation results in unseen poses. The project page with code and dataset can be found at https://www.liuyebin.com/closet.

1. Introduction

Animating 3D clothed humans requires the modeling of pose-dependent deformations in various poses. The diversity of clothing styles and body poses makes this task extremely challenging. Traditional methods are based on either simple rigging and skinning [4,20,33] or physics-based simulation [14,23,24,49], which heavily rely on artist efforts or computational resources. Recent learning-based methods [10,36,39,59] resort to modeling the clothing deformation directly from raw scans of clothed humans. Despite the promising progress, this task is still far from being solved due to the challenges in clothing representations, generalization to unseen poses, and data acquisition, etc.

For the modeling of pose-dependent garment geometry, the representation of clothing plays a vital role in a learning-based scheme. As the relationship between body poses and clothing deformations is complex, an effective representation is desirable for neural networks to capture pose-dependent deformations. In the research of this line, meshes [2,12,38], implicit fields [10,59], and point clouds [36,39] have been adopted to represent clothing. In accordance with the chosen representation, the clothing deformation and geometry features are learned on top of a fixed-resolution template mesh [8,38], a 3D implicit sampling space [10,59], or an unfolded UV plane [2,12,36,39]. Among these representations, the mesh is the most efficient one but is limited to a fixed topology due to its discretization scheme. The implicit fields naturally enable continuous feature learning in a resolution-free manner but are too flexible to satisfy the body structure prior, leading to geometry artifacts in unseen poses. The point clouds enjoy the compact nature and topology flexibility and have shown promising results in the recent state-of-the-art solutions [36,39] to represent clothing, but the feature learning on UV planes still leads to discontinuity artifacts between body parts.
To model the pose-dependent deformation of clothing, body templates such as SMPL [34] are typically leveraged to account for articulated motions. However, a body template alone is not ideal, since the body template only models the minimally-clothed humans and may hinder the learning of actual pose-dependent deformations, especially in cases of loose clothing. To overcome this issue, recent implicit approaches [59] make attempts to learn skinning weights in the 3D space to complement the imperfect body templates. However, their pose-dependent deformations are typically coarse due to the difficulty in learning implicit fields. For explicit solutions, the recent approach [32] suggests learning coarse templates implicitly at first and then the pose-dependent deformations explicitly. Despite its effectiveness, such a workaround requires a two-step modeling procedure and hinders end-to-end learning.

In this work, we propose CloSET, an end-to-end method to tackle the above issues by modeling Clothed humans on a continuous Surface with Explicit Template decomposition. We follow the spirit of recent state-of-the-art point-based approaches [32, 36, 39] as they show the efficiency and potential in modeling real-world garments. We take steps forward in the following aspects for better point-based modeling of clothed humans. First, we propose to decompose the clothing deformations into explicit garment templates and pose-dependent wrinkles. Specifically, our method learns a garment-related template and adds the pose-dependent displacement upon them, as shown in Fig. 1. Such a garment-related template preserves a shared topology for various poses and enables better learning of pose-dependent wrinkles. Different from the recent solution [32] that needs two-step procedures, our method can decompose the explicit templates in an end-to-end manner with more garment details. Second, we tackle the seam artifact issues that occurred in recent point-based methods [36, 39]. Instead of using unfolded UV planes, we propose to learn point features on a body surface, which supports a continuous and compact feature space. We achieve this by learning hierarchical point-based features on top of the body surface and then using barycentric interpolation to sample features continuously. Compared to feature learning in the UV space [39], on template meshes [8, 38], or in the 3D implicit space [57, 59], our body surface enables the network to capture not only fine-grained details but also long-range part correlations for pose-dependent geometry modeling. Third, we introduce a new scan dataset of humans in real-world clothing to facilitate the research of clothed human modeling and animation from real-world scans.

2. Related Work

Representations for Modeling Clothed Humans. A key component in modeling clothed humans is the choice of representation, which mainly falls into two categories: implicit and explicit representations.

Implicit Modeling. Implicit methods [3, 10–12, 15, 21, 29, 41, 42, 48, 59, 67, 70, 77] represent surfaces as the level set of an implicit neural scalar field. Recent state-of-the-art methods typically learn the clothing deformation field with a canonical space decomposition [9, 30, 45, 59, 65] or part-based modeling strategies [16, 25, 46, 55, 75]. Compared to mesh templates, implicit surfaces are not topologically constrained to specific templates [57, 58], and can model various clothes with complex topology. However, the learning space of an implicit surface is the whole 3D volume, which makes training and interpolation difficult, especially when the numbers of scan data are limited.

Explicit Modeling. Mesh surfaces, the classic explicit representation, currently dominate the field of 3D modeling [6–8, 13, 23, 24, 26–28, 38, 43, 49, 60, 62, 64, 66, 69, 72] with their compactness and high efficiency in downstream tasks such as rendering, but they are mostly limited to a fixed topology and/or require scan data registered to a template. Thus, mesh-based representations forbid the learning of a universal model for topologically varying clothing types. Though some approaches have been proposed to allow varying mesh topology [44, 47, 63, 68, 79], they are still limited in their expressiveness. Point clouds enjoy both compactness and topological flexibility. Previous work generates sparse point clouds for 3D representation [1, 19, 31, 74]. However, the points need to be densely sampled over the surface to model surface geometry accurately. Due to the difficulty of generating a large point set, recent methods group points into patches [5, 17, 18, 22]. Each patch maps the 2D UV space to the 3D space, allowing arbitrarily dense sampling within this patch. SCALE [36] successfully applies this idea to modeling clothed humans, but produces notable discontinuity artifacts near patch boundaries.

• We propose a point-based clothed human modeling method by decomposing clothing deformations into explicit garment templates and pose-dependent wrinkles in an end-to-end manner. These learnable templates provide a garment-aware canonical space so that pose-dependent deformations can be better learned and applied to unseen poses.

• We propose to learn point-based clothing features on a continuous body surface, which allows a continuous feature space for fine-grained detail modeling and helps to capture long-range part correlations for pose-dependent geometry modeling.

• We introduce a new high-quality scan dataset of clothed humans in real-world clothing to facilitate the research of clothed human modeling and animation from real-world scans.
Pretty [39] further utilizes a single fine-grained UV map for the whole body surface, leading to a more topologically flexible representation. However, the discontinuity of the UV map may lead to seam artifacts in POP. Very recently, FTTE [32] suggests learning implicit coarse templates [76] at first and then explicit fine details. Despite the efficiency, it requires a two-step modeling procedure. Concurrently, SkiRT [37] proposes to improve the body template by learning the blend skinning weights with several data terms. In contrast, our method applies regularization to achieve template decomposition in an end-to-end manner and learns the point-based pose-dependent displacement more effectively.

**Pose-dependent Deformations for Animation.** In the field of character animation, traditional methods utilize rigging and skinning techniques to repose characters [4, 20, 33, 34, 50], but they fail to model realistic pose-dependent clothing deformations such as wrinkles and sliding motions between clothes and body. We conclude two key ingredients in modeling pose-dependent clothing: (i) pose-dependent feature learning; (ii) datasets with realistic clothing.

**Pose-dependent Feature Learning.** Some traditional methods directly incorporate the entire pose parameters into the model [28, 38, 49, 71]. Such methods easily overfit on pose parameters and introduce spurious correlations, causing bad generalization to unseen poses. Recent work explores poses conditioning with local features, either with point clouds [36, 39] or implicit surfaces [59, 67], and shows superiority in improving geometry quality and in eliminating spurious correlations. Among them, the most relevant to ours is POP [39], which extracts local pose features by utilizing convolution on a UV position map. Despite its compelling performance, POP suffers from artifacts inherent to its UV-based representation, hence the convolution on the UV map produces discontinuity near UV islands’ boundaries [39]. We address this issue by discarding the UV-based scheme and returning to the actual 3D body surface. We attach the features to a uniform set of points on a T-pose body template, and process them via a PointNet++ [53] structure for pose-dependent modeling. As validated by our experiments, our pose embedding method leads to both qualitative and quantitative improvements.

**Clothed Human Datasets.** Another challenging issue for training an animatable avatar is the need for datasets of clothed humans in diverse poses. There are considerable efforts seeking to synthesize clothed datasets with physics-based simulation [14, 23, 24, 39, 49]. Although they are diverse in poses, there remains an observable domain gap between synthetic clothes and real data. Acquiring clothed scans with realistic details of clothing deformations [38, 51, 56, 73, 78] is crucial for the development of learning-based methods in this field.

### 3. Method

As illustrated in Fig. 2, the proposed method CloSET learns garment-related and pose-dependent features on body surfaces (see Sec. 3.1), which can be sampled in a continuous manner and fed into two decoders for the generation of explicit garment templates and pose-dependent wrinkles (see Sec. 3.2).

#### 3.1. Continuous Surface Features

As most parts of the clothing are deformed smoothly in different poses, a continuous feature space is desirable to model the garment details and pose-dependent garment geometry. To this end, we approach first learns features on top of a body template surface, i.e., a SMPL [34] or SMPL-X [50] model in a T-pose. Note that these features are not limited to those on template vertices as they can be continuously sampled from the body surface via barycentric interpolation. Hence, our feature space is more continuous than UV-based [2, 39] spaces, while being more compact than 3D implicit feature fields [59, 67].

To model pose-dependent clothing deformations, the underlying unclothed body model is taken as input to the geometry feature encoder. For each scan, let $V^u = \{v_n^u\}_{n=1}^N$ denote the posed vertex positions of the fitted unclothed body model, where $N = 6890$ for SMPL [34] and $N = 10475$ for SMPL-X [50]. These posed vertices act as the pose code and will be paired with the template vertices $V^t = \{v_n^t\}_{n=1}^N$ of the body model in a T-pose, which shares the same mesh topology with $V^u$. These point pairs are processed by the pose encoder $F_p$ to generate the pose-dependent geometry features $\{\phi_p(v_n^t) \in \mathbb{R}^{C_p}\}_{n=1}^N$ at vertices $V^t$, i.e.,

$$\{\phi_p(v_n^t) \in \mathbb{R}^{C_p}\}_{n=1}^N = F_p(V^t, V^u). \quad (1)$$

To learn hierarchical features with different levels of receptive fields, we adopt PointNet++ [53] as the architecture of the pose encoder $F_p$, where vertices $V^t$ are treated as the input point cloud in the PointNet++ network, while vertices $V^u$ act as the feature of $V^t$. As the template vertices $V^t$ are constant, the encoder $F_p$ can focus on the feature learned from the posed vertices $V^u$. Moreover, the PointNet++ based $F_p$ first abstracts features from the template vertices $V^t$ to sparser points $\{V_i^t\}_{i=1}^L$ at $L$ levels, where the number of $V_i^t$ decreases with $l$ increasing. Then, the features at $\{V_i^t\}_{i=1}^L$ are further propagated back to $V^t$ successively. In this way, the encoder can capture the long-range part correlations of the pose-dependent deformations.

Similar to POP [39], our method can be trained under multi-outfit or outfit-specific settings. When trained with multiple outfits, the pose-dependent deformation should be aware of the outfit type, and hence requires the input of the garment features. Specifically, the garment-related features $\{\phi_g(v_n^t) \in \mathbb{R}^{C_g}\}_{n=1}^N$ are also defined on template
vertices \( V^t \), which are learned by feeding the garment code \( \{\phi_g(v^t_n)\}_{n=1}^N \) to a smaller PointNet++ encoder \( F_g \). Note that the garment-related features \( \phi_g \) are shared for each outfit across all poses and optimized during the training. Since both the pose-dependent and garment-related geometry features are aligned with each other, we denote them as the surface features \( \{\phi(v^t_n)\}_{n=1}^N \) for simplicity. Note that the input of \( \phi_g(p^t_i) \) has no side effect on the results when trained with only one outfit, as the garment features are invariant to the input poses.

**Continuous Feature Interpolation.** In the implicit modeling solutions [42, 57, 59], features are learned in a spatially continuous manner, which contributes to the fine-grained modeling of clothing details. To sample continuous features in our scheme, we adopt barycentric interpolation on the surface features \( \phi \). As illustrated in Fig. 3, for any point \( p^t_i = V^t(b_i) \) lying on the template surface, where \( b_i = [n_{i1}, n_{i2}, n_{i3}, b_{i1}, b_{i2}, b_{i3}] \) denotes the corresponding vertex indices and barycentric coordinates in \( V^t \). Then, the corresponding surface features can be retrieved via barycentric interpolation, i.e.,

\[
\phi(p^t_i) = \sum_{j=1}^3 (b_{ij} \, \phi(v^t_{n_{ij}})). \tag{2}
\]

In this way, the point features are not limited to those learned on the template vertices and are continuously defined over the whole body surface without the seaming discontinuity issue in the UV plane.

### 3.2. Point-based Clothing Deformation

Following previous work [36, 39], our approach represents the clothed body as a point cloud. For any point on the surface of the unclothed body model, the corresponding features are extracted from surface features to predict its displacement and normal vector.

**Explicit Template Decomposition.** Instead of predicting the clothing deformation directly, our method decomposes the deformations into two components: garment-related template displacements and pose-dependent wrinkle displacements. To achieve this, the garment-related template is learned from the garment-related features and shared across all poses. Meanwhile, the learning of pose-dependent wrinkles are conditioned on both garment-related and pose-dependent features. Specifically, for the point \( p^u_i = V^u(b_i) \) at the unclothed body mesh, it has the same vertex indices and barycentric coordinates \( b_i \) as the point \( p^t_i \) on the template surface. The pose-dependent and garment-related features of the point \( p^u_i \) are first sampled according to Eq. (2) based on \( p^t_i = V^t(b_i) \) and then further fed into the garment decoder \( D_g \) and the pose decoder \( D_p \) for displacement predictions, i.e.,

\[
\begin{align*}
    r^g_i &= D_g(\phi_g(p^t_i), p^t_i), \\
    r^p_i &= D_p(\phi_g(p^t_i), \phi_g(p^t_i), p^t_i),
\end{align*}
\tag{3}
\]

where \( \oplus \) denotes the concatenation operation, \( r^g_i \) and \( r^p_i \) are the displacements for garment templates and pose-dependent wrinkles, respectively. Finally, \( r^g_i \) and \( r^p_i \) will be added together as the clothing deformation \( r_i = r^g_i + r^p_i \).
Local Transformation. Similar to [36,39], the displacement \( r_i \) is learned in a local coordinate system. It is further transformed to the world coordinate system by applying the following transformation, i.e., \( x_i = T_i r_i + p_i^b \), where \( T_i \) denotes the local transformation calculated based on the unclothed body model. Following [36,39], the transformation matrix \( T_i \) is defined at the point \( p_i^b \) on the unclothed body model, which naturally supports the barycentric interpolation. Similarly, the normal \( n_i \) of each point is predicted together with \( r_i \) from \( D \) and transformed by \( T_i \).

3.3. Loss Functions

Following previous work [36,39], the point-based clothing deformation is learned with the summation of loss functions: \( L = L_{data} + \lambda_{rgl} L_{rgl} \) where \( L_{data} \) and \( L_{rgl} \) denote the data and regularization terms respectively, and the weight \( \lambda_{rgl} \) balances the loss terms.

Data Term. The data term \( L_{data} \) is calculated on the final predicted points and normals, i.e., \( L_{data} = \lambda_f L_f + \lambda_n L_n \). Specifically, \( L_f \) is the normalized Chamfer distance to minimize the bi-directional distances between the point sets of the prediction and the ground-truth scan:

\[
L_f = \text{Chamfer} \left( \left\{ x_i \right\}_{i=1}^M, \left\{ \hat{x}_j \right\}_{j=1}^{N_s} \right) = \frac{1}{M} \sum_{i=1}^M \min_{j} \| x_i - \hat{x}_j \|_2^2 + \frac{1}{N_s} \sum_{j=1}^{N_s} \min_{i} \| x_i - \hat{x}_j \|_2^2, \tag{4}
\]

where \( \hat{x}_j \) is the point sampled from the ground-truth surface, \( M \) and \( N_s \) denote the number of the predicted and ground-truth points, respectively.

The normal loss \( L_n \) is an averaged \( L1 \) distance between the normal of each predicted point and its nearest ground-truth counterpart:

\[
L_n = L1 \left( \left\{ n_i \right\}_{i=1}^M, \left\{ \hat{n}_j \right\}_{j=1}^{N_s} \right) = \frac{1}{M} \sum_{i=1}^M \| n_i - \hat{n}_j \|, \tag{5}
\]

where \( \hat{n}_j \) is the normal of its nearest point in the ground-truth point set.

Note that we do not apply data terms on the garment templates, as we found such a strategy leads to noisy template learning in our experiments.

Regularization Term. The regularization terms are added to prevent the predicted deformations from being extremely large and regularize the garment code. Moreover, following the previous implicit template learning solution [30, 76], we also add regularization on the pose-dependent displacement \( r_i \) to encourage it to be as small as possible. As the pose-dependent displacement represents the clothing deformation in various poses, such a regularization implies that the pose-invariant deformation should be retained in the template displacement \( p_i^b \), which forms the garment-related template shared by all poses. Overall, the regularization term can be written as follows:

\[
L_{rgl} = \frac{1}{M} \sum_{i=1}^M \| r_i \|_2^2 + \frac{\lambda_{rd}}{M} \sum_{i=1}^M \| r_i - p_i^b \|_2^2 + \frac{\lambda_{gc}}{N} \sum_{n=1}^N \| \phi_{gc}(v_n^t) \|_2^2. \tag{6}
\]

4. Experiments

Network Architecture. For a fair comparison with POP [39], we modify the official PointNet++ [53] (PN++) architecture so that our encoders have comparable network parameters as POP [39]. The modified PointNet++ architecture has 6 layers for feature abstraction and 6 layers for feature propagation (i.e., \( L = 6 \)). Since the input point cloud \( V^t \) has constant coordinates, the farthest point sampling in PointNet++ is only performed at the first forward process, and the sampling indices are saved for the next run. In this way, the runtime is significantly reduced for both training and inference so that our pose and garment encoders can have similar network parameters and runtime speeds to POP. Note that the pose and garment encoders in our method can also be replaced with recent state-of-the-art point-based encoders such as PointMLP [40] and PointNeXt [54]. More details about the network architecture and implementation can be found in the Supp.Mat.

Datasets. We use CAPE [38], ReSynth [39], and our newly introduced dataset THuman-CloSET for training and evaluation. CAPE [38] is a captured human dataset consisting of multiple humans in various motions. The outfits in this dataset mainly include common clothing such as T-shirts. We follow SCALE [36] to choose blazerlong (with outfits of blazer jacket and long trousers) and shortslong (with outfits of short T-shirt and long trousers) from subject 03375 to validate the efficacy of our method. ReSynth [39] is a synthetic dataset introduced in POP [39]. It is created by using physics simulation, and contains challenging outfits such as skirts and jackets. We use the official training and test split as [39]. THuman-CloSET is our newly introduced dataset, containing high-quality clothed human scans captured by a dense camera rig. We introduce THuman-CloSET for the reason that existing pose-dependent clothing datasets [38, 39] are with either relatively tight clothing or synthetic clothing via physics simulation. In THuman-CloSET, there are more than 2,000 scans of 15 outfits with a large variation in clothing style, including T-shirts, pants, skirts, dresses, jackets, and coats, to name a few. For each outfit, the subject is guided to perform different poses by imitating the poses in CAPE. Moreover, each subject has a scan with minimal clothing in A-pose. THuman-CloSET contains well-fitted body models in the form of SMPL-X [50]. Note that
Table 1. Quantitative comparison with previous point-based methods on ReSynth. † denotes the methods using 1/8 training data.

<table>
<thead>
<tr>
<th>Method</th>
<th>Chamfer-$L_2$ ↓</th>
<th>Normal diff. ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Max</td>
</tr>
<tr>
<td><strong>outfit-specific</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCALE [36]</td>
<td>1.491</td>
<td>8.451</td>
</tr>
<tr>
<td><strong>multi-outfit</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (POP [39]) †</td>
<td>1.490</td>
<td>7.859</td>
</tr>
<tr>
<td>Baseline w. PN++ †</td>
<td>1.290</td>
<td>5.940</td>
</tr>
<tr>
<td>CloSET (Ours) †</td>
<td><strong>1.240</strong></td>
<td><strong>5.543</strong></td>
</tr>
</tbody>
</table>

Table 2. Quantitative comparison of different methods on the proposed THuman-CloSET dataset in the outfit-specific setting.

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>SCANimate</th>
<th>SNARF</th>
<th>POP</th>
<th>CloSET</th>
</tr>
</thead>
<tbody>
<tr>
<td>sweater-000</td>
<td>1.06</td>
<td>1.64</td>
<td>7.11</td>
<td>2.09</td>
</tr>
<tr>
<td>longshirt-001</td>
<td>1.42</td>
<td>1.85</td>
<td>6.66</td>
<td>2.21</td>
</tr>
<tr>
<td>skirt-005</td>
<td>1.93</td>
<td>1.74</td>
<td>9.39</td>
<td>2.31</td>
</tr>
</tbody>
</table>

The loose clothing makes the fitting of the underlying body models quite challenging. For more accurate fitting of the body models, we first fit a SMPL-X model on the scan of the subject in minimal clothing and then adopt its shape parameters for fitting the outfit scans in different poses. More details can be found in the Supp.Mat. In our experiments, we use the outfit scans in 100 different poses for training and use the remaining poses for evaluation. We hope our new dataset can open a promising direction for clothed human modeling and animation from real-world scans.

**Metrics.** Following previous work [39], we generate 50K points from our method and point-based baselines and adopt the Chamfer Distance (see Eq. (4)) and the $L_1$ normal discrepancy (see Eq. (5)) for quantitative evaluation. By default, the Chamfer distance (CD) and normal discrepancy (NML) are reported in the unit of $10^{-4}m^2$ and $10^{-1}$, respectively. To evaluate the implicit modeling methods, the points are sampled from the surface extracted using Marching Cubes [35].

4.1. Comparison with the State-of-the-art Methods

We compare results with recent state-of-the-art methods, including point-based approaches SCALE [36], POP [39], and SkiRT [37], and implicit approaches SCANimate [59] and SNARF [10].

**ReSynth.** Tab. 1 reports the results of the pose-dependent clothing predictions on unseen motion sequences from the ReSynth [39] dataset, where all 12 outfits are used for evaluation. As can be seen, the proposed approach has the lowest mean and max errors, which outperforms all other approaches including POP [39]. Note that our approach needs fewer data for the pose-dependent deformation modeling. By using only 1/8 data, our approach achieves a performance comparable to or even better than other models trained with full data. Tab. 3 also reports outfit-specific performances on 3 selected subject-outfit types, including jackets, skirts, and dresses. In comparison with the recent state-of-the-art method SkiRT [37], our method achieves better results on challenging skirt/dress outfits and comparable results on non-skirt clothing.

**THuman-CloSET.** The effectiveness of our method is also validated on our real-world THuman-CloSET dataset. The sparse training poses and loose clothing make this dataset very challenging for clothed human modeling. Tab. 2 reports the quantitative comparisons of different methods on three representative outfits. Fig. 4 also shows example results of different methods, where we follow previous work [39] to obtain meshed results via Poisson surface reconstruction. We can see that our method generalizes better to unseen poses and produces more natural pose-dependent wrinkles than other methods. In our experiments, we found that SNARF [10] fails to learn correct skinning weights due to loose clothing and limited training poses. As discussed in FITE [32], there is an ill-posed issue of jointly optimizing the canonical shape and the skinning fields, which becomes more severe in our dataset.

4.2. Ablation Study

**Evaluation of Continuous Surface Features.** The point features in our method are learned on the body surface, which provides a continuous and compact feature learning space. To validate this, Tab. 4 summarizes the feature learning space of different approaches and their performances on two representative outfits from CAPE [38]. Here, we only include the proposed Continuous Surface Features (CSF) in Tab. 4 by applying the continuous features on POP [39] for fair comparisons with SCALE [36] and POP [39]. As discussed previously, existing solutions learn features either in a discontinuous space (e.g., CAPE [38] on the fixed resolu-

Table 3. Quantitative comparison of different methods on the ReSynth dataset in the outfit-specific setting. The garment styles are non-skirt, skirt, and dress for carla-004, christine-027, and felice-004, respectively.

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>SCANimate</th>
<th>POP</th>
<th>SkiRT</th>
<th>CloSET</th>
</tr>
</thead>
<tbody>
<tr>
<td>carla-004</td>
<td>0.90</td>
<td>1.52</td>
<td>0.51</td>
<td>1.02</td>
</tr>
<tr>
<td>christine-027</td>
<td>3.21</td>
<td>1.66</td>
<td>1.72</td>
<td>0.97</td>
</tr>
<tr>
<td>felice-004</td>
<td>20.79</td>
<td>2.94</td>
<td>7.34</td>
<td>1.24</td>
</tr>
</tbody>
</table>
Figure 4. Comparison of different clothed human modeling methods on the proposed real-world scan dataset.

Figure 5. Comparison of the approach learned on UV planes (POP [39]) and the approach learned on continuous surfaces (Ours). Our solution alleviates the seam artifacts of POP.

Table 4. Comparison of the modeling ability of different approaches and their feature learning space on the CAPE dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Features</th>
<th>Chamfer $L_2$ ↓</th>
<th>Normal diff. ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAPE [38]</td>
<td>Mesh</td>
<td>1.96</td>
<td>1.37</td>
</tr>
<tr>
<td>NASA [16]</td>
<td>3D space</td>
<td>1.37</td>
<td>0.95</td>
</tr>
<tr>
<td>SCALE [36]</td>
<td>Global</td>
<td>1.46</td>
<td>1.03</td>
</tr>
<tr>
<td>SCALE [36]</td>
<td>UV plane</td>
<td>1.07</td>
<td>0.89</td>
</tr>
<tr>
<td>POP [39]</td>
<td>UV plane</td>
<td>0.78</td>
<td>0.57</td>
</tr>
<tr>
<td>CSF Surface</td>
<td></td>
<td>0.71</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Table 5. Ablation study on the effectiveness of continuous surface features (CSF) and Explicit Template Decomposition (ETD) on a dress outfit (felice-004 from ReSynth [39]).

<table>
<thead>
<tr>
<th>Method</th>
<th>POP POP + ETD</th>
<th>CSF CSF + ETD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD</td>
<td>7.34</td>
<td>7.05</td>
</tr>
<tr>
<td>NML</td>
<td>1.24</td>
<td>6.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.16</td>
</tr>
</tbody>
</table>

Evaluation of Explicit Template Decomposition. The decomposed templates help to capture more accurate pose-dependent deformations and produce more natural wrinkles in unseen poses, especially for outfits that differ largely from the body template. To validate the effectiveness of our decomposition strategy, Tab. 5 reports the ablation experiments on a dress outfit (felice-004) of the ReSynth [39].
5. Conclusions and Future Work

In this work, we present CloSET, a point-based clothed human modeling method that is built upon a continuous surface and learns to decompose explicit garment templates for better learning of pose-dependent deformations. By learning features on a continuous surface, our solution gets rid of the seam artifacts in previous state-of-the-art point-based methods [39]. Moreover, the explicit template decomposition helps to capture more accurate and natural pose-dependent wrinkles. To facilitate the research in this direction, we also introduce a high-quality real-world scan dataset with diverse outfit styles and accurate body model fitting.

Limitations and Future Work. Due to the incorrect skinning weight used in our template, the issue of the non-uniform point distribution remains for the skirt and dress outfits. Combining our method with recent learnable skinning solutions [37, 59] could alleviate this issue and further improve the results. Currently, our method does not leverage information from adjacent poses. Enforcing temporal consistency and correspondences between adjacent frames would be interesting for future work. Moreover, incorporating physics-based losses into the learning process like SNUG [61] would also be a promising solution to address the artifacts like self-intersections.

Acknowledgements. This work was supported by the National Key R&D Program of China (2022YFF0902200), the National Natural Science Foundation of China (No.62125107 and No.61827805), and the China Postdoctoral Science Foundation (No.2022M721844).
References


509


[69] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. Modeling clothing as a separate layer for an animatable human avatar. ACM TOG, dec 2021. 2


