SCALE: Modeling Clothed Humans with a Surface Codec of Articulated Local Elements

Qianli Ma$^{1,2}$ Shunsuke Saito$^1\dagger$ Jinlong Yang$^1$ Siyu Tang$^2$ Michael J. Black$^1$

$^1$Max Planck Institute for Intelligent Systems, Tübingen, Germany $^2$ETH Zürich

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

Abstract

Learning to model and reconstruct humans in clothing is challenging due to articulation, non-rigid deformation, and varying clothing types and topologies. To enable learning, the choice of representation is the key. Recent work uses neural networks to parameterize local surface elements. This approach captures locally coherent geometry and non-planar details, can deal with varying topology, and does not require registered training data. However, naively using such methods to model 3D clothed humans fails to capture fine-grained local deformations and generalizes poorly. To address this, we present three key innovations: First, we deform surface elements based on a human body model such that large-scale deformations caused by articulation are explicitly separated from topological changes and local clothing deformations. Second, we address the limitations of existing neural surface elements by regressing local geometry from local features, significantly improving the expressiveness. Third, we learn a pose embedding on a 2D parameterization space that encodes posed body geometry, improving generalization to unseen poses by reducing non-local spurious correlations.

We demonstrate the efficacy of our surface representation by learning models of complex clothing from point clouds. The clothing can change topology and deviate from the topology of the body. Once learned, we can animate previously unseen motions, producing high-quality point clouds, from which we generate realistic images with neural rendering. We assess the importance of each technical contribution and show that our approach outperforms the state-of-the-art methods in terms of reconstruction accuracy and inference time. The code is available for research purposes at https://qianlim.github.io/SCALE.

1. Introduction

While models of humans in clothing would be valuable for many tasks in computer vision such as body pose and shape estimation from images and videos [9, 15, 31, 32, 35, 36] and synthetic data generation [60, 61, 71, 83], most existing approaches are based on “minimally-clothed” human body models [2, 30, 42, 49, 54, 75], which do not represent clothing. To date, statistical models for clothed humans remain lacking despite the broad range of potential

† Now at Facebook Reality Labs.
applications. This is likely due to the fact that modeling 3D clothing shapes is much more difficult than modeling body shapes. Fundamentally, several characteristics of clothed bodies present technical challenges for representing clothing shapes.

The first challenge is that clothing shape varies at different spatial scales driven by global body articulation and local clothing geometry. The former requires the representation to properly handle human pose variation, while the latter requires local expressiveness to model folds and wrinkles. Second, a representation must be able to model smooth cloth surfaces and also sharp discontinuities and thin structures. Third, clothing is diverse and varies in terms of its topology. The topology can even change with the motion of the body. Fourth, the relationship between the clothing and the body changes as the clothing moves relative to the body surface. Finally, the representation should be compatible with existing body models and should support fast inference and rendering, enabling real-world applications.

Unfortunately, none of the existing 3D shape representations satisfy all these requirements. The standard approach uses 3D meshes that are draped with clothing using physics simulation [3, 38, 41]. These require manual clothing design and the physics simulation makes them inappropriate for inference. Recent work starts with classical rigged 3D meshes and blend skinning but uses machine learning to model clothing shape and local non-rigid shape deformation. However, these methods often rely on pre-defined garment templates [8, 37, 45, 53], and the fixed correspondence between the body and garment template restricts them from generalizing to arbitrary clothing topology. Additionally, learning a mesh-based model requires registering a common 3D mesh template to scan data. This is time consuming, error prone, and limits topology change [56]. New neural implicit representations [12, 46, 51], on the other hand, are able to reconstruct topologically varying clothing types [13, 16, 65], but are not consistent with existing graphics tools, are expensive to render, and are not yet suitable for fast inference. Point clouds are a simple representation that also supports arbitrary topology [21, 39, 77] and does not require data registration, but highly detailed geometry requires many points.

A middle ground solution is to utilize a collection of parametric surface elements that smoothly conform to the global shape of the target geometry [20, 25, 80, 82, 84]. As each element can be freely connected or disconnected, topologically varying surfaces can be effectively modeled while retaining the efficiency of explicit shape inference. Like point clouds, these methods can be learned without data registration.

However, despite modeling coherent global shape, existing surface-element-based representations often fail to generate local structures with high-fidelity. The key limiting factor is that shapes are typically decoded from global latent codes [25, 80, 82], i.e. the network needs to learn both the global shape statistics (caused by articulation) and a prior for local geometry (caused by clothing deformation) at once. While the recent work of [24] shows the ability to handle articulated objects, these methods often fail to capture local structures such as sharp edges and wrinkles, hence the ability to model clothed human bodies has not been demonstrated.

In this work, we extend the surface element representation to create a clothed human model that meets all the aforementioned desired properties. We support articulation by defining the surface elements on top of a minimal clothed body model. To densely cover the surface, and effectively model local geometric details, we first introduce a global patch descriptor that differentiates surface elements at different locations, enabling the modeling of hundreds of local surface elements with a single network, and then regress local non-rigid shapes from local pose information, producing folding and wrinkles. Our new shape representation, Surface Codec of Articulated Local Elements, or SCALE, demonstrates state-of-the-art performance on the challenging task of modeling the per-subject pose-dependent shape of clothed humans, setting a new baseline for modeling topologically varying high-fidelity surface geometry with explicit shape inference. See Fig. 1.

In summary, our contributions are: (1) an extension of surface element representations to non-rigid articulated object modeling; (2) a revised local elements model that generates local geometry from local shape signals instead of a global shape vector; (3) an explicit shape representation for clothed human shape modeling that is robust to varying topology, produces high-visual-fidelity shapes, is easily controllable by pose parameters, and achieves fast inference; and (4) a novel approach for modeling humans in clothing that does not require registered training data and generalizes to various garment types of different topology, addressing the missing pieces from existing clothed human models. We also show how neural rendering is used together with our point-based representation to produce high-quality rendered results. The code is available for research purposes at https://qianlim.github.io/SCALE.

2. Related Work

Shape Representations for Modeling Humans. Surface meshes are the most commonly used representation for human shape due to their efficiency and compatibility with graphics engines. Not only human body models [2, 42, 49, 75] but also various clothing models leverage 3D mesh representations as separate mesh layers [17, 26, 27, 37, 53, 67] or displacements from a minimally clothed body [8, 45, 48, 69, 74, 79]. Recent advances in deep learn-
Articulated Shape Modeling. Articulation often dominates large-scale shape variations for articulated objects, such as hands and human bodies. To efficiently represent shape variations, articulated objects are usually modeled with meshes driven by an embedded skeleton [2, 42, 62]. Mesh-based clothing models also follow the same principle [26, 37, 45, 48, 53], where shapes are decomposed into articulated deformations and non-rigid local shape deformations. While the former are explained by body joint transformations, the latter can be efficiently modeled in a canonical space. One limitation, however, is that registered data or physics-based simulation is required to learn these deformations on a template mesh with a fixed topology. In contrast, recent work on articulated implicit shape modeling [18, 47, 66] does not require surface registration. In this work we compare with Deng et al. [18] on a clothed human modeling task from point clouds and show the superiority of our approach in terms of generalization to unseen poses, fidelity, and inference speed.

Local Shape Modeling. Instead of learning 3D shapes solely with a global feature vector, recent work shows that learning from local shape variations leads to detailed and highly generalizable 3D reconstruction [11, 13, 22, 52, 55, 64, 70]. Leveraging local shape priors is effective for 3D reconstruction tasks from 3D point clouds [11, 13, 22, 28, 55, 70] and images [52, 64, 76]. Inspired by this prior work, SCALE leverages both local and global feature representations, which leads to high-fidelity reconstruction as well as robust generalization to unseen poses.

3. SCALE

Figure 2 shows an overview of SCALE. Our goal is to model clothed humans with a topologically flexible point-based shape representation that supports fast inference and animation with SMPL pose parameters [42]. To this end, we model pose-dependent shape variations of clothing using a collection of local surface elements (patches) that are associated with a set of pre-defined locations on the body. Our learning-based local pose embedding further improves the generalization of pose-aware clothing deformations (Sec. 3.1). Using this local surface element representation, we train a model for each clothing type to predict a set of 3D points representing the clothed body shape given an unclothed input body. Together with the predicted point normals and colors, the dense point set can be meshed or realistically rendered with neural rendering (Sec. 3.2).

3.1. Articulated Local Elements

While neural surface elements [25, 80, 82, 84] offer locally coherent geometry with fast inference, the existing formulations have limitations that prevent us from applying them to clothed-human modeling. We first review the existing neural surface elements and introduce our formulation that addresses the drawbacks of the prior work.

Review: Neural Surface Elements. The original methods that model neural surface elements [25, 80] learn a function to generate 3D point clouds as follows:

$$f_w(p; z) : \mathbb{R}^D \times \mathbb{R}^Z \rightarrow \mathbb{R}^3,$$

(1)

where $f_w$ is a multilayer perceptron (MLP) parameterized by weights $w$, $p \in \mathbb{R}^D$ is a point on the surface element, and $z \in \mathbb{R}^Z$ is a global feature representing object shape. Specifically, $f_w$ maps $p$ on the surface element to a point on the surface of a target 3D object conditioned by a shape code $z$. Due to the inductive bias of MLPs, the resulting 3D point clouds are geometrically smooth within the ele-
ment [73]. While this smoothness is desirable for surface modeling, to support different topologies, AtlasNet [25] requires multiple surface elements represented by individual networks, which increases network parameters and memory cost. The cost is linearly proportional to the number of patches. As a result, these approaches limit expressiveness for topologically complex objects such as clothed humans.

Another line of work represents 3D shapes using a collection of local elements. PointCapsNet [84] decodes a local shape code \( \{ z_k \}_{k=1}^{K} \), where \( K \) is the number of local elements, into local patches with separate networks:

\[
    f_{\text{pcn}}(p; z_k) : \mathbb{R}^D \times \mathbb{R}^Z \rightarrow \mathbb{R}^3. 
\]

While modeling local shape statistics instead of global shape variations improves the generalization and training efficiency for diverse shapes, the number of patches is still difficult to scale up as in AtlasNet for the same reason.

Point Completion Network (PCN) [82] uses two-stage decoding: the first stage predicts a coarse point set of the target shape, then these points are used as basis points for the second stage. At each basis location \( b_k \in \mathbb{R}^3 \), points \( p \) from a local surface element (a regular grid) are sampled and fed into the second decoder as follows:

\[
    f_{\text{pcn}}(b_k, p; z) : \mathbb{R}^3 \times \mathbb{R}^D \times \mathbb{R}^Z \rightarrow \mathbb{R}^3. 
\]

Notably, PCN utilizes a single network to model a large number of local elements, improving the expressiveness with an arbitrary shape topology. However, PCN relies on a global shape code \( z \) that requires learning global shape statistics, resulting in poor generalization to unseen data samples as demonstrated in Sec. 4.3.

Articulated Local Elements. For clothed human modeling, the shape representation needs to be not only expressive but also highly generalizable to unseen poses. These requirements and the advantages of the prior methods lead to our formulation:

\[
    g_{\text{pcn}}(u_k, p; z_k) : \mathbb{R}^{D_1} \times \mathbb{R}^{D_2} \times \mathbb{R}^Z \rightarrow \mathbb{R}^3, 
\]

where \( u_k \in \mathbb{R}^{D_1} \) is a global patch descriptor that provides inter-patch relations and helps the network \( g_{\text{pcn}} \) distinguish different surface elements, and \( p \in \mathbb{R}^{D_2} \) are the local (intra-patch) coordinates within each surface element. Importantly, our formulation achieves higher expressiveness by efficiently modeling a large number of local elements using a single network as in [82] while improving generality by learning local shape variations with \( z_k \).

Moreover, unlike the existing methods [24, 25, 80, 84], where the networks directly predict point locations in \( \mathbb{R}^3 \), our network \( g_{\text{pcn}}(\cdot) \) models residuals from the minimally-clothed body. To do so, we define a set of points \( t_k \in \mathbb{R}^3 \) on the posed body surface, and predict a local element (in the form of residuals) for each body point: \( r_{k,i} = g_{\text{pcn}}(u_k, p_i; z_k) \), where \( p_i \) denotes a sampled point from a local element. In particular, an \( r_{k,i} \) is relative to a local coordinate system\(^1\) that is defined on \( t_k \). To obtain a local element’s position in the world coordinate \( x_{k,i} \), we apply articulations to \( r_{k,i} \) by the known transformation \( T_k \) associated with the local coordinate system, and add it to \( t_k \):

\[
    x_{k,i} = T_k \cdot r_{k,i} + t_k. 
\]

Our network \( g_{\text{pcn}}(\cdot) \) also predicts surface normals as an additional output for meshing and neural rendering, which are also transformed by \( T_k \). The residual formulation with explicit articulations is critical to clothed human modeling as the network \( g_{\text{pcn}}(\cdot) \) can focus on learning local shape variations, which are roughly of the same scale. This leads to the successful recovery of fine-grained clothing deformations.

\(^1\)See SupMat. for more details on the definition of the local coordinates.
as shown in Sec. 4. Next, we define local and global patch descriptors as well as the local shape feature $z_k$.

**Local Descriptor.** Each local element approximates a continuous small region on the target 2-manifold. Following [82], we evenly sample $M$ points on a 2D grid and use them as a local patch descriptor: $p = (p_i, q_i) \in \mathbb{R}^2$, with $p_i, q_i \in [0, 1], \ i = 1, 2, \ldots, M$. Within each surface element, all sampled points share the same global patch descriptor $u_k$ and patch-wise feature $z_k$.

**Global Descriptor.** The global patch descriptor $u_k$ in Eq. (4) is the key to modeling different patches with a single network. While each global descriptor needs to be unique, it should also provide proximity information between surface elements to generate a globally coherent shape. Thus, we use 2D location on the UV positional map of the human body as a global patch descriptor: $u_k = (u_k, v_k)$. While the 3D positions of a neutral human body can also be a global descriptor as in [24], we did not observe any performance gain. Note that $T_k$ and $t_k$ in Eq. (4) are assigned based on the corresponding 3D locations on the UV positional map.

**Pose Embedding.** To model realistic pose-dependent clothing deformations, we condition the proposed neural network with pose information from the underlying body as the local shape feature $z_k$. While conditioning every surface element on global pose parameters $\theta$ is possible, in the spirit of prior work [37, 45, 53, 78], we observe that such global pose conditioning does not generalize well to new poses and the network learns spurious correlations between body parts (a similar issue was observed and discussed in [49] for parametric human body modeling). Thus, we introduce a learning-based pose embedding using a 2D positional map, where each pixel consists of the 3D coordinates of a unique point on the underlying body mesh normalized by a transformation of the root joint. This 2D positional map is fed into a UNet [63] to predict a 64-channel feature $z_k$ for each pixel. The advantage of our learning-based pose embedding is two-fold: first, the influence of each body part is clothing-dependent and by training end-to-end, the learning-based embedding ensures that reconstruction fidelity is maximized adaptively for each outfit. Furthermore, 2D CNNs have an inductive bias to favor local information regardless of theoretical receptive fields [44], effectively removing non-local spurious correlations. See Sec. 4.3 for a comparison of our local pose embedding with its global counterparts.

### 3.2. Training and Inference

For each input body, SCALE generates a point set $X$ that consists of $K$ deformed surface elements, with $M$ points sampled from each element: $|X| = KM$. From its corresponding clothed body surface, we sample a point set $Y$ of size $N$ (i.e., $|Y| = N$) as ground truth. The network is trained end-to-end with the following loss:

$$
\mathcal{L} = \lambda_d \mathcal{L}_d + \lambda_a \mathcal{L}_a + \lambda_r \mathcal{L}_r + \lambda_c \mathcal{L}_c,
$$

where $\lambda_d, \lambda_a, \lambda_r, \lambda_c$ are weights that balance the loss terms.

First, the Chamfer loss $\mathcal{L}_d$ penalizes bi-directional point-to-point L2 distances between the generated point set $X$ and the ground truth point set $Y$ as follows:

$$
\mathcal{L}_d = d(x, y) = \frac{1}{KM} \sum_{k=1}^{K} \sum_{i=1}^{M} \min_{j} \|x_{k,i} - y_{j}\|_2^2 + \frac{1}{N} \sum_{j=1}^{N} \min_{k,i} \|x_{k,i} - y_{j}\|_2^2.
$$

For each predicted point $x_{k,i} \in X$, we penalize the $L1$ difference between its normal and that of its nearest neighbor from the ground truth point set: $\mathcal{L}_n = \frac{1}{KM} \sum_{k=1}^{K} \sum_{i=1}^{M} \|n(x_{k,i}) - n(\arg\min_{y_j \in Y} d(x_{k,i}, y_j))\|_1$,

where $n(\cdot)$ denotes the unit normal of the given point. We also add $L2$ regularization on the predicted residual vectors to prevent extreme deformations:

$$
\mathcal{L}_r = \frac{1}{KM} \sum_{k=1}^{K} \sum_{i=1}^{M} \|r_{k,i}\|_2^2.
$$

When the ground-truth point clouds are textured, SCALE can also represent RGB color inference by predicting another 3 channels, which can be trained with an $L1$ reconstruction loss:

$$
\mathcal{L}_c = \frac{1}{KM} \|c(x_{k,i}) - c(\arg\min_{y_j \in Y} d(x_{k,i}, y_j))\|_1,
$$

where $c(\cdot)$ represents the RGB values of the given point.

**Inference, Meshing, and Rendering.** SCALE inherits the advantage of existing patch-based methods for fast inference. Within a surface element, we can sample arbitrarily dense points to obtain high-resolution point clouds. Based on the area of each patch, we adaptively sample points to keep the point density constant. Furthermore, since SCALE produces oriented point clouds with surface normals, we can apply off-the-shelf meshing methods such as Ball Pivot [5] and Poisson Surface Reconstruction (PSR) [33, 34]. As the aforementioned meshing methods are sensitive to hyperparameters, we present a method to directly render the SCALE outputs into high-resolution images by leveraging neural rendering based on point clouds [1, 57, 81]. In Sec. 4, we demonstrate that we can render the SCALE outputs using SMPLpix [57]. See SupMat for more details on the adaptive point sampling and our neural rendering pipeline.
4. Experiments

4.1. Experimental Setup

Baselines. To evaluate the efficacy of SCALE’s novel neural surface elements, we compare it to two state-of-the-art methods for clothed human modeling using meshes (CAPE [45]) and implicit surfaces (NASA [18]). We also compare with prior work based on neural surface elements: AtlasNet [25] and PCN [82]. Note that we choose a minimally-clothed body with a neutral pose as a surface element for these approaches as in [24] for fair comparison. To fully evaluate each technical contribution, we provide an ablation study that evaluates the use of explicit articulation, the global descriptor $u_k$, the learning-based pose embedding using UNet, and the joint-learning of surface normals.

Datasets. We primarily use the CAPE dataset [45] for evaluation and comparison with the baseline methods. The dataset provides registered mesh pairs (clothed and minimally clothed body) of multiple humans in motion wearing common clothing (e.g. T-shirts, trousers, and a blazer). In the main paper we choose blazerlong (blazer jacket, long trousers) and shortlong (short T-shirt, long trousers) with subject 03375 to illustrate the applicability of our approach to different clothing types. The numerical results on other CAPE subjects are provided in the SupMat. In addition, to evaluate the ability of SCALE to represent a topology that significantly deviates from the body mesh, we synthetically generate point clouds of a person wearing a skirt using physics-based simulation driven by the motion of the subject 00134 in the CAPE dataset. The motion sequences are randomly split into training (70%) and test (30%) sets.

Metrics. We numerically evaluate the reconstruction quality of each method using Chamfer Distance (Eq. (7)), in $m^2$ and the $L_1$-norm of the unit normal discrepancy (Eq. (8)), evaluated over the 12,768 points generated by our model. For CAPE [45], as the mesh resolution is relatively low, we uniformly sample the same number of points as our model on the surface using barycentric interpolation. As NASA [18] infers an implicit surface, we extract an iso-surface using Marching Cubes [43] with a sufficiently high resolution (512$^3$), and sample the surface. We sample and compute the errors three times with different random seeds and report the average values.

Implementation details. We use the SMPL [42] UV map with a resolution of $32 \times 32$ for our pose embedding, which yields $K = 798$ body surface points (hence the number of surface elements). For each element, we sample $M = 16$ square grid points, resulting in 12,768 points in the final output. We uniformly sample $N = 40,000$ points from each clothed body mesh as the target ground truth scan. More implementation details are provided in the SupMat.

Figure 3: Qualitative comparison with mesh and implicit methods. Our method produces coherent global shape, salient pose-dependent deformation, and sharp local geometry. The meshed results are acquired by applying PSR [34] to SCALE’s point+normal prediction. The patch color visualization assigns a consistent set of colors to the patches, showing correspondence between the two bodies.

4.2. Comparison: Mesh and Implicit Surface

Block I of Tab. 1 quantitatively compares the accuracy and inference runtime of SCALE, CAPE [45] and NASA [18]. CAPE [45] learns the shape variation of articulated clothed humans as displacements from a minimally clothed body using MeshCNN [59]. In contrast to ours, by construction of a mesh-based representation, CAPE requires registered templates to the scans for training. While NASA, on the other hand, learns the composition of articulated implicit functions without surface registration, it requires watertight meshes because the training requires ground-truth occupancy information. Note that these two approaches are unable to process the skirt sequences as the thin structure of the skirt is non-trivial to handle using the fixed topology of human bodies or implicit functions.

For the other two clothing types, our approach not only achieves the best numerical result, but also qualitatively demonstrates globally coherent and highly detailed reconstruction results as shown in Fig. 3. On the contrary, the mesh-based approach [45] suffers from a lack of details and fidelity, especially in the presence of topological change. Despite its topological flexibility, the articulated implicit function [18] is outperformed by our method by a large margin, especially on the more challenging blazerlong data (22% in Chamfer-$L_2$). This is mainly due to the artifacts caused by globally incoherent shape predictions for unseen poses, Fig. 3. We refer to the SupMat for extended qualitative comparison with the baselines.

The run-time comparison illustrates the advantage of fast
Table 1: Results of pose dependent clothing deformation prediction on unseen test sequences from the 3 prototypical garment types, of varying modeling difficulty. Best results are in **boldface**.

<table>
<thead>
<tr>
<th>Methods / Variants</th>
<th>Chamfer-$L_2$ ($\times 10^{-3} m^2$) ↓</th>
<th>Normal diff. ($\times 10^{-1}$) ↓</th>
<th>Inference time (s) ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAPE [45]</td>
<td>1.96</td>
<td>1.37</td>
<td>–</td>
</tr>
<tr>
<td>NASA [18]</td>
<td>1.37</td>
<td>0.95</td>
<td>–</td>
</tr>
<tr>
<td>Ours (SCALE, full model)</td>
<td><strong>1.07</strong></td>
<td><strong>0.89</strong></td>
<td><strong>2.69</strong></td>
</tr>
<tr>
<td></td>
<td><strong>1.22</strong></td>
<td><strong>1.12</strong></td>
<td><strong>0.94</strong></td>
</tr>
<tr>
<td></td>
<td>0.009 (+ 1.1)</td>
<td>0.013</td>
<td>–</td>
</tr>
</tbody>
</table>

I. Comparison with SoTAs, Sec. 4.2

II. Global vs Local Elements, Sec. 4.3

- a). global $\mathbf{z}$ [24] + AN [25] → $\mathbf{z}$ + AN [25]
- d). pose param $\mathbf{\theta}$ + PCN [82] + Arti.
- e). w/o Arti. $\mathbf{T}_k$
- f). w/o $\mathbf{u}_k$
- g). UNet $\rightarrow$ PointNet, with $\mathbf{u}_k$
- h). UNet $\rightarrow$ PointNet, w/o $\mathbf{u}_k$
- i). w/o Normal pred.

III. Ablation Study: Key Components, Sec. 4.4

We compare existing neural surface representations [24, 25, 82] in Fig. 4 and block II of Tab. 1. Following the original implementation of AtlasNet [25], we use a global encoder that provides a global shape code $\mathbf{z} \in \mathbb{R}^{1024}$ based on PointNet [58]. We also provide a variant of AtlasNet [25] and PCN [82], where the networks predict residuals on top of the input body and then are articulated as in our approach. AtlasNet with the explicit articulation (b) significantly outperforms the original AtlasNet without articulation (a). This shows that our newly introduced articulated surface elements are highly effective for modeling articulated objects, regardless of neural surface formulations. As PCN also efficiently models a large number of local elements using a single network, (c) and (d) differ from our approach only in the use of a global shape code $\mathbf{z}$ instead of local shape codes. While (c) learns the global code in an end-to-end manner, (d) is given global pose parameters $\mathbf{\theta}$ *a priori*. Qualitatively, modeling local elements with a global shape code leads to noisier results. Numerically, our method outperforms both approaches, demonstrating the importance of modeling local shape codes. Notably, another advantage of modeling local shape codes is its parameter efficiency. The global approaches often require high dimensional latent codes (e.g. 1024), leading to the high usage of network parameters (1.06M parameters for the networks above). In contrast, our local shape modeling allows us to efficiently model shape variations with significantly smaller latent codes (64 in SCALE) with nearly half the trainable parameters (0.57M) while achieving the state-of-the-art modeling accuracy.

4.3. Global vs. Local Neural Surface Elements

We further evaluate our technical contributions via an ablation study. As demonstrated in Sec. 4.3 and Tab. 1 (e), explicitly modeling articulation plays a critical role in the success of accurate clothed human modeling. We also observe a significant degradation by replacing our UNet-based pose embedding with PointNet, denoted as (g) and (h). This indicates that the learning-based pose embedding with a 2D CNN is more effective for local feature learning despite the conceptual similarity of these two architectures that incorporate spatial proximity information. Interestingly, the lack of a global descriptor derived from the UV map, denoted as (f), has little impact on numerical accuracy. As the similar ablation study between (g) and (h) shows significant improvement by adding $\mathbf{u}_k$ in the case of the PointNet architecture, this result implies that our UNet local encoder implicitly learns the global descriptor as part of the local codes $\mathbf{z}_k$. As shown in Tab. 1 (i), another interesting observation is that the joint training of surface normals improves reconstruction accuracy, indicating that the multi-task learning of geometric features can be mutually beneficial.
5. Conclusion

We introduce SCALE, a highly flexible explicit 3D shape representation based on pose-aware local surface elements with articulation, which allows us to faithfully model a clothed human using point clouds without relying on a fixed-topology template, registered data, or watertight scans. The evaluation demonstrates that efficiently modeling a large number of local elements and incorporating explicit articulation are the key to unifying the learning of complex clothing deformations of various topologies.

Limitations and future work. While the UV map builds a correspondence across all bodies, a certain patch produced by SCALE is not guaranteed to represent semantically the same region on the cloth in different poses. Jointly optimizing explicit correspondences [7, 72] with explicit shape representations like ours remains challenging yet promising. Currently, SCALE models clothed humans in a subject-specific manner but our representation should support learning a unified model across multiple garment types. While we show that it is possible to obviate the meshing step by using neural rendering, incorporating learnable triangulation [40, 68] would be useful for applications that need meshes.

Acknowledgements: We thank Sergey Prokudin for insightful discussions and the help with the SMPLpix rendering. Qianli Ma is partially funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 276693517 SFB 1233.

Disclosure: MJB has received research gift funds from Adobe, Intel, Nvidia, Facebook, and Amazon. While MJB is a part-time employee of Amazon, his research was performed solely at, and funded solely by, Max Planck. MJB has financial interests in Amazon, Datagen Technologies, and Meshcapade GmbH.
References


Chen-Hsuan Lin, Chen Kong, and Simon Lucey. Learning


