Deep Parametric Surfaces for 3D Outfit Reconstruction from Single View Image

Hugo Bertiche[1,2], Meysam Madadi[1,2] and Sergio Escalera[1,2]
[1] Universitat de Barcelona, Spain
[2] Computer Vision Center, Spain

Abstract— **We present a methodology to retrieve analytical surfaces parametrized as a neural network. Previous works on 3D reconstruction yield point clouds, voxelized objects or meshes. Instead, our approach yields 2-manifolds in the euclidean space through deep learning. To this end, we implement a novel formulation for fully connected layers as parametrized manifolds that allows continuous predictions with differential geometry. Based on this property we propose a novel smoothness loss. Results on CLOTH3D++ dataset show the possibility to infer different topologies and the benefits of the smoothness term based on differential geometry.**

I. Introduction

3D human reconstruction from still images has been widely explored over the last few years. This is due to the wide applicability it has in entertainment and video game industries, and recently, in the VR/AR domains as well. Understanding of 3D scenarios allows for a higher level of human-computer interaction. While body pose and shape regression has undergone a significant progress by the scientific community, new research lines focus on recovering garments along the body. This task comes with additional challenges, such as more complex geometries, different topologies and color variability.

Deep learning has shown success in 3D inference from RGB images in the past [13], [35], [8], relying on point clouds, meshes and voxel representations. Within human-centric domains, parametric mesh models greatly simplify the task for face, body and hands [7], [22], [12], [31], [27]. Some authors propose leveraging this for garment representation as an extension of parametric body models, or, alternatively, defining a few models of this kind for a few garment types [16], [25], [5], [14], [15], [6]. Nonetheless, previous solutions present some of these drawbacks: single mesh for body and outfit, low or heterogeneous garment resolution or poor garment type –or topology– representativity. Our methodology infers outfits with different topologies as continuous parametric surfaces in the 3D space.

In this paper, we propose a methodology to learn analytical surfaces by parametrizing a fully connected layer in the shape of a 2-manifold. Since garments are 2-manifolds on the Euclidean space, we believe this is the appropriate strategy for this domain. Furthermore, our predictions have a differential geometry and we show how to leverage this to design a novel smoothness loss. Finally, we show this approach can be directly applied for surface color prediction. Our contributions are as follow:

- **Continuous surfaces.** As opposed to the current trends in 3D domain, we do not rely on discrete representations such as meshes, point clouds or voxels. Instead, our architecture is formulated as parametric surfaces, a continuous representation for outfits.
- **Differential geometry.** Since our predictions are continuous, we are able to leverage the differentiability of the parametric surfaces to guide the learning. More specifically, we propose a novel surface smoothness loss term based on this property.

II. Related Work

RGB to 3D garment reconstruction. Prior work based on parametric 3D body models encoding shape and pose deformations separately, being learnt from thousands of scans of real people [11], [19]. These body models provide a good prior for 3D garment reconstruction. However, these models are trained to just capture the human body. There are attempts to reconstruct clothed body from video inputs [3], [2], RGB-D data [38] and multi-view images [6], [36]. Although, in these approaches, richer inputs clearly provide more information than a single image, the developed pipelines yield additional setup/hardware constraints and extra computation, limiting applicability. Recently, new approaches based on deep learning [34], [32], [29], [28], [1], [4], [40], [18] addressed single-view dressed body estimation. However, for all these methods, heavy manual post-processing is needed to extract the clothing surface from the reconstructed result. Furthermore, the reconstructed garments still lack realism. Closer to our work, there are few approaches that propose to reconstruct garment as a layer separated from the body. DeepGarment [10] proposes to use physics based simulations as supervision for learning a garment shape estimation model. However, it only works for seen garments and does not provide realistic results. Lehnar *et al.* [17] design a method to synthesize garment wrinkles onto a coarse garment mesh following a given pose. This method, however, needs a computationally demanding step to register the template cloth to the captured 4D scan. Additionally, the method is limited to a fixed topology and cannot scale well to large deformations. Multi-Garment Net [6] learns per-category garment reconstruction from images using 3D scanned data. However, this method typically requires 8 input RGB images

This work has been partially supported by the Spanish project PID2019-105093GB-I00 (MINECO/FEDER, UE) and CERCA Programme/Generalitat de Catalunya, by ICREA under the ICREA Academia programme, and by Amazon Research Awards.

and fails to reconstruct complex clothing topology such as skirts and dresses.

RGB to 3D garment and texture reconstruction. Some recent works [18], [4], [17] propose to use 2D UV map representation for estimating geometry and color details. Particularly, the Tex2Shape method of Alldieck *et al.* [4] aims to reconstruct high quality 3D geometry by regressing displacements in an unwrapped UV space. Nevertheless, this type of approach is limited by the topology of the template mesh (need of different mesh topology for skirts and dresses) and the topology of the UV parametrization (e.g. visible seam artifacts around texture seams). Other works [33], [32] propose volumetric voxel representations for colored 3D reconstruction. For example, Im2Avatar [32] performs textured single-image reconstruction, using colored voxels as the output representation. Other approaches [29], [23], [28] use implicit functions representation to recover shape and texture of clothed human bodies. Unlike explicit representations (e.g. meshes, voxels, point clouds) these methods learn functions to parametrize a 3D volume or surface. For instance, PiFu [29] learns an implicit surface function based on aligned image features. This model generates clothes details but does not predict a realistic texture of the occluded regions of the dressed person (e.g. back of the person). Also, it is less robust to pose variations. Here, we propose a novel architecture that exploits the shape and topology of the human body to explicitly predict analytical 3D surfaces as garments from still images with a differentiable geometry. In addition, we show that it can be directly applied to color prediction as well.

III. Learning deep parametric surfaces

The goal of this work is to reconstruct 3D outfits from RGBA images (RGB plus alpha channel for transparency). Thus, given an image $\mathbf{I} \in \mathbb{R}^{H\times W\times 4}$, we want to predict a 3D outfit as a mesh $M = \{\mathbf{V}, \mathbf{F}\}$, where $\mathbf{V} \in \mathbb{R}^{N\times 3}$ are its vertices and $\mathbf{F} \in \mathbb{R}^{M\times 3}$ is its triangulated faces. We assume a segmentation mask for the subject is given. Our model consists on two parts, visual and 3D. For the visual part, we present CNN-PointNet, an approach inspired on PointNet [26], a successful digit classification on MNIST. We encode images as point clouds of high-level color features with spatial attributes. For 3D, we propose a novel approach to explicitly predict analytical 3D surfaces. Fig. 1 shows the proposed pipeline.

A. CNN-PointNet

Qi *et al.* [26] showed how PointNet can be applied to MNIST dataset for digit recognition. They represent binary images α as point clouds $B = \{(i, j) \mid \alpha_{ij} = 1\} \in \mathbb{R}^{|B|\times 2}$ with $|B| = \sum \alpha$. They encode each *point* independently through a MLP and aggregate point features by max pooling the point clouds. Inspired by [26], one can represent RGBA images as:

$$A = \{\mathbf{x}_i \in \mathbb{R}^F \mid \alpha_i > 0\}. \tag{1}$$

where, in a naive representation, $\mathbf{x}_i$ is a feature array containing the 3 image channels and image coordinates for the i-th pixel. Here, α_i corresponds to the alpha channel. Then with a high capacity PointNet architecture, it is possible to obtain per point image embeddings for posterior 3D reconstruction. Nonetheless, it is more efficient and effective to leverage the image data structure and obtain higher-level color and spatial feature arrays prior to the point cloud encoding. Thus, we apply two convolutional branches: one with 3×3 filters on the RGB image and the other with 1×1 filters on the XY pixel coordinate image. We use a stride of 1 and no pooling between layers. This architecture allows to obtain per-pixel feature arrays. We latter concatenate the two branches and apply Eq. 1 to obtain the image point cloud A. Finally, we apply an additional fully connected layer per point followed by a max pooling across point clouds (as PointNet). The output of this part is an image embedding $U \in \mathbb{R}^F$. CNN-PointNet contains a much lower number of parameters than conventional CNNs ($\sim$ 1M vs. $\sim$ 100M), leverages the provided image mask and, in our experiments, shows better performance and faster convergence. Note we assume camera calibration parameters and subject distance to the camera are known. Thus, we normalize the pixel coordinates before feeding the network as:

$$\mathbf{p}' = d_x \cdot (\mathbf{p} - \mathbf{j}_0)/w, \tag{2}$$

where $\mathbf{p}$ are the pixel coordinates, $\mathbf{j}_0$ is the human skeleton root joint image plane coordinates, d_x is the distance between the camera and the root joint and $w = 640$ is the image width.

B. Predicting analytical surfaces

Assume a trained MLP f able to perfectly map image embedding $\mathbf{U}$ to infinitely dense outfit point cloud $\mathbf{Y}' \in \mathbb{R}^{3N}$ (with $N \to \infty$). The last layer of this MLP is a linear transformation as $\mathbf{Y}' = \mathbf{W}\mathbf{X}$ where $\mathbf{X} \in \mathbb{R}^F$ is the input feature array to the last layer and $\mathbf{W} \in \mathbb{R}^{3N\times F}$ is its weight matrix. By reshaping and reordering axes, we can interpret the matrix $\mathbf{W}$ as an infinite set of $\mathbf{W}_i \in \mathbb{R}^{3\times F}$ sub-matrices. Then, vertex i is given by $\mathbf{v}_i = \mathbf{W}_i\mathbf{X}$. Thus, for two output vertices $\mathbf{v}_i$ and $\mathbf{v}_j$ infinitesimally close to each other we have:

$$\epsilon \geq |\mathbf{v}_i - \mathbf{v}_j| = |\mathbf{W}_i\mathbf{X} - \mathbf{W}_j\mathbf{X}| \rightarrow \epsilon' \geq |\mathbf{W}_i - \mathbf{W}_j|, \tag{3}$$

which means that submatrices $\mathbf{W}_i$ form a continuous 2-manifold on the weights space $\mathcal{W} \subset \mathbb{R}^{3\times F}$. We propose to parametrize this manifold instead of learning each $\mathbf{W}_i$, as it would be standard with a MLP. To this end, we need to define a parametric subspace P as a 2-manifold and learn a function as $f_P : P \rightarrow \mathcal{W}$ such that:

$$P = \{\mathbf{p}_i \mid \mathbb{E}_i[\mathbf{Y}] = \|\mathbf{Y}_i - f_P(\mathbf{p}_i)\mathbf{U}\| = 0\}, \tag{4}$$

where f_P is the MLP mapping from $\mathbf{p}_i$ to $\mathbf{W}_i$, $\mathbf{Y}_i$ is a 3D point on the ground truth outfit surface and $\mathbb{E}_i$ is the expected error. This allows training the model as a point cloud predictor while implicitly enforces these predictions to follow a 2-manifold in the $\mathbb{R}^3$ space. It also removes the need of an increase on network parameters to predict denser point

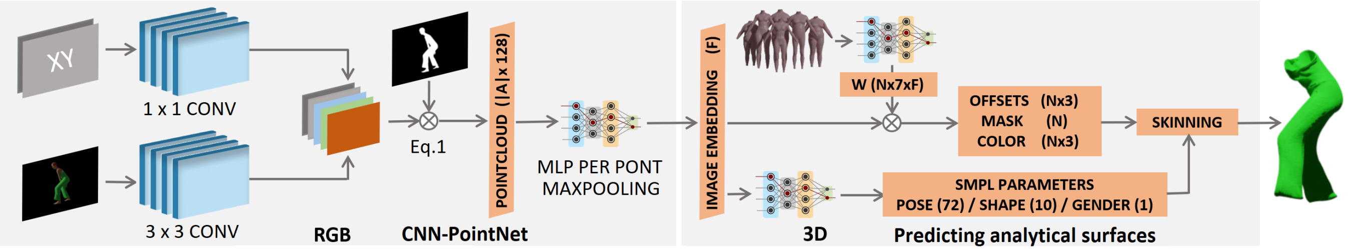

Fig. 1: Model pipeline. Dimensions are shown in brackets. Left part: RGB channels are convolved to obtain per-pixel high-level features. Normalized pixel coordinates (XY) are processed by 1×1 convolutions to increase the dimensionality. The output of both branches is concatenated and encoded as a point cloud following Eq. 1. We apply an additional fully connected layer per point to the concatenation of visual and spatial features before max pooling the whole *point cloud*. This yields a global image embedding. Right part: an MLP extracts submatrices $\mathbf{W}$ from the parametric subspace (human body) as explained in Sec. III-B. These submatrices map the image embedding to the offsets connecting skin-to-cloth for each body point sampled. The same approach estimates the mask (for different garments) and color. Finally, with estimated SMPL parameters we obtain final predictions by skinning the body with offsets and mask.

clouds. One can train the model with point clouds of size N and predict point clouds or meshes of size $K >> N$ during inference by sampling more points from P. To the best of our knowledge, we are the first to propose a parametrized fully connected layer to predict analytical surfaces. Note that we omit biases for simplicity.

Parametric Subspace. From Eq. 3 we can further conclude that the submatrix space also has the same topology as the predicted garments. Therefore, it is important to choose a parametric subspace P that also follows this topology. In the current literature we often see works that rely on the assumption that cloth follows body topology [4], [5], [6], [17], [20], [25], [30], [37]. While this might not be accurate for all possible garments, reported results show it is a valid assumption that simplifies the problem. Another trend in this domain is to learn garments in rest pose and later leverage SMPL skinning to obtain posed garments. We define the parametric subspace P as the body surface in rest pose and learn the mapping of each skin point to the submatrix that yields the corresponding unposed garment vertex location for a given image embedding. We further simplify the problem by predicting an offset from $\mathbf{p}_i$ (body surface) to the garment instead of absolute vertex locations.

Point sampling. For each training step, we sample a different set of points $\mathbf{p}_i$ from the body surface P. Given a triangulated body mesh, this is done by randomly generating valid barycentric coordinates $\omega \in \mathbb{R}^{N\times3}$ and sampling N face indices from a weighted distribution proportional to each face area. These weights ω allow interpolation of any vertex property: location, normals, UV map coordinates and blend weights (for posterior skinning), among others. This standard interpolation is computed for the i-th face as:

$$\mathbf{p}_i = \sum_j^3 \omega_{ij} \cdot \mathbf{v}_{t_{ij}}, \tag{5}$$

where $t \in \mathbb{Z}^{N\times3}$ is the set of triangular face indices and $\mathbf{v}$ is the set of body mesh vertices (location plus extra properties). In practice, we want to use different body shapes, and therefore, sampling vertex location alone can be ambiguous. Different surface points of different body shapes can share location, and will therefore be mapped to the same submatrix. To help the network resolve this ambiguity, we concatenate point location ($\mathbb{R}^3$), surface normal ($\mathbb{R}^3$) and blend weights ($\mathbb{R}^{24}$ which can be understood as an skeleton based coordinate system). Therefore, the parametric subspace P is defined as a 2-manifold in the $\mathbb{R}^{30}$ space. We remove head, hands and feet from the subspace prior to the sampling. The sampled blend weights are later used for skinning garments along the body. This sampling strategy yields a matrix $\mathbf{P} \subset P$.

Body mask and color. Following the methodology explained above, a garment surface is sampled along the whole body. This forces the network to learn non-homogeneous large offsets from body areas far from real outfit surface, for instance by sampling from lower leg to predict a short skirt. This causes a slower convergence and noisy predictions. We propose overcoming this by learning a continuous mask through the body surface conditioned on the image embedding. The presented methodology can be adapted to predict a single value m_i representing a mask for each $\mathbf{p}_i$. As ground truth for this mask, we compute 2 nearest neighbours from the whole outfit in rest pose (provided in the dataset) against all sampled body points $\mathbf{p}_i$. Then, $m_i = 1$ if matched with an outfit vertex, otherwise, $m_i = 0$. Note this works best if outfit vertices outnumber points $\mathbf{P}$.

Finally, we want to also predict textured/colored garments. Once again, we follow the same methodology to predict RGB color $\mathbf{c}_i \in \mathbb{R}^3$ for each point $\mathbf{p}_i$. Ground truth is obtained by assigning to each $\mathbf{p}_i$ the color of the closest point in the outfit in rest pose (garment UV map and texture/color provided).

SMPL Skirt. For skirts and dresses, this approach fails, since skirt-like garments do not closely follow the body topology. We propose to use a different parametric subspace for these garments. We append a skirt-like shape to the waist of SMPL model and assign blend weights such that it follows root joint orientation (legs relative motion w.r.t. skirt is too noisy). Fig. 2 shows the proposed SMPL modification. In this case, we remove the legs before sampling (plus head

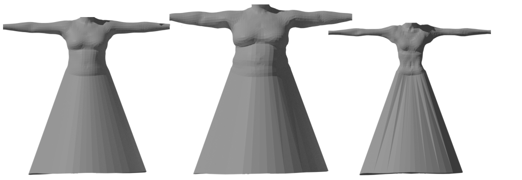

Fig. 2: Proposed SMPL modification. Skirts and dresses do not follow body topology. We deal with this by using a modified body as the parametric subspace for skirts/dresses.

and hands). During training we choose body topology (skirt-like or not) based on ground truth outfit type. To be able to apply this strategy at inference time, we train a binary topology classifier (skirt-like or not) on the input RGBA sample (details in Sec. IV-B).

Final prediction. Combining the predictions for offsets, mask and RGB color yields unposed textured outfits. To compute the final vertex locations we need to predict as well SMPL parameters. We do so by an standard regression from image embeddings $\mathbf{U}$ learnt with an L2 loss on predicted SMPL joints and body mesh vertices. With an estimation of pose, shape and gender, we perform an skinning w.r.t. SMPL skeleton for the predicted garment points.

C. Training

Training can be tackled as a point cloud reconstruction problem. We use the point sampling strategy explained before to sample not only from the body as parameters $\mathbf{P}$, but also from the ground truth garment meshes in order to obtain uniform 3D labels, to ease the implementation. Therefore, the inputs of the network consist of an image $\mathbf{I} \in \mathbb{R}^{w \times h \times 5}$, as a result of concatenating RGB channels with normalized pixel coordinates (Eq. 2), an image mask $\mathbf{M} \in \{0,1\}^{w \times h}$, obtained by binarizing the alpha channel, and the parametric points $\mathbf{P} \in \mathbb{R}^{N \times 30}$. From $\mathbf{I}$ and $\mathbf{M}$, we extract the image embedding $\mathbf{U} \in \mathbb{R}^{F}$, and from the parameters $\mathbf{P}$ we obtain the set of $\mathbf{W}_{XYZ} \in \mathbb{R}^{N \times 3 \times F}$ for vertex locations, $\mathbf{W}_{M} \in \mathbb{R}^{N \times 1 \times F}$ for the predicted body surface mask or $\mathbf{W}_{RGB} \in \mathbb{R}^{N \times 3 \times F}$ for point color. We train a different CNN to obtain image embeddings specifics for each task.

In the current literature we find Chamfer Distance loss (CD) and Earth Moving Distance (EMD) as valid candidates for 3D garment prediction. It is showed EMD abilities to predict more uniform point clouds with less outliers thanks to the one-to-one correspondence under which it is formulated. Nonetheless, due to the construction of our model as a parametric 3D surface predictor and the uniform sampling in the parametric subspace, CD already achieves uniform predictions and no outliers. Furthermore, CD implementation is much simpler and allows masking out distances as required. We choose a masked CD to train our model as:

$$\mathcal{L}_{CD} = \sum_{\mathbf{y}'_i \in \mathbf{Y}'_M} \min_{\mathbf{y}_j \in \mathbf{Y}} \|\mathbf{y}'_i - \mathbf{y}_j\|_2^2 + \sum_{\mathbf{y}_j \in \mathbf{Y}} \min_{\mathbf{y}'_i \in \mathbf{Y}'_M} \|\mathbf{y}'_i - \mathbf{y}_j\|_2^2 , \tag{6}$$

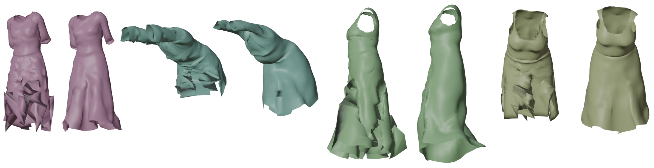

Fig. 3: Our proposed smooth loss leverages the differential geometry of the predicted surfaces to enforce smoothness by penalizing its second derivatives. Left: before smoothness, right: after smoothness.

where $\mathbf{y}'_i \in \mathbf{Y}'_m$ is the i-th predicted point after masking, $\mathbf{y}_j \in \mathbf{Y}$ is the j-th ground truth point. This loss will be back-propagated through the skinning performed to obtain $\mathbf{Y}'$, which allows implicit learning of unposed deformations. To learn mask and color, ground truth is computed for each point $\mathbf{p}_i \in \mathbf{P}$, and therefore the model can be trained with regular L1 for the mask and L2 for the color.

Due to higher cloth dynamics complexity of skirt-like garments, CD becomes sensitive to incorrect matches between ground truth garments $\mathbf{Y}$ and predictions $\mathbf{Y}'$. Empirically, we observed noisy and highly distorted predictions. We regularize this behaviour by leveraging the analytical nature of our predictions. Given a fixed image embedding $\mathbf{U}$, predicted garments are parametrized surfaces as $\mathbf{Y}'_{\mathbf{U}} = \{\mathbf{y}'_{\mathbf{U}}(\mathbf{p}) = f_{XYZ}(\mathbf{p})\mathbf{U} \quad \forall \mathbf{p} \in \mathbf{P} \mid f_M(\mathbf{p})\mathbf{U} > 0\}$, where $f_{XYZ} : \mathbf{p} \rightarrow \mathbf{W}_{XYZ}$ and $f_M : \mathbf{p} \rightarrow \mathbf{W}_M$. In Eq. 5 we describe $\mathbf{p}$ as a function of barycentric weights ω, therefore, $\mathbf{y}'_{\mathbf{U}}(\mathbf{p}(\omega)) = \mathbf{y}'_{\mathbf{U}}(\omega)$. Predictions have a differential geometry since $\mathbf{y}'_{\mathbf{U}}(\omega)$ is differentiable almost everywhere. This allows us to define the following surface regularization term:

$$\mathcal{L}_{smooth} = \sum_{\mathbf{P}}^{\mathbf{P}} \frac{\partial^2 \mathbf{y}'_{\mathbf{U}}}{\partial \omega^2}(\mathbf{p}), \tag{7}$$

which enforces locally smooth predictions, and implicitly yields more uniform surfaces. This loss term discourages the network from generating highly distorted garments to minimize CD. Differential geometry is a significant novelty that appears as a direct result of our proposed formulation. We apply $\mathcal{L}_{CD} + 0.01\mathcal{L}_{smooth}$ as final loss.

IV. Experiments

In this section we explain the results achieved through the proposed methodology.

A. Dataset

We evaluate our approach on CLOTH3D++ dataset [21]. This dataset is created by realistic renderings of CLOTH3D [5] (purely 3D sequences of dressed human performing actions) through an unbiased ray-tracing engine and PBR materials for the different fabrics. Additionally, garments are randomly textured or colored to present high visual variability. Renderings are given as RGBA videos, where alpha channel can be used as masks for the subject in the video. CLOTH3D++ contains around 8000 different subjects and oufits, and over 2M different frames and poses.

TABLE I: Comparison of standard CNN and proposed CNN-PointNet for the RGB part.

	N. params	Epoch time	CD loss
VGG-16	208M	4h	0.046
Ours	1M	1.5h	0.029

For evaluation, we follow the protocol presented by the authors. We choose CLOTH3D++ due to its size and variability in terms of outfits and garment topologies. Additionally, CLOTH3D++ provides of a segmentation mask as the alpha channel.

B. Setup

We train our model with an initial batch size of 8 samples incremented by a factor of 2 every two epochs up to 256. We train the network for 10 epochs with an initial set of 50000 samples. Afterwards, we finetune during 2 more epochs using all the data and a batch size of 256. To train the offsets and color, we use the obtained ground truth body mask to compute the loss. This speeds up convergence. For the offsets we additionally provide ground truth SMPL parameters during training. Otherwise, the model tries to correct the error due to pose and shape through these offsets, which yields unrealistic predictions on test time. As mentioned in Sec. III-B, skirt-like garments have a different parametric subspace than trousers. Therefore, we train two different models for these two garment types (skirt-like or not). Then, at inference time, to select the corresponding trained model, we apply an independent classifier to identify skirt and dresses from single RGB images. To do so, we train PSPNet [39] to semantically segment the garments and body. Afterwards, to classify garments, we average estimated garment probabilities per pixel, multiply them with garment pixel ratio and sort them. Then we select winner classes based on a top-down strategy, e.g. a jumpsuit and dress can not co-occur. By this strategy, 2-class garment classification is predicted by 97% accuracy at inference time.

C. Evaluation metric

We evaluate predicted outfit surface using Surface-to-Surface distance (S2S), an extension of CD as proposed in [21]. It is computed based on the nearest face rather than nearest vertex.

We also use the following metrics: average per vertex Euclidean error in mm for SMPL body surface, F1 score for classification of six garment categories, and the percentage of correctly selected vertices for outfit vertex mask.

D. Ablation study

CNN-PointNet. To show the benefits of the proposed CNN-PointNet for RGB, we train a model to learn to regress a $N = 2000$ garment point cloud with a standard CNN and CNN-PointNet. For the standard CNN we choose a VGG-16 and apply the image alpha channel at the feature volume of each convolutional block. Tab. I shows a much higher CNN-PointNet performance after 10 epochs regarding number of parameters, training time and CD reconstruction loss.

TABLE II: Ablation study. S2S error per garment shows smoothness regularization greatly improve the results on all garment categories. We additionally include results obtained with CLOTH3D++ baselines[21] and SMPLicit[9] for comparison. Errors in millimeters.

	Top	T-shirt	Trousers	Jumpsuit	Skirt	Dress	All
Ours	8.0	12.4	10.5	10.3	9.5	8.7	9.9
Ours (Eq. 7)	6.3	11.4	9.2	8.5	8.1	7.5	8.6
Baseline[21]	23.9	43.3	40.6	22.1	32.8	29.3	31.3
SMPLicit[9]	41.4	36.2	31.8	-	68.1	-	42.9

Smoothness loss. Fine-tuning our models with the smoothness term $\mathcal{L}_{smooth}$ greatly improves the quality of the predicted surface for all garments. One can see the impact of the smoothness regularization per garment category in Tab. II. Fig. 3 shows the qualitative improvement on two different validation samples after just a few tens of training steps. The effect of this loss is specially visible for long dresses and skirts.

Tab. III shows ablation results for SMPL body surface, mask prediction accuracy, garment classification, and effect of Eq. 7 regularizer on surface reconstruction quality. Next, we analyse these results on different subsets of the test set: seen vs. unseen pose, texture pattern vs. plain color.

Body surface error (2nd row in Tab. III). Our simple SMPL pose and shape parameter estimation from embedded parametric subspace has an error of 80.6mm on body surface vertices, being slightly higher for unseen poses and texture patterns compared to seen poses and plain color, respectively.

Mask and garment classification accuracy (3rd-4th rows in Tab. III). Mask and garment classification based on semantic segmentation and vertex masking show good performance, specially on seen pose and plain color subsets.

Outfit surface error (5th-6th rows in Tab. III). Outfit surface error is more dependent on the body pose than image texture. The S2S error on unseen poses is significantly higher than seen poses, while the error is slightly higher on texture patterns than plain color. The use of the smoothing regularizer consistently reduces the error in all cases.

Qualitative. Fig. 4 shows predictions from our model. Each of them corresponds to a single RGBA frame. While the models was trained with just a few thousands of points per sample, these visualizations contain tens of thousands of points (we can sample any arbitrary amount of points during inference). We observe that the main limitation of our model is in the capturing of colored patterns. As it can be seen in the figure, our model is able to predict mainly plain colors. Nonetheless, it shows to be able to predict correctly different colors for upper and lower body garments. We take advantage of this to split predictions into upper and lower by thresholding the gradient of the color w.r.t. the sampled barycentric weights (Eq. 5). That is, we complement the predicted mask with an additional one as $\mathbf{M}' = \{\mathbf{p} \in \mathbf{P} \mid \partial \mathbf{c}_{\mathbf{U}} / \partial \omega < \epsilon\}$, where $\mathbf{c}_{\mathbf{U}}$ is the parametric expression of the color for a given image embedding $\mathbf{U}$. This process removes from the prediction those points where there

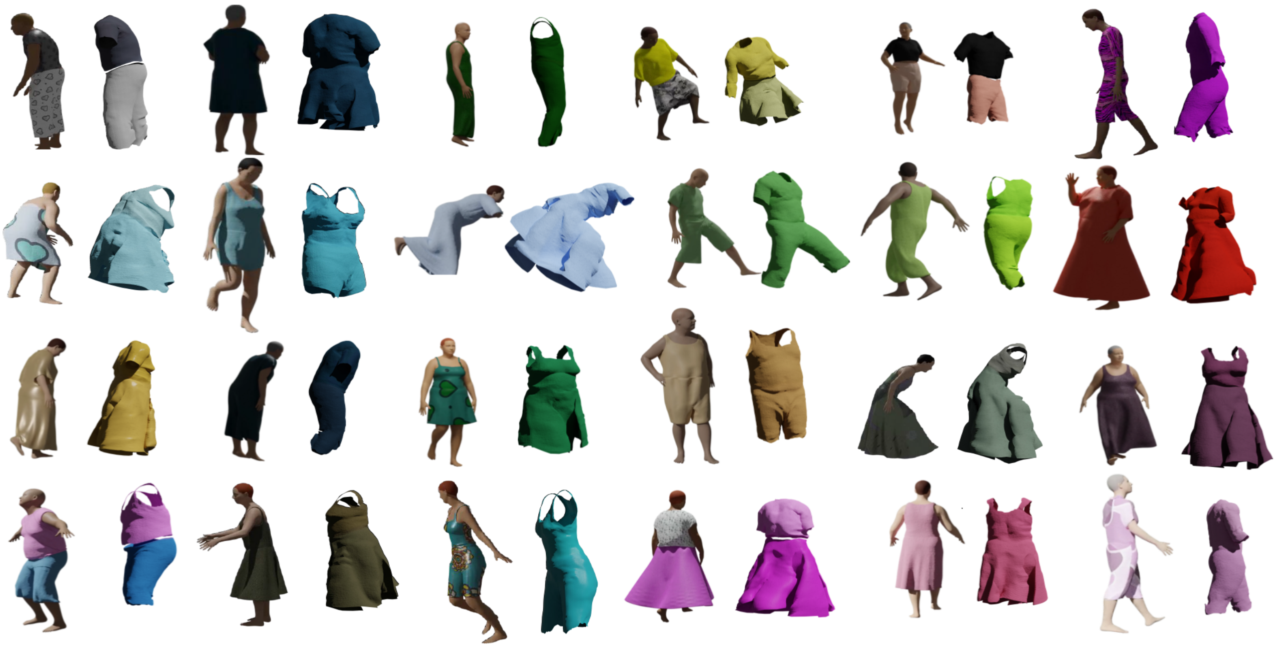

Fig. 4: Qualitative results of our model from single frames. For each sample, we show the RGBA frame (left) and the 3D prediction with color (right). Subjects size is proportional to its corresponding size in the RGBA frame in pixels.

TABLE III: Ablation results for SMPL body surface, mask prediction accuracy, garment classification, and effect of Eq. 7 regularizer on surface reconstruction quality.

	Seen pose	Unseen pose	Texture pattern	Plain color	All
Body surface error (mm)	78.1	88.4	81.5	78.8	80.6
Mask accuracy (%)	95.2	86.7	92.0	94.4	93.1
Garment classification accuracy (F1)	0.92	0.86	0.89	0.97	0.90
Outfit surface error w/o regularization (S2S, mm)	8.4	14.8	10.0	9.7	9.9
Outfit surface error with regularization (S2S, mm)	7.0	13.2	8.5	8.2	8.6

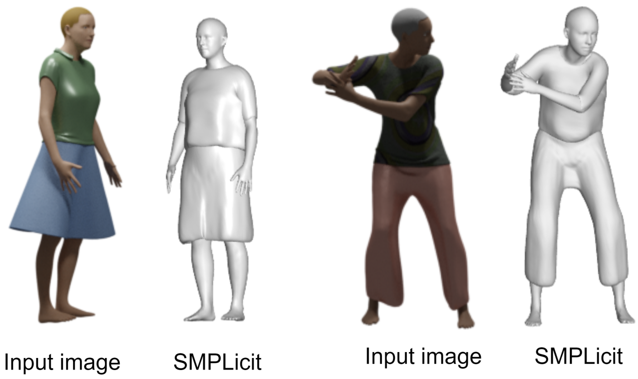

Fig. 5: Results obtained with SMPLicit. While predictions are consistent in terms of pose and garment type, we observe significant differences in terms of body shape, garment size and geometry.

is a sudden change in color.

E. State-of-the-art comparison

We first compare our approach against the baselines proposed by CLOTH3D++ authors[21]. Results are presented in Tab. II in a per-garment type fashion. As observed, our approach shows a significant improvement w.r.t. the CLOTH3D++ baselines. We believe most of this improvement comes from a more accurate topology prediction, which has a high impact on the S2S metric. Also, due to the use of a specific model and rigging for the skirt. While CLOTH3D++ baseline attaches the skirts to the legs, producing distorted artifacts in the predictions.

We also compare our approach against the work of [9]. Similar to our methodology, SMPLicit allows handling a great variability of garments with different topologies. To this end, they propose the use of implicit surfaces for garment representation, inspired in [24]. Similarly, when applied for inference to images, a search for a latent code –from which the implicit surface is computed– is performed as an optimization problem. Note how direct inference already has advantages over optimization based problems in terms of efficiency and applicability. Moreover, implicit surfaces require marching cubes algorithm to extract the predictions as a 3D mesh, this further increases computational cost. Due to the computational overhead, computing predictions for CLOTH3D test set using SMPLicit is unfeasible, thus, we do this for a representative subset of data. Tab. II presents the quantitative results obtained. As observed, the error for the predictions obtained with SMPLicit is larger than our methodology. Fig. 5 depicts qualitative results to better understand the numerical error. As can be seen, SMPLicit predictions are accurate in terms of pose and garment types. Nonetheless, garment size and geometry do not fully correspond to the image. Furthermore, seems like SMPLicit is limited to overly smoothed surfaces. Also, note how our approach allows predicting color as well and it is potentially scalable to other surface properties. Nonetheless, on the other hand, our approach is designed for images with alpha channel as input, while SMPLicit does not have this requirement.

V. Conclusions

We presented a novel architecture for RGBA to 3D garment reconstruction. We proposed a parametrization of a fully connected layer to explicitly predict analytical surfaces. We showed how this can be beneficial during training by exploiting the differential geometry of the predictions. This also allows predictions as dense as desired at inference time. Compared to other approaches, the proposed method has less computational requirements, being the number of parameters of the network decoupled from the dimensionality of the predictions. We also showed how our approach can be used for color prediction, albeit it cannot capture complex color patterns. We believe this to be a possible future research line.

References

[1] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2019.

[2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018.

[3] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. CVPR Spotlight Paper.

[4] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2293–2303, 2019.

[5] H. Bertiche, M. Madadi, and S. Escalera. Cloth3d: Clothed 3d humans. In *European Conference on Computer Vision*, pages 344–359. Springer, 2020.

[6] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5420–5430, 2019.

[7] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016.

[8] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016.

[9] E. Corona, A. Pumarola, G. Alenya, G. Pons-Moll, and F. Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11875–11885, 2021.

[10] R. Danerek, E. Dibra, A. C. Öztireli, R. Ziegler, and M. Gross. Deepgarment : 3d garment shape estimation from a single image. *Computer Graphics Forum*, 36, 2017.

[11] A. Dragomir, S. Praveen, K. Daphne, T. Sebastian, R. Jim, and D. James. Scape: shape completion and animation of people. *ACM Trans. Graphics*, 24, july 2005.

[12] B. Egger, W. A. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020.

[13] H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017.

[14] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black. Drape: Dressing any person. *ACM Trans. Graph.*, 31(4):35–1, 2012.

[15] E. Gundogdu, V. Constantin, A. Seifoddini, M. Dang, M. Salzmann, and P. Fua. Garnet: A two-stream network for fast and accurate 3d cloth draping. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019.

[16] B. Jiang, J. Zhang, Y. Hong, J. Luo, L. Liu, and H. Bao. Bcnet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision*, pages 18–35. Springer, 2020.

[17] Z. Lahner, D. Cremers, and T. Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 667–684, 2018.

[18] V. Lazova, E. Insafutdinov, and G. Pons-Moll. 360-degree textures of people in clothing from a single image. *2019 International Conference on 3D Vision (3DV)*, pages 643–653, 2019.

[19] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):248:1–248:16, oct 2015.

[20] Q. Ma, S. Tang, S. Pujades, G. Pons-Moll, A. Ranjan, and M. J. Black. Dressing 3d humans using a conditional mesh-vae-gan. *arXiv preprint arXiv:1907.13615*, 2019.

[21] M. Madadi, H. Bertiche, W. Bouzouita, I. Guyon, and S. Escalera. Learning cloth dynamics: 3d + texture garment reconstruction benchmark. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track, PMLR*, volume 133, pages 57–76, 2021.

[22] M. Madadi, H. Bertiche, and S. Escalera. Smplr: Deep learning based smpl reverse for 3d human pose and shape recovery. *Pattern Recognition*, 106:107472, 2020.

[23] M. Oechsle, L. M. Mescheder, M. Niemeyer, T. Strauss, and A. Geiger. Texture fields: Learning texture representations in function space. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4530–4539, 2019.

[24] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019.

[25] C. Patel, Z. Liao, and G. Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7365–7375, 2020.

[26] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[27] Y. Rong, T. Shiratori, and H. Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020.

[28] N. Ryota, S. Shunsuke, H. Zeng, C. Weikai, M. Chongyang, L. Hao, and M. Shigeo. Siclope: Silhouette-Based Clothed People. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[29] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[30] I. Santesteban, M. A. Otaduy, and D. Casas. Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum*, volume 38, pages 355–366. Wiley Online Library, 2019.

[31] R. Shoja Ghiass, O. Arandjelović, and D. Laurendeau. Highly accurate and fully automatic 3d head pose estimation and eye gaze estimation using rgb-d sensors and 3d morphable models. *Sensors*, 18(12):4280, 2018.

[32] Y. Sun, Z. Liu, Y. Wang, and S. E. Sarma. Im2avatar: Colorful 3d reconstruction from a single image. *arXiv preprint arXiv:1804.06375*, 2018.

[33] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 209–217, 2017.

[34] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and

C. Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[35] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67, 2018.

[36] Y. Xu, S. Yang, W. Sun, L. Tan, K. Li, and H. Zhou. 3d virtual garment modeling from rgb images. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 37–45, 2019.

[37] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Analyzing clothing layer deformation statistics of 3d human motions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 237–253, 2018.

[38] T. Yu, Z. Zheng, Y. Zhong, J. Zhao, Q. Dai, G. Pons-Moll, and Y. Liu. Simulcap: Single-view human performance capture with cloth simulation. *arXiv preprint arXiv:1903.06323*, 2019.

[39] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[40] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. Deephuman: 3d human reconstruction from a single image. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.