Deep Self-supervised UV Parameterization for Template-free Textured 3D Digitization of Garments

Shanthika Naik *
IIIT Hyderabad
shanthika.naik@research.iiit.ac.in

Chandradeep Pokhariya *
IIIT Hyderabad
chandradeep.pokhariya@research.iiit.ac.in

Astitva Srivastava
IIIT Hyderabad
astitva.srivastava@research.iiit.ac.in

Avinash Sharma
IIIT Hyderabad
asharma@iiit.ac.in

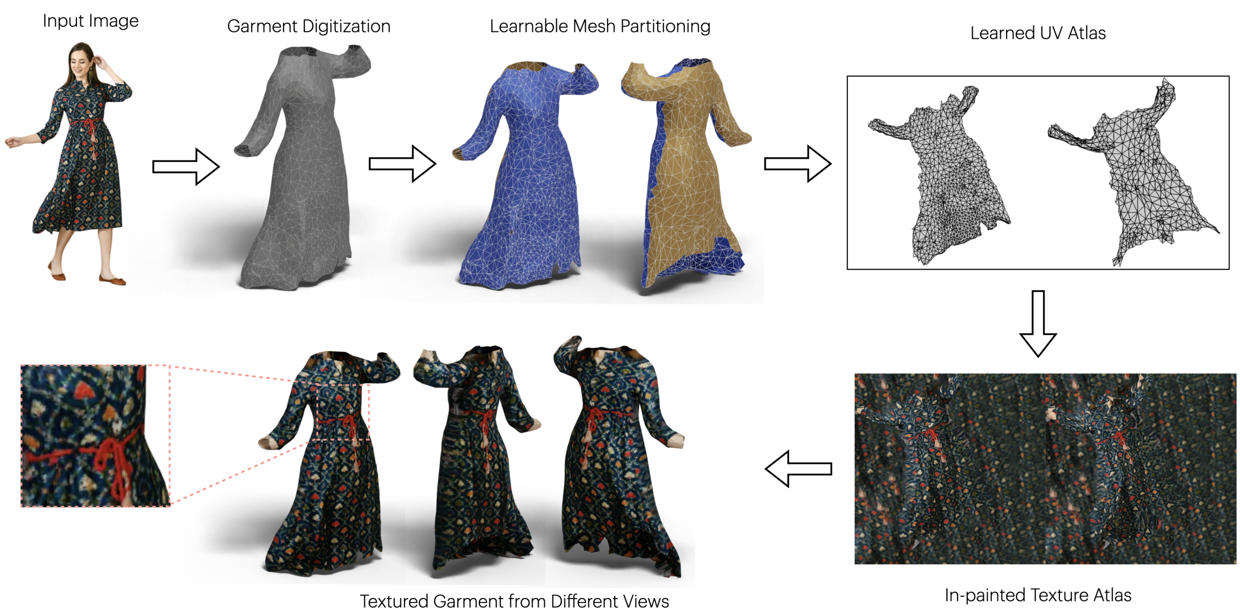

Figure 1: 3D template-free textured digitization of loose garment obtained with our framework from a single RGB image.

Abstract

We propose a novel framework for template-free textured 3D garment digitization from RGB image(s). We employ implicit geometry modeling and novel diffusion-guided garment mesh extraction & partitioning followed by a novel, self-supervised neural UV parametrization, and texture filling to obtain mesh and associated texture atlas representation of garments. The key novelty of our method is the learnable discretization-agnostic UV parametrization of arbitrary garment surfaces along with automated partitioning of garments. Existing approaches for UV parameterization either fail to parameterize the bounded surface because of noisy mesh or are time-consuming in iterative optimization. Because of the optimization behaviour, the loss function needs to be minimized again, even if the minimal discretization of the surface is changed. We overcome these issues by learning a neural representation of the 3D to a 2D mapping function. We also show that, with diffusion-based global embedding, the parameterization can be generalized to unseen garment shapes. We perform a comprehensive

*These authors contributed equally to this work.

empirical analysis on the publicly available datasets and report superior results to existing methods for garment extraction, as well as UV parametrization.

1. Introduction

3D digitization of garments from single or multi-view images is a challenging research problem with key applications in domains such as fashion e-commerce, gaming, and multi-media. This problem is typically perceived as closely related to another well-attempted problem of human body digitization from image(s) [12, 19, 8]. Modelling the human body is not straightforward due to challenges like varying articulation, self-occlusions, depth-ambiguity etc. However, unlike the human body, which has a more-or-less similar topology across different subjects, garment modelling poses even tougher challenges as different clothes have arbitrary topologies and styles owing to inherent artistic ideation in the designs. Additionally, the underlying cloth surface is subject to non-rigid deformations, which vary according to associated material attributes, stitching patterns, pose & shape of the underlying body, and external environmental factors such as wind, illumination, etc. Garment designers rely on professional software solutions which employ traditional template-based 3D garment modelling. Hence, initial learning-based techniques naturally acclimate to template-based strategy for garment digitization [1, 20, 6, 21]. Apart from providing a strong geometrical prior for garments, a template also enables retaining a high-quality visual appearance by leveraging the associated fixed UV parameterization. However, template-based digitization is restricted to limited clothing styles and fails to generalize to the styles for which the template is unavailable; hence, it is not accustomed to new frontiers of artistic creativity and is not scalable in fast-fashion scenarios. A very recent work xCloth [16] attempted *template-free* textured reconstruction of arbitrary garments using monocular images. Nevertheless, their method is not generalizable to multi-view RGB images, and garment geometry highly depends on the fitting of the parametric SMPL [11] body prior (as they adopt [8] for 3D reconstruction). Moreover, their reconstruction method, the sparse PeeledHuman [7] representation, which typically yields noisy geometry due to per-pixel(ray) depth prediction and also requires significant post-processing, including Poisson hole-filling around peel-map boundaries. Another key limitation of their work is that the UV mapping is obtained using conventional, numerical optimization methods that are not computationally scalable to high-resolution meshes. Thus, we aim to address these limitations by proposing a single/multi-view implicit garment reconstruction and learnable UV parameterization on top.

Some of the existing efforts for learning the UV parameterization in a self-supervised setup attempt to model the surface using multiple patches [4, 17] resulting in many seams. A recent work [18] proposes to leverage the power of NeRF to generate an object-centric cube-map textured representation but works only for genus-0 surfaces and requires a significantly large number of multi-view images. Thus, it is not suitable for single or few-view reconstruction, and moreover, a basic cube-map representation is not ideal for modelling garments with arbitrary styles. A recent work in [2] attempts to learn UV parameterization in a self-supervised setup. However, it can only generalize to the objects of similar topology (e.g., human faces) oriented in the same poses, which doesn't hold for in-the-wild garment modelling. Please refer to the Supplementary material for a detailed literature review.

In this paper, we propose a novel framework for the template-free textured 3D digitization of arbitrary garments from single/multi-view images. We employ implicit geometry modelling and diffusion-guided semantic segmentation to extract and partition 3D garments. Subsequently, we introduce a novel, self-supervised neural UV parameterization method for arbitrary surfaces and apply it over the extracted garment partitions to obtain texture atlas representation, followed by texture filling. Figure 1 depicts the intermediate output of our proposed framework. The key novelty of our method is the learnable discretization-agnostic UV parameterization of arbitrary garment surfaces along with automated partitioning of garments. Existing approaches for UV parameterization either fail to parameterize the bounded surface because of noisy mesh or are time-consuming in iterative optimization. Hence, their loss function needs to be minimized again, even if the minimal discretization of the surface is changed. Our framework overcomes these issues by learning a neural representation of the 3D to 2D mapping function. We also show that, with the use of diffusion-based global embedding, the parameterization can be generalized to the unseen garment shapes. We perform a comprehensive empirical analysis on the publicly available datasets and report superior results to existing methods for cloth extraction, as well as UV parameterization.

2. Proposed Framework

Figure 2 provides an overview of the proposed framework consisting of three key modules. Given a single monocular RGB image $\mathcal{I}$, our "Garment Extraction" module reconstructs the mesh $\mathcal{M}$ and extracts the 3D garment $\mathcal{M}_{gar}$ by learning the semantic segmentation over the mesh surface. Subsequently, the extracted garment $\mathcal{M}_{gar}$ is then fed to our novel "Neural Surface Parameterization" module, which learns to parameterize the underlying surface. This is achieved by finding the underlying seams of the garment followed by neural parameterization, which is learned in a

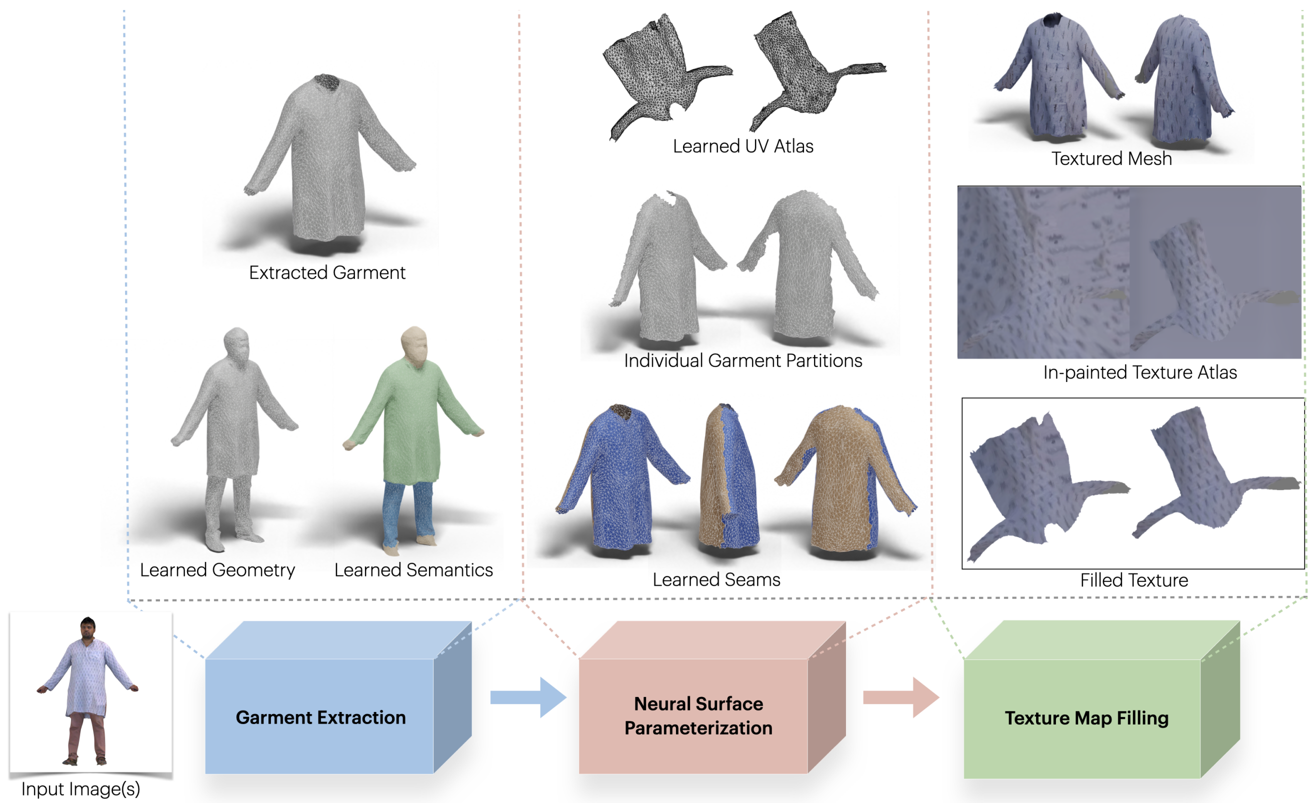

Figure 2: An overview of the proposed framework.

complete self-supervised setup. Finally, we use the pixel-aligned texture filling and in-painting in our last module "Texture Map Filling" to get the final texture map $\mathcal{T}$.

2.1. Garment Extraction

This module adopts a two-step approach to recover the 3D geometry of the garment. In the first step, we learn the implicit function to reconstruct 3D geometry of the clothed body from input RGB image(s) and then extract the garment using semantic understanding in the second step. For the first step, inspired by the popular implicit representation [12], we formulate the problem to estimate the continuous occupancy field in a 3D volume. Given enough training examples, the network learns to estimate the iso-surface of the underlying geometry, which can be extracted with the help of the marching cubes algorithm. This approach is helpful in generating smooth geometry, which is crucial for later stages. Once the geometry of a clothed human is reconstructed, we extract the garment by learning semantic segmentation over mesh for different cloth labels, namely "top-wear", "bottom-wear", and "body". [21] uses a similar approach for cloth extraction but learns the semantic field along with the geometry using implicit function learning. However, this approach in absence of a boundary field produces poor results for boundaries of segmentation as discussed in subsection 4.1. Thus, we propose to disentangle the learning of the semantic labels from geometric reconstruction by learning the segmentation directly over the mesh surface. For this task, we leverage the diffusion-guided propagation of features and hence train DiffusionNet [14] architecture as the underlying baseline and minimize $\mathcal{L}_{semantic}$.

$$\mathcal{L}_{semantic} = \sum_{s=1}^{S} \mathcal{P}^{s}_{semantic} \log(\widehat{\mathcal{P}}^{s}_{semantic}) \tag{1}$$

Here $\widehat{\mathcal{P}^s}_{semantic}$ is the probability of assigning the label to a pixel s whereas $\mathcal{P}^s_{semantic}$ is the ground truth label. We use $S = 3$ in our experiments.

2.2. Neural Surface Parameterization

Our main goal is to recover the textured garment of arbitrary topology, so UV parameterization of extracted garment mesh is necessary. This module consists of our novel self-supervised neural surface parameterization approach, where we first attempt to learn to estimate valid seems for partitioning the mesh into multiple regions, followed by self-supervised neural UV parameterization of each region. An outline of the proposed approach is shown in Figure 3.

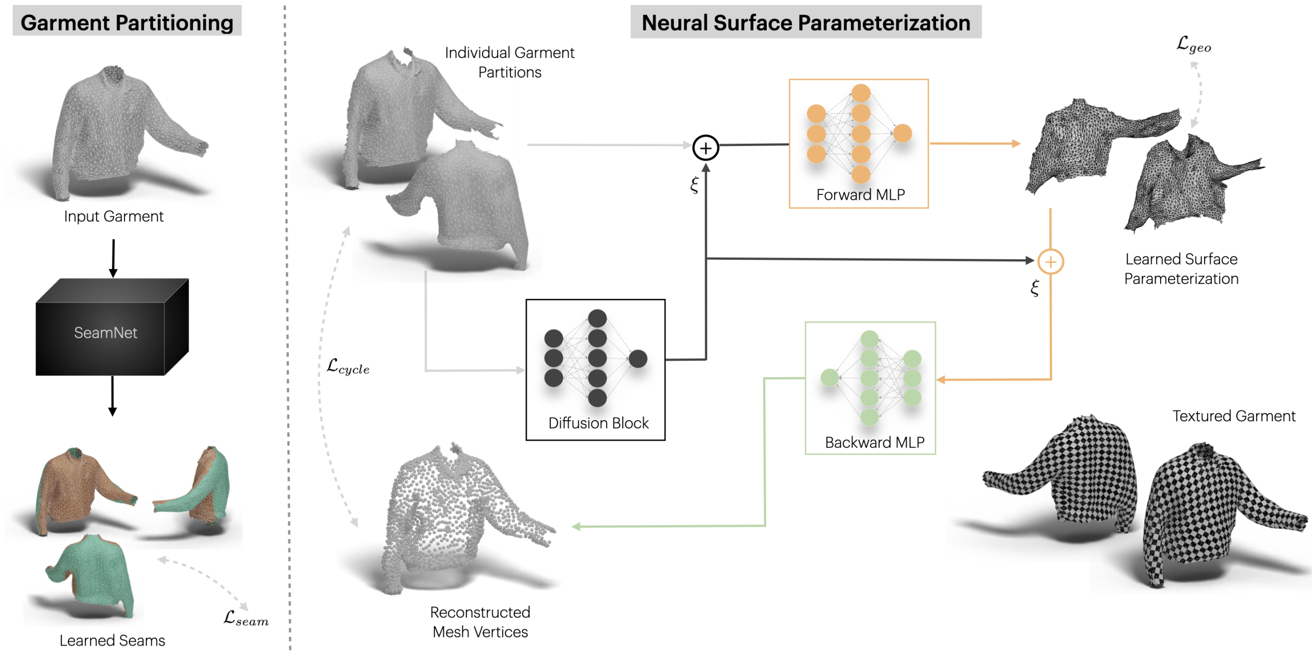

Figure 3: Overview of the Neural Surface Parameterization module.

2.2.1 Learnable Garment Partitioning

Let the extracted garment geometry is represented as mesh $\mathcal{M}_{gar}$. We need to partition the mesh surface before UV parametrization as unbounded and compact surfaces need to be divided into multiple partitions so that the distortion and overlaps in the UV space are minimal. We propose to estimate mesh partitioning as a vertex labeling task in a supervised setup. More specifically, let $\mathcal{M}^k_{gar} \subset \mathcal{M}_{gar}$ be the k-th partition from a set of total K mesh partitions. Inspired by the [14], we propose SeamNet which takes the extracted garment $\mathcal{M}_{gar}$ and estimate the mesh partitioning by assigning the probability $\widehat{\mathcal{P}}^k_{seam}$ to k_{th} partition. We train SeamNet in a supervised setting using the following objective function:

$$\mathcal{L}_{seam} = \sum_{k=1}^{K} \mathcal{P}^k_{seam} \log(\widehat{P}^k_{seam}) \qquad (2)$$

Here, $\mathcal{P}^k_{seam}$ is the ground truth seam labels.

2.2.2 Self-Supervised Neural Parameterization

Each individual mesh partitions $\mathcal{M}^k_{gar}$ is then passed to our novel neural parameterization network, which learns to map the surface to 2D UV space. Let $f : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ be the mapping function that maps each mesh vertex 3D position in world space to 2D position in UV space. We learn this mapping function by representing it with the neural network. Inspired by the fundamental nature of texture mapping, we learn to map the individual facets of the 3D surface to UV space instead of mapping it in per-vertex manner. We argue that this formulation considers the surface triangulation providing more constraint for mapping function.

More specifically, let $(\mathcal{V}_k, \mathcal{F}_k)$ are the vertices and facets of $\mathcal{M}^k_{gar}$ and we intend to recover UV mapping for all $\mathcal{V}_k$, represented as $\mathcal{U}_k$ where each element of this set is the UV position of corresponding mesh vertex i.e., $\mathcal{U}^i_k \in \mathbb{R}^2$. We input $\mathcal{V}_k$ to the diffusion module (adopted from [14]) to get a global shape embedding $\xi \in \mathbb{R}^{64}$. Note that we model each facet of the mesh as a set of three participating vertices (for a triangulated mesh) defining a facet feature (of dimension $\mathbb{R}^{3\times3}$). We concatenate global shape embedding ξ with these per-face features of $\mathcal{M}^k_{gar}$ to obtain a $\mathbb{R}^{3\times(3+64)}$ vector which can be reshaped to $\mathbb{R}^{201}$. We pass this per-face feature vector to the forward mapping MLP $\mathcal{F}_{forward}$ which outputs the $\mathbb{R}^{3\times2}$ dimensional UV position for each vertex of the respective facet. However, each vertex may participate in multiple faces; we propose to compute a mean UV position for each vertex of the mesh by taking an average over independently predicted UV positions across multiple faces sharing that vertex. In order to learn this mapping in a self-supervised manner, we employ cycle-consistency loss by learning an inverse mapping $f^{-1} : \mathbb{R}^2 \rightarrow \mathbb{R}^3$. We use the network architecture similar to the $\mathcal{F}_{forward}$ to train a backward MLP $\mathcal{F}_{backward}$ which takes $\mathbb{R}^{3\times(2+64)}$ dimensional input vector after concatenating $\mathcal{U}^i_k$ with ξ and maps it to the 3D world positions $\mathcal{V}_k$. Additionally, we also impose another loss function to regularize the UV mapping by minimizing the difference between geodesic and Euclidean distance between every pair of vertices over $\mathcal{M}^k_{gar}$. Let, $\mathcal{D}_{geodesic} \in \mathbb{R}^{\mathcal{V}_k \times \mathcal{V}_k}$ is a matrix of pairwise geodesic distance of each vertex with another vertex position and $\mathcal{D}_{euclidean} \in \mathbb{R}^{\mathcal{U}_k \times \mathcal{U}_k}$ is a matrix of pairwise euclidean distance of each mapped UV with another UV position.

We define the complete objective function as follows:

$$\mathcal{L}_{np} = \lambda_{cycle}\mathcal{L}_{cycle} + \lambda_{geo}\mathcal{L}_{geo} \tag{3}$$

$$\mathcal{L}_{cycle} = \frac{1}{|\mathcal{V}_k|} \sum_{v \in \mathcal{V}_k, u \in \mathcal{U}_k} (v - \mathcal{F}_{forward}(u))^2 \tag{4}$$

$$\mathcal{L}_{geo} = |\mathcal{D}_{geodesic} - \mathcal{D}_{euclidean}|^2 \tag{5}$$

2.3. Texture Map Filling

Once the neural parameterization is done, and the UV atlas is generated, the final task is to fill the UV map faithfully using the input RGB image(s). This is achieved by doing pixel-aligned texture filling followed by structured inpainting.

We leverage the information present in the input image to fill the learned UV atlas. Since our geometry reconstruction module uses implicit function learning, the alignment between parameterized mesh regions and image space is knowm apriori. Thus, we can fill the UV map by assigning the RGB value of each texel (UV map pixel) by projecting the corresponding 3D point on the mesh to the corresponding input image space.

However, some texels remain unfilled as they correspond to 3D points on mesh surfaces that are unobserved in RGB images either due to a lack of enough input camera views or caused by self-occlusion by body/cloth. However, this should be a small portal of texels if we are given enough multi-view images. Nevertheless, it is important to recover complete texture filling; hence, similar to [16], we use structured in-painting to fill the missing regions for each independent UV map by simply providing the filled texture map as initialization. This allows us to in-paint the high-frequency textures for the underlying garment.

3. Experiments & Results

3.1. Implementation Details

Here we provide implementation details, including architectural details and various parameters used for training the networks across multiple modules. We employ the original PIFu [12] architecture to learn to reconstruct the initial clothed body model from input RGB image(s). We used DiffusionNet [14] architecture with 4 blocks, channel width of 128 and Eigen basis 128. Relu activations are used for intermediate layers, log softmax is used for the semantic label prediction, and softmax for SeamNet. The DiffusionNet block used in the parameterization module uses tanh activation function in the final layer. We partition the garments into $K = 2$ partitions, namely, “front” and “back”. The forward and backward MLP (i.e., $\mathcal{F}_{forward}$ and $\mathcal{F}_{backward}$) use 8 hidden layers with LeakyReLU activation function and tanh at the final output layer. The lambda weights used in Eq.5 are $\lambda_{cycle} = \lambda_{geo} = 1.0$. ADAM optimizer is used in the training of all the networks, with a learning rate of 0.001, and batch size is set to 1. RTX 2080Ti GPU is used for training of all the networks. In case of single input, view is available, we fill both the front and back texture maps with the same input image and inpaint the remaining part using the structure inpainting [5] to avoid seams.

3.2. Evaluation Metrics

We are using the following evaluation metrics for the evaluation of our learned model.

Intersection Over Union (mIOU): IOU is a ratio of the area of overlap between the predicted segmentation and ground truth segmentation to the area of union between them. Since we learn the segmentation directly over the mesh surface, we calculate the IOU of predicted face labels and ground truth face labels. Finally, we report mean IOU across all semantic segmentation labels.

Accuracy (mACC): Accuracy is the percent of segmentation labels correctly assigned. Here mean accuracy over the faces across all labels is calculated and reported as a measure of performance of the semantic segmentation task.

Area and Angle Stretch: We use the angle and area stretch of open-source software Blender to plot the distortion plot in the UV parameterization.

	PIFu		Our method	
Loss	mIOU↑	mACC↑	mIOU↑	mACC↑
3dHuman	0.8352	0.7238	0.9232	0.8354
THUman2	0.7605	0.8823	0.8758	0.9387

Table 1: Comparison of results Implicit field learning and our method on segmentation task.

3.3. Datasets

We use following datasets for training and evaluation.

THUman2.0: This dataset contains 500 high-quality human scans captured by a dense DSLR rig. We manually curated segmentation labels.

3DHuman: This dataset contains around 250 meshes of people in diverse body shapes in various garment styles. It

Figure 4: 3D garment digitization generalization results on in-the-wild internet images.

	Optcuts	BFF	Ours
2k	1.795 sec	0.1 sec	**0.09** sec
8k	4.80 sec	1.00 sec	**0.35** sec
35k	53.5 sec	2.00 sec	**1.91** sec
100k	10 min +	14.0 sec	**7.33** sec

Table 2: Comparison of the inference time of our method with existing methods at various input vertices resolution.

covers a wide variety of clothing styles, from relatively tight garments to very loose garments. Similar to THUman2.0 we, manually curate the segmentation labels.

3.4. Quantitative Evaluation

We evaluate our neural parameterization with existing methods (BFF [13], ABF [15], LSCM [9], OptCut [10], Blender SmartUV [3]) in terms of inference time and report in Table 2 that our method is superior to them. We take a mesh (with only front partition) from the 3DHumans dataset and decimate it to 2k vertices. We train our neural parameterization network on this low resolution mesh and infer on different discretization levels of the same mesh. Note that other deterministic parameterization methods can not use this functionality as they have to minimize the loss function again even if there is a minimal change in mesh connectivity. This leads to the increase in their run-time as big as

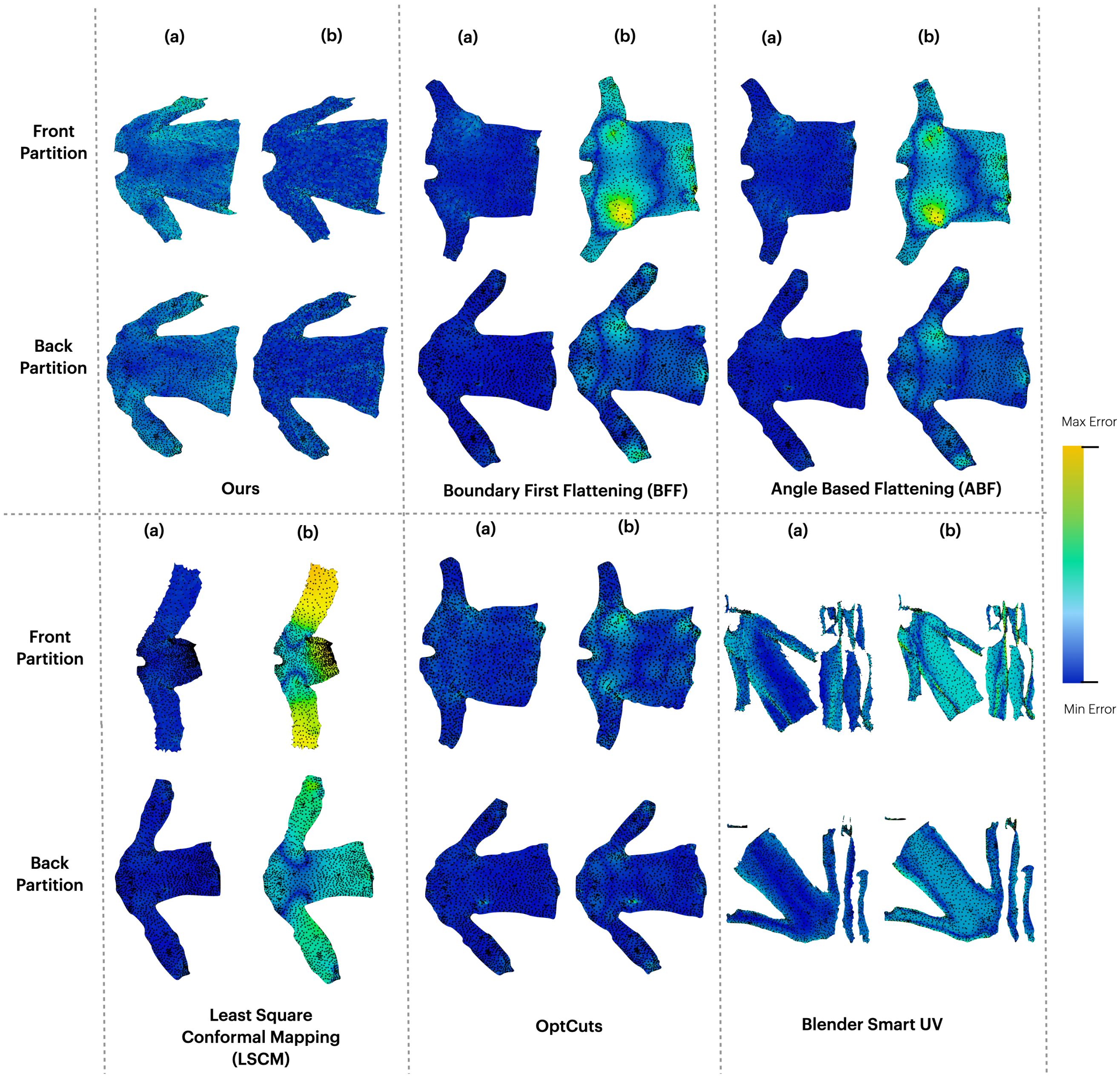

Figure 5: Comparison of UV parameterization computed with optimization based various deterministic methods ([13], [15], [9], [10], [3]) and our self-supervised learning approach. Columns (a) and (b) point to the angle and area stretch.

10 minutes as shown in Table 2. However, our method is discretization agnostic and thus can UV parameterize high resolution meshes with low inference time.

3.5. Qualitative Evaluation

We showcase the qualitative results of our framework on internet images to claim generalizability, as shown in Figure 4. Note that our framework can yield highly plausible high-frequency textural details while faithfully reconstructing the UV atlas. Please refer to our supplementary material for more qualitative results including comparison with SOTA methods.

Since our key novelty is learning based surface parameterization, we compare our method with the existing optimization based UV parameterization methods. Note that in case of texture mapping, we want to minimize both area and angle distortion so that texture editing and other texture manipulation tasks do not get affected much. As shown in

Figure 5, row (i) & (ii) are the front and back partitions of garment mesh, while columns (a)& (b) are the angle stretching and area stretching respectively. We can infer that our method is on-par with other methods in terms of minimizing the angle distortion while performs superior in case of minimizing the area deformation.

4. Discussion

4.1. Ablation Study

We perform ablation study on the choice of architecture to obtain the semantic segmentation of the reconstructed mesh. In first case, inspired by REEF [21], we use implicit learning based method to learn the semantic information in the form of semantic field. We observe that, though this method produces reasonable results, it fails to produce clear boundaries of segmentation. In fact, REEF [21] learns a separate boundary field to remedy this, however this requires additional annotated boundary ground truth data. We show that our segmentation network with DiffusionNet architecture produces far superior result and report the mIOU and mACC in Table 1. We provide further ablative study on neural parameterization loss functions as well as effect of global shape embeddings in the supplementary material.

4.2. Limitations & Future Work

Extrinsic camera calibrations need to be known for multi-view implicit reconstruction. This restricts applicability to calibrated input RGB images. Additionally, all existing methods for human digitization struggle while reconstructing fine grained geometrical details like thin threads and fine wrinkles. The texture map mitigates this challenge up to some extent by retaining such details in textural appearance. Similarly, small garment accessories like belts and ties are difficult to recover separately due to their small scale and proximity to garment surface. Another key limitation of existing methods, including ours is that the digitization is dependent on input illumination. Learning accurate surface reflectance for accurate appearance modeling in varying illumination will be interesting to explore as part of future work. Future, it will be interesting to explore reconstruction of same garments across multiple images of varying human subjects as well as body poses in different illumination.

5. Conclusion

We propose a novel framework for the template-free textured 3D digitization of arbitrary garments from single/multi-view images. The key novelty of our method is the learnable self-supervised discretization-agnostic UV parameterization of arbitrary garment surfaces along with automated partitioning of garments. This enables generalization of our framework to various garment types captured across multiple in-the-wild internet images. We perform a comprehensive empirical analysis on the publicly available datasets and report superior results to existing methods for garment extraction and as well as UV parameterization.

References

[1] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images, 2019.

[2] Zhiqin Chen, Kangxue Yin, and Sanja Fidler. Auv-net: Learning aligned uv maps for texture transfer and synthesis. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[3] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.

[4] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. Atlasnet: A papier-mâché approach to learning 3d surface generation, 2018.

[5] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Image completion using planar structure guidance. *ACM Trans. Graph.*, 33(4), jul 2014.

[6] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image, 2020.

[7] Sai Sagar Jinka, Rohan Chacko, Avinash Sharma, and P.J. Narayanan. Peeledhuman: Robust shape representation for textured 3d human body reconstruction. In *2020 International Conference on 3D Vision (3DV)*, pages 879–888, 2020.

[8] Sai Sagar Jinka, Astitva Srivastava, Chandradeep Pokhariya, Avinash Sharma, and P. J. Narayanan. Sharp: Shape-aware reconstruction of people in loose clothing. 2022.

[9] Bruno Lévy, Sylvain Petitjean, Nicolas Ray, and Jérome Maillot. Least squares conformal maps for automatic texture atlas generation. *ACM Trans. Graph.*, 21(3):362–371, jul 2002.

[10] Minchen Li, Danny M. Kaufman, Vladimir G. Kim, Justin Solomon, and Alla Sheffer. Optcuts: Joint optimization of surface cuts and parameterization. *ACM Transactions on Graphics*, 37(6), 2018.

[11] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), oct 2015.

[12] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019.

[13] Rohan Sawhney and Keenan Crane. Boundary first flattening. *ACM Trans. Graph.*, 37(1), dec 2017.

[14] Nicholas Sharp, Souhaib Attaiki, Keenan Crane, and Maks Ovsjanikov. Diffusionnet: Discretization agnostic learning on surfaces. arXiv, 2020.

[15] Alla Sheffer, Bruno Lévy, Maxim Mogilnitsky, and Alexander Bogomyakov. Abf++: fast and robust angle based flattening. *ACM Transactions on Graphics (TOG)*, 24(2):311–330, 2005.

[16] Astitva Srivastava, Chandradeep Pokhariya, Sai Sagar Jinka, and Avinash Sharma. xcloth: Extracting template-free textured 3d clothes from a monocular image. arXiv, 2022.

[17] Francis Williams, Teseo Schneider, Claudio Silva, Denis Zorin, Joan Bruna, and Daniele Panozzo. Deep geometric prior for surface reconstruction, 2018.

[18] Fanbo Xiang, Zexiang Xu, Miloš Hašan, Yannick Hold-Geoffroy, Kalyan Sunkavalli, and Hao Su. Neutex: Neural texture mapping for volumetric neural rendering, 2021.

[19] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction, 2020.

[20] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images, 2020.

[21] Heming Zhu, Lingteng Qiu, Yuda Qiu, and Xiaoguang Han. Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images, 2022.