SPnet: Estimating Garment Sewing Patterns from a Single Image of a Posed User

Seungchan Lim Sumin Kim Sung-Hee Lee

Korea Advanced Institute of Science and Technology (KAIST)

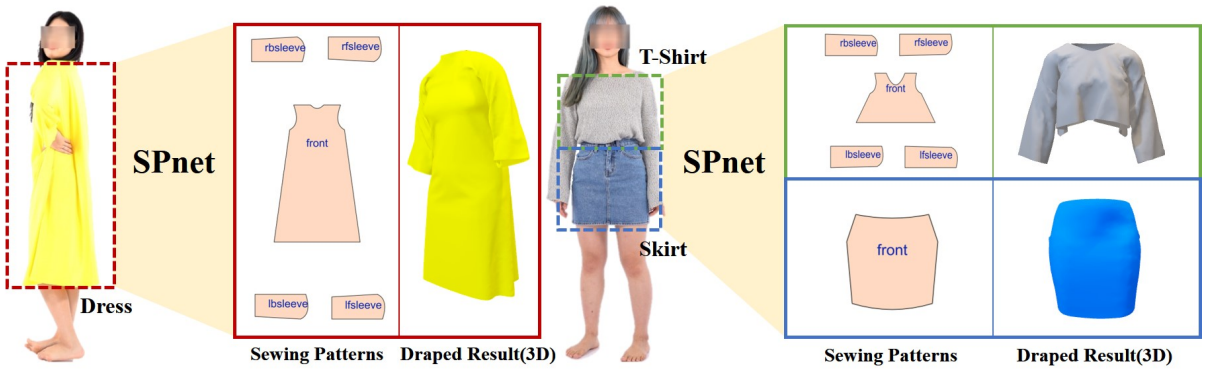

Figure 1: ***SPnet*** *can predict a garment sewing pattern from a single image with an arbitrary pose, enabling the creation of a natural-looking 3D garment mesh.*

Abstract

This paper presents a novel method for reconstructing 3D garment models from a single image of a posed user. Previous studies that have primarily focused on accurately reconstructing garment geometries to match the input garment image may often result in unnatural-looking garments when deformed for new poses. To overcome this limitation, our work takes a different approach by inferring the fundamental shape of the garment through sewing patterns from a single image, rather than directly reconstructing 3D garments. Our method consists of two stages. Firstly, given a single image of a posed user, it predicts the garment image worn on a T-pose, representing the baseline form of the garment. Then, it estimates the sewing pattern parameters based on the T-pose garment image. By simulating the stitching and draping of the sewing pattern using physics simulation, we can generate 3D garments that can adaptively deform to arbitrary poses. The effectiveness of our method is validated through ablation studies on the major components and a comparison with other methods.

CCS Concepts

• ***Computing methodologies*** → ***Shape modeling;***

1. Introduction

With growing interest in virtual clothing reconstruction in computer graphics, researchers have developed techniques to enhance the accuracy, interactivity, and speed of garment reconstruction. Regarding input data, various modalities have been investigated, such as single or multiple images [ZCJ*20]. The range of expressible garments and the naturalness of their shape on a reposed body are significantly influenced by how the garments are represented and deformed. While tight-fitting garments can be efficiently represented as displacements from the skin [BTTPM19], loose garments such as skirts and dresses require a separate modeling approach independent from the body. One effective method to achieve this is by utilizing a 3D parametric template to represent the garments [JZH*20, CPA*21]. However, when the template model is deformed using the linear blend skinning (LBS) method, which constrains garment deformation to a linear relation with the body pose, the quality of the deformation can be significantly compromised [BTTPM19, JZH*20, CPA*21].

DOI: 10.2312/egs.20241035

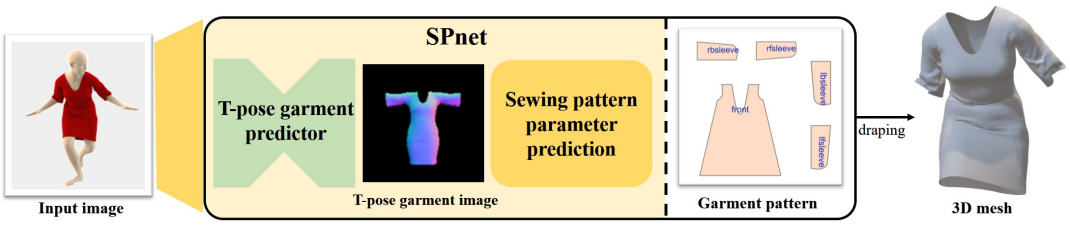

Figure 2: *Overview of our framework.*

Recent studies have focused on sewing patterns, which are collections of 2D patches that compose garments through stitching. Sewing patterns are not only informative for defining the basic shape of a garment but also enable the generation of garments with natural wrinkles when using sewing patterns as input to cloth simulators instead of outputting the mesh directly. Korosteleva and Lee [KL22] proposed a deep-learning framework to infer sewing patterns from the point cloud data of garment. Other works estimated parameterized sewing patterns from a single image by optimization [YPA*18] or supervised learning [LXL*23].

In this paper, we propose *SPnet*, a novel deep learning framework aimed at predicting stable garment **s**ewing **p**attern more efficiently compared to previous works. To address the challenge of predicting sewing patterns from images in varied poses, which is complicated by garment occlusions and deformations, we developed a two-stage deep learning framework. Initially, our T-pose garment predictor converts clothing images from any pose into a T-pose, revealing the garment's true shape free from pose effects. This helps distinguish between pose and garment characteristics, considering that shapes and wrinkles vary with poses. Subsequently, we predict the necessary parameters for the garment pattern from the T-pose prediction. These parameters are then used to create and simulate the garment pattern on an avatar, resulting in a natural-looking garment. Our approach's effectiveness is validated through ablation studies and comparisons with related studies.

2. Method

Our goal is to predict garment sewing patterns from a single near-front image of a person in an arbitrary pose. To address the challenge of predicting these patterns, we employ a two-step sequential approach consisting of T-Pose garment prediction and sewing pattern parameter prediction steps, as illustrated in Figure 2.

T-pose garment predictor Figure 3 shows the framework of the T-pose garment predictor, which infers the T-pose garment image G^t as a normal map from a single near-front image I^s. Since the source image I^s contains not only the clothing information of the person but also various complex features like face and hair, instead of directing feeding the source image to the network, we extract from the source image only the essential information for the garment prediction as network inputs. Specifically, we extract a pose map P^s that represents the user's body and pose [GNK18], and a garment normal map G^s that represents the shape of the posed garment by using PIFuHD [SSSJ20] and garment segmentation [JSS*20]. As

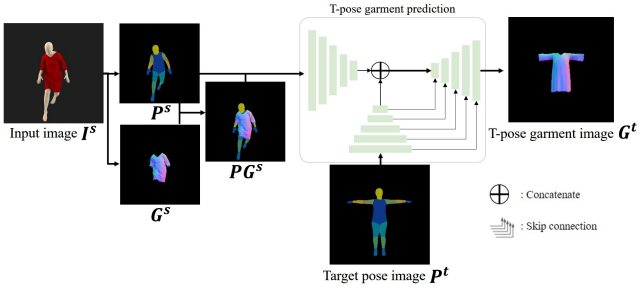

Figure 3: *The structure of the T-pose garment predictor.*

inputs to the network, we found that feeding P^s and PG^s, which overlays G^s onto P^s, leads to the best results. In addition, the T-pose P^t is provided to the network as the target pose for the garment.

This network structure was inspired by [YLG*21]. In [YLG*21], the pose map, garment label, and silhouette image are derived from the input image, enabling the encoding of the spatial interplay between the body and silhouette. In our case, for learning the relationship between garments of various sizes and poses of the human body, we use the normal map as the garment information. The convolutional layers encode the relationship between body parts and the corresponding clothes and wrinkles into a latent vector. Upon combining the encoded information with the latent vector of the target pose map P^t, the decoded output corresponds to the garment-worn image G^t in the T-pose. During the training process, the encoded feature obtained from the T-pose at each stage is passed to the corresponding component within the decoder via a skip connection to retain the T-pose information within the network, and instance normalization in each convolutional layer is applied to generalize and stabilize feature extraction for various poses of input images. We train T-pose garment predictor by minimizing the L1 distance of the predicted T-pose garment map and the ground truth.

Sewing pattern parameter predictor Given the T-pose garment image, we proceed to predict the sewing pattern parameters S^t. As shown in Figure 4, to predict the sewing pattern parameters from the normal map G^t, we use a convolutional autoencoder [BKC17] to encode the T-pose garment information into a latent vector, from which the sewing pattern parameters are extracted through MLP layers. Furthermore, the prediction of garment pattern parameters involves various variables, each reflecting distinct characteristics as

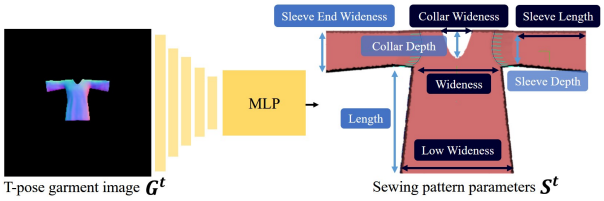

Figure 4: *The structure of the sewing pattern parameter predictor.*

detailed in [KL21]. Since we predict the clothing pattern parameters from front-view images, we ensure that the back of the pattern mirrors the characteristics of the front. We train sewing pattern parameter predictor by minimizing the L1 distance of the normalized pattern parameters and the ground truth.

Since the meaning and number of sewing pattern parameters differ for three distinct garment types: T-shirts, pants, and skirts, we conducted separate training for each of these garment types. We used the cloth characteristics of [KL21] to generate the garment patterns and simulate them on a T-pose SMPL avatar with Qualoth simulator [CK05]. To fit the garments to different poses, we use the CMU motion dataset to position the avatar and render the posed avatar using Arnold. The dataset consists of T-shirt, pants, and skirt categories, with a total of 2679, 1469, and 1772 different sizes and poses of garments, respectively. The network structure for SPnet is detailed in the supplementary material.

3. Results and Experiments

Figure 5 shows the prediction results from our held out synthetic dataset images through SPnet. Comparing the predicted sewing patterns from images with non-static poses and their respective ground truth, it shows that they are largely similar. Regardless of garment type and size, both the predicted patterns and draping images show that they reflect the characteristics of the input image well.

Ablation study To validate the effectiveness of the T-pose garment predictor, we conducted an ablation test by removing the T-pose garment predictor and training the sewing pattern parameter predictor to directly infer sewing pattern parameters from the pose maps P^s and garment normal images PG^s. In Table 1, the average error values of garment parameters for T-shirt, skirt, and pants are lower in the proposed framework. This demonstrates the necessity of an intermediate step of predicting the T-pose from input images.

Comparison To demonstrate the effectiveness of our garment generation approach, we predicted garment patterns from datasets of real-world photographs where sewing patterns and 3D garment forms do not exist, as shown in Figure 1. We then qualitatively evaluated the results of draping the patterns with state-of-the-art works. In Figure 6, when comparing the results of creating 3D clothes from wild images, both the BCNet [JZH*20] and SMPLicit [CPA*21] models generate clothes in a smooth form without creating details such as wrinkles. Additionally, they struggle to generate images, such as loose-fit clothes. In contrast, our method can generate shapes that well reflect the appearance of the clothing in the input image and can express elements that represent natural clothes.

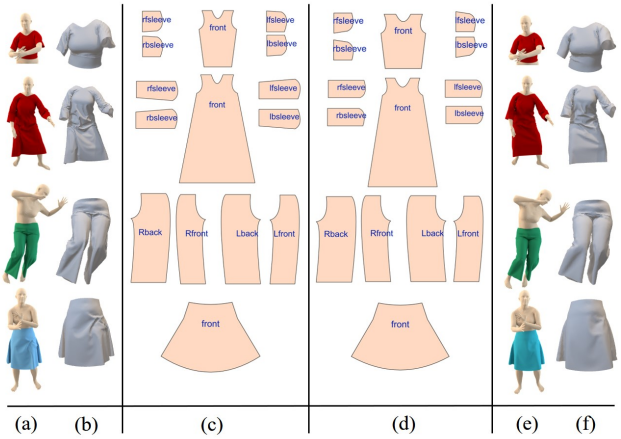

Figure 5: *SPnet results for synthetic garments. (a) Input images. (b) 3D models of input image. (c) Ground truth of garment sewing pattern. (d) Predicted garment sewing patterns. (e) Results of draping predicted patterns to the input pose by rendering and (f) generated garment geometry.*

Figure 6: *Qualitative evaluation. (a) BCNet [JZH*20]. (b) SMPLicit [CPA*21]. (c) Sewformer [LXL*23] (d) Ours.*

In the case of Sewformer [LXL*23], they can generate natural-looking garments, but there are limitations in predicting specific parts of clothing, such as the sleeves of T-shirts. While they predict the edges of garment patterns, our approach focuses on predicting pattern parameters based on each garment template. As a result, even though our predictions may depend on the templates, we can achieve stable garment pattern predictions. Furthermore, due to the

	T-shirt							
	sleeve length	low wideness	length	collar wideness	wideness	front collar depth	sleeve end wideness	sleeve depth
WOT	2.67 (2.06)	0.19 (0.14)	0.11 (0.1)	0.12 (0.09)	0.06 (0.04)	1.07 (0.8)	0.09 (0.05)	0.14 (0.09)
Ours	**1.69 (1.3)**	**0.17 (0.14)**	**0.08 (0.09)**	**0.12 (0.09)**	**0.06 (0.04)**	**1.03 (0.81)**	**0.08 (0.06)**	**0.11 (0.08)**

	Skirt			Pants		
	wideness	length	curve front	length	crotch depth	low wideness
WOT	0.14 (0.11)	0.11 (0.09)	1.55 (1.27)	4.9 (5.22)	1.76 (1.5)	3.95 (3.14)
Ours	**0.12 (0.1)**	**0.08 (0.08)**	**0.99 (0.91)**	**4.16 (4.71)**	**1.34 (1.57)**	**2.81 (2.33)**

Table 1: *Ablation study with or without the T-pose garment predictor (WOT). Each value represents the average error (along with the standard deviation) obtained by subtracting the predicted parameter value from the ground truth. The "sleeve length" of the T-shirt and the "length", "crotch depth", and "low wideness" of the pants are in centimeters, while the remaining parameters' units are the scaling from the respective clothing template. The experiment was conducted on a dataset reserved for validation, consisting of T-shirts of arbitrary sizes (536 items), skirts (354 items), and pants (294 items).*

lower number of output parameters, our model can be trained with two GTX 3090 GPUs, in contrast to Sewformer, which required eight A100 GPUs.

When creating clothes of Sewformer and our model, we used FrankMocap [RSJ21] for extracting pose and shape parameters of human model, which is also used by SMPLicit. Since BCNet incorporates its own pose estimator within its framework which outputs cloth mesh and not sewing patterns, we cannot apply FrankMocap as the pose estimator for BCNet. For evaluation of the garment's visual representation, we compared the garment-generated works from wild images in a frontal view. Comparisons with other wild images are available in the supplementary material.

4. Limitations and Conclusion

Our novel method SPnet used for predicting garment sewing patterns from single images employs a two-stage network training process, enabling the simulation of natural-looking garments with detailed wrinkles. However, our model is limited for images where the human subject is similar as the training data, which uses the female SMPL model with normal physique. Consequently, it struggles with predictions like sleeve width for thinner arms. Also, while we can represent realistic details like wrinkles, there are limitations in generating additional elements like pockets, zippers, and hoods. Despite these constraints, SPnet not only effectively matches predicted patterns in our synthetic dataset but also successfully applies to several real-world images, creating realistic 3D garments.

Acknowledgement

This work was supported by NRF, Korea (2022R1A4A5033689), by IITP, MSIT, Korea (2022-0-00566), and by MSIT and Gwangju Metropolitan City, Korea (AI industrial convergence cluster development project)

References

[BKC17] BADRINARAYANAN V., KENDALL A., CIPOLLA R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence 39*, 12 (2017), 2481–2495. 2

[BTTPM19] BHATNAGAR B. L., TIWARI G., THEOBALT C., PONS-MOLL G.: Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 5420–5430. 1

[CK05] CHOI K.-J., KO H.-S.: Stable but responsive cloth. In *ACM SIGGRAPH 2005 Courses*. 2005, pp. 1–es. 3

[CPA*21] CORONA E., PUMAROLA A., ALENYA G., PONS-MOLL G., MORENO-NOGUER F.: Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 11875–11885. 1, 3

[GNK18] GÜLER R. A., NEVEROVA N., KOKKINOS I.: Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 7297–7306. 2

[JSS*20] JIA M., SHI M., SIROTENKO M., CUI Y., CARDIE C., HARIHARAN B., ADAM H., BELONGIE S.: Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16* (2020), Springer, pp. 316–332. 2

[JZH*20] JIANG B., ZHANG J., HONG Y., LUO J., LIU L., BAO H.: Bcnet: Learning body and cloth shape from a single image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16* (2020), Springer, pp. 18–35. 1, 3

[KL21] KOROSTELEVA M., LEE S.-H.: Generating datasets of 3d garments with sewing patterns. *arXiv preprint arXiv:2109.05633* (2021). 3

[KL22] KOROSTELEVA M., LEE S.-H.: Neuraltailor: reconstructing sewing pattern structures from 3d point clouds of garments. *ACM Transactions on Graphics (TOG) 41*, 4 (2022), 1–16. 2

[LXL*23] LIU L., XU X., LIN Z., LIANG J., YAN S.: Towards garment sewing pattern reconstruction from a single image. *ACM Transactions on Graphics (TOG) 42*, 6 (2023), 1–15. 2, 3

[RSJ21] RONG Y., SHIRATORI T., JOO H.: Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 1749–1759. 4

[SSSJ20] SAITO S., SIMON T., SARAGIH J., JOO H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 84–93. 2

[YLG*21] YOON J. S., LIU L., GOLYANIK V., SARKAR K., PARK H. S., THEOBALT C.: Pose-guided human animation from a single image in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 15039–15048. 2

[YPA*18] YANG S., PAN Z., AMERT T., WANG K., YU L., BERG T., LIN M. C.: Physics-inspired garment recovery from a single-view image. *ACM Transactions on Graphics (TOG) 37*, 5 (2018), 1–14. 2

[ZCJ*20] ZHU H., CAO Y., JIN H., CHEN W., DU D., WANG Z., CUI S., HAN X.: Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16* (2020), Springer, pp. 512–530. 1