Computational Visual Media
DOI 10.1007/s41095-xxx-xxxx-x Vol. x, No. x, month year, xx–xx

Short Communication

Machine Learning for Digital Try-on: Challenges and Progress

Junbang Liang[1] (✉), and Ming C. Lin[1]

Abstract Digital try-on system, as a promising utility for E-Commerce, has the potential to become one of the revolutionary technologies that change people's lives. However, its development is limited by some practical constraints, such as accurate sizing of the body and vivid try-on demonstrations. With recent advances in machine learning, these challenging problems become increasingly more tractable. We enumerate a set of three open challenges towards a complete and easy-to-use try-on system that can be enabled by recent advances in machine learning. For each of them, we define the problem, introduce state-of-the-art approaches, and provide future directions. A digital try-on system enabled by machine learning techniques can further enhance the consumer's E-shopping experience and provide notable economic benefits to the society.

Keywords Machine-Learning, Digital Try-On, Garment Modeling, Human Body Estimation, Material Estimation.

1 Introduction

E-Commerce has been growing at a rapid pace in recent years. Consumers today are more likely to shop online than visiting retail stores. It is much more complicated, however, when it comes to buying clothes. People need to know how a garment fits on them, how it looks, and how it feels. Digital try-on systems can potentially satisfy these needs, providing direct visual impression, and possibly customized cloth sizing as well. Therefore, it has drawn much attention as one attractive alternative to improve the user experience and popularize online fashion shopping.

However, the technology is still far from being practical and easy-to-use to replace physical try-on. Currently, most of the try-on systems are either image-editing, copy-pasting, or only showing template demonstrations, while the ultimate goal is a fast and realistic try-on system adaptive to any customers' body. There is still a substantial technological gap in modeling and demonstrating garment fitting in the digital vs. the real worlds, including fast and realistic demonstration, accurate modeling of human body and garments, faithful estimation of garment materials, and lossless transformation of garments between virtual and physical worlds.

In this paper, we present some open research issues that contribute to this technological gap, including:

1. Accurate estimation of human shapes and sizes through consumer devices;
2. Faithful recovery of garment materials via (online) images;
3. Ease of design and manipulation on sewing patterns and garment pieces by end-users.

Although traditional methods have made important progresses on these under-constrained problems, learning-based approaches have shown tremendous potential in making notable impact. Compared to traditional methods, machine-learning algorithms are usually a lot faster since the training and the optimization are done offline. They are also good at generalization to unseen images without the need of tedious data pre-processing. While extensive study is made on 2D image learning, machine-learning on 3D human body and shapes with high variations is still far from being mature, which is the reason why the open issues described above still remain illusive.

For each problem listed above, we motivate its importance, provide a problem definition, and present state-of-the-art approaches with potential improvements. We believe that the solutions to these challenging problems will lead to significant advances in digital try-on, as well as other areas of E-commerce.

1 University of Maryland, College Park, MD 20785, USA. E-mail: liangjb@cs.umd.edu, lin@cs.umd.edu.

Manuscript received: 2014-12-31; accepted: 2015-01-30.

2 Open Problems

In this section we first introduce the three major challenges that limit digital try-on technology from being widely adopted and accepted by shoppers. There are several reasons why they still prefer physical try-on. First, consumers are unsure if what they buy online will fit their bodies well. Although there exist general sizing systems for individuals, its lack of consistency and standardization across different brands and garment materials can often make it difficult to sizing the clothes, especially for those with non-standard body shapes and proportion. Accurate estimation of human body shapes is the key to make digital try-on work. Second, the fabric material is usually one of the key considerations when shopping for clothes. Different fabric materials affect how the garments look and fit on a body, how consumers would wear it, and whether or not they would buy it. However, the correspondences between the actual material and its digital representation are not well understood. It is also challenging to acquire full fabric material and digital cloning from the real-world examples.

Visual effects from the customers' view is as critical as other factors. There are two approaches to display garments: 2D image-based and 3D mesh with (photo-realistic) rendering. They have different advantages and drawbacks, but both need a large garment database for support. While creating a 3D garment takes considerable efforts, 2D images often suffer from the lack of variation and are much more difficult to make customized changes. In either case, the try-on system would need a user-friendly design and manipulation backend to suffice the needs. Last, but not least, a fast and vivid animation of the garments in motion along with the body movement will greatly improve the user experience. Although it is not so critical as other factors, it would effectively reduce the perceptual gap between the real world and the digital one for online shopping. Previous work has proposed using cloud computing to improve the animation speed, but there is still a notable technology gap towards high-quality, interactive 3D animation of clothes.

3 Human Shape Estimation

As mentioned above, accurate human shape estimation is key to enabling digital try-on. Human body reconstruction, consisting of pose and shape estimation, has been widely studied in a variety of areas, including digital surveillance, computer animation, special effects, and virtual/augmented environments. Yet, it remains a challenging and popular topic of interest. While direct 3D body scanning can provide excellent and sufficiently accurate results, its adoption is somewhat limited by the required specialized hardware. As input, RGB images are widely available and can be easily captured using commodity mobile devices for digital try-on. Although purely image-based try-on methods are proposed [29], learning-based 3D body estimations are more widely applicable in that the 3D body is articulated and can be re-posed and re-targeted.

We define the human-body reconstruction problem informally as, given one or a set of RGB images, estimate the human body geometry and sizes, and output (preferably) a 3D humanoid mesh. Traditional algorithms often formulate it as an optimization problem, and compute the silhouette difference as its major part of the target function [7]. Therefore, these methods either require the human to wear tight clothes, or alternatively relax the target function to be unilateral on uncovered body parts [3] or to point correspondences [13]. The use of machine learning methods in this problem was game-changing. First, it moved the algorithm from online to offline, significantly reduced the inquiry latency. Second, by using a parametric human model [17], one can easily construct a regression network for the parameters while the losses needed can also be inferred from them. While early works propose network models for only 2D/3D body skeletons [4, 18, 24], more recent works introduce techniques to regress the entire human body – either using a parametric human model [2, 11] or voxel-based representation [20, 22, 28]. Given the fact that the annotations in most real-world datasets contain only joint positions, the learning process has been refined in various ways [1, 12, 21, 25]. The current state of the art is the recent work by [15][1] that emphasizes on shape learning, while many other works often focus on body-joint losses, but neglect the effect of body shapes.

The key contribution of [15] is a multi-view multi-stage framework to address the ambiguity issue caused by camera projection (Fig. 1). Their model is iteratively run for several stages of error correction. Within each stage, the multi-view image input is passed on one at a time. At each step, the shared-parameter prediction block computes the correction based on the image feature and the input guesses. The camera and the human body parameters are estimated at the same time, projecting the predicted 3D joints back to 2D for loss computation. The estimated pose and shape parameters are shared among all views, while each view maintains its camera calibration and the global rotation. Their proposed framework uses a recurrent structure, making it a universal model applicable to the input of any number of views. At the same time, it couples the shareable information across different views so that the human body pose and shape can be optimized using image features from all views. Different

[1]Their data and code is available at: https://gamma.umd.edu/researchdirections/virtualtryon/humanmultiview

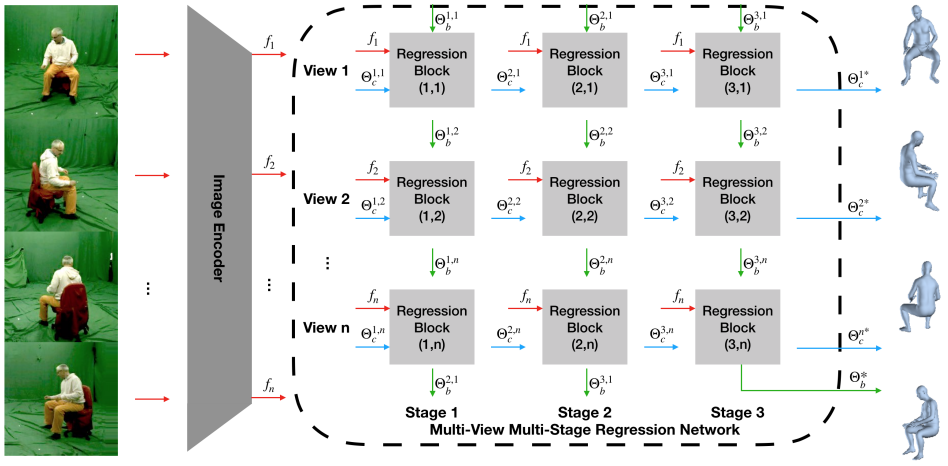


Fig. 1 The network structure from [15]. By using an iterative value correction structure, the visual information from different views is effectively integrated together to jointly provide a unified human shape.

(a) The input image. (b) Recovered Body [15].

Fig. 2 Prediction results of the state of the art [15]. The model can capture the shape of the human body by learning from synthetic data. The recovered legs and chest are close to the person in the image.

from static multi-view CNNs which have to fix the number of inputs, they make use of the RNN-like structure in a cyclic form to accept any number of views, and avoid the gradient vanishing by predicting the corrective values instead of the updated parameters at each regression block.

Experiments have shown that this trained model can provide equally good pose estimation as the state of the art using single-view images, while providing considerable improvement on pose estimation using multi-view inputs and a better shape estimation across all datasets. One example is demonstrated in Fig. 2. Moreover, a physically-based synthetic data generation pipeline is introduced to enrich the training data, which is very helpful for shape estimation and regularization of end effectors that traditional datasets do not capture. While synthetic data improves the diversity of human bodies with ground-truth parameters, a larger garment dataset and a more convenient registration process are needed to minimize the performance gap between real-world images and synthetic data. In addition, other variables such as hair, skin color, and 3D backgrounds are subtle elements that can influence the perceived realism of the synthetic data at the higher expense of a more complex data generation pipeline. With the recent progress in image style transfer using GAN, a promising direction is to transfer the synthetic result to more realistic images to further improve the learning result.

4 Garment Material Cloning

Garment material plays an important role in digital try-on systems. Physical recreation of the fabric not only gives a compelling visual simulacrum of the cloth, but also affects how the garment feels and fits on the body. However, modeling material is a challenging task: the visual effect and physical interaction of the garment is determined not only by the type of materials it is made of, but also the way of sewing and yarning. Due to this factor, researchers often focus on the physical properties it behaves, rather than the underlying semantic primitives.

Following such assumption, we model the garment material cloning problem as below. Given a sufficient amount of data, model its physical behaviors and compute the corresponding properties, such that the same or similar visual effect can be reproduced on computers. It has two implication: first we need to define a physical model of the material, next we estimate the parameters in the model.

There are many works to model clothes, including spring-mass systems and finite elements. Finite elements method is the most popular model since it can produce realistic results. While one can use isotropic properties such as Yang's Modulus and Poisson Ratio, anisotropic model is the better choice since it can support different behaviors caused by yarning.

4.1 Learning-Based Estimation

While traditional optimization methods [27] often take a long time to compute the material parameters, machine-learning methods can do predictions in real-time by a simple feed-forward operation, which is more feasible in applications that need fast feedback, such as garment prototyping. The state-of-the-art model from Yang *et al*. [26][2] used CNN combined with LSTM to recover the material parameters from videos.

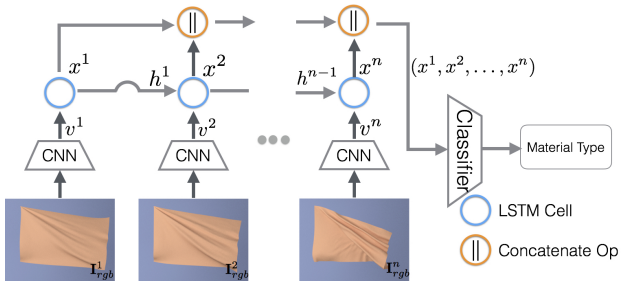


Fig. 3 The network model from [26]. The cloth material is estimated by learning motion patterns of image features given by CNN.

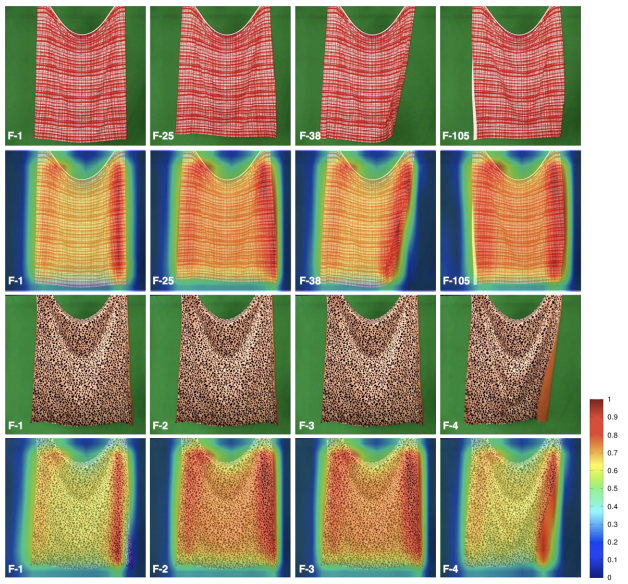


Fig. 4 Learned CNN conv5-layer activation visualization from [26]. Experiments show that the trained model is able to capture moving parts of the cloth even in an unseen video.

To constrain both the input and solution space, they choose one of the materials as the basis and the material sub-space is constructed by multiplying this material basis with a positive coefficient. To construct an optimal material parameter sub-space, a material parameter sensitivity analysis is conducted to examine the sensitivity of the material parameters κ with respect to the amount of deformation $D(\kappa)$. Physically based cloth simulations are used to generate a much larger number of data samples within these sub-spaces that would otherwise be difficult or time-consuming to capture. The cloth meshes are generated through physically based simulation, then rendered to 2D images with a randomly assigned texture. With the data samples, they combine the image signal feature extraction method, CNN, with the temporal sequence learning method, LSTM, to learn the mapping from visual "appearance" to "material". As shown in Fig. 3, the CNN layer is used to extract both low- and high-level visual features, while the LSTM layer focuses on learning the mapping between the material properties of the cloth and its sequential movement.

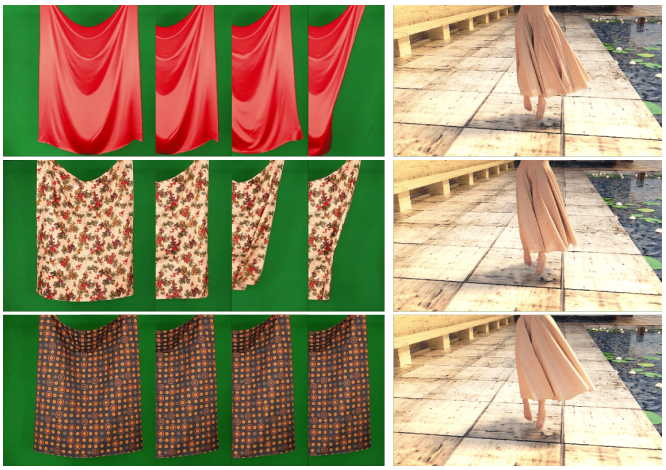


Fig. 5 Yang *et al*. [26] predicted the material type of the cloth in input videos from the left and cloned those material on to the skirt. The simulated skirt are shown on the right.

They demonstrated the proposed framework with the application of "material cloning". With the trained deep neural network model being able to capture the cloth motions (Fig. 4), the material type can be inferred from a video recording the motion of the cloth in a fairly small amount of time. The recovered material type can be "cloned" on another piece of cloth or a piece of garment as shown in Fig. 5.

In this work, the videos contain only a single piece of cloth and the recorded cloth is not interacting with any other object. While this is not always the case in real-world scenarios, this method provides new insights on addressing this challenging problem. A natural extension would be to learn from videos of cloth directly interacting with the human body, under varying lighting conditions and partial occlusion.

[2]Their data and code is available at: `http://gamma.cs.unc.edu/VideoCloth`

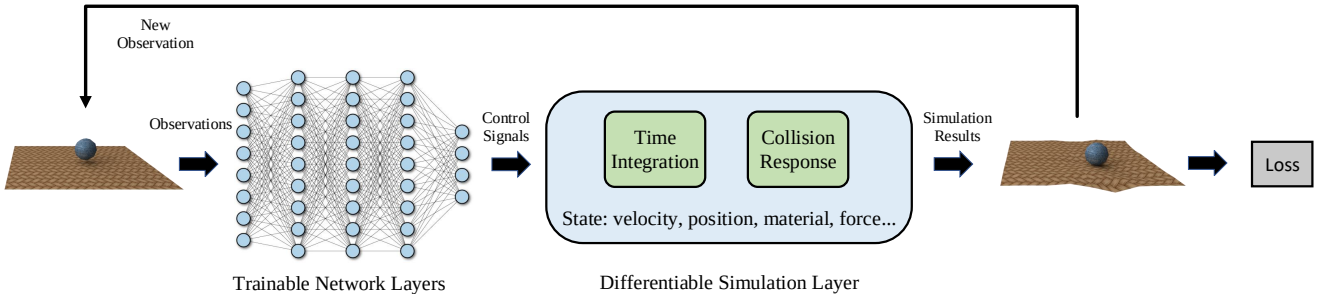


Fig. 6 Differentiable simulation embedding example from [19]. The loss can be backpropagated through the physics simulator to the neural network, enabling learning tasks such as material estimation and motion control.

4.2 Optimization Using Differentiable Physics

Another approach to obtain the fabric material is to measure the geometry difference directly during parameter optimization. Assuming that the environment is known to the system, the computation of the estimated motion and its gradient w.r.t. the material parameters can be achieved using differentiable simulation. A typical usage of differentiable simulation is motion control (Fig. 6), where it measures the difference to the target and backpropagate the loss to the network. Similar processes can be applied to material estimation as well. By measuring the distance to the target as the loss and computing the corresponding gradients, either in pixel space or in 3D space, the material parameters can be learned or optimized to achieve the desired cloth motion or visual effect. Recent differentiable physics work include rigid body [5, 6], cloth [14], and particle-grid system [8, 9]. The state-of-the-art work is from [14][3], where they proposed a method to differentiate cloth simulation. It is the first work to tackle high dimensional simulation problem and to propose a general differentiable collision handling algorithm. Later, a follow-up work [19] extends the algorithm to be applicable to coupled dynamics with rigid bodies.

In general, they follow the computation flow of the common approach to cloth simulation: discretization using the finite element method, integration using implicit Euler, and collision response on impact zones. They use implicit differentiation in the linear solve and the optimization, in order to compute the gradient with respect to the input parameters. The discontinuity introduced by the collision response is negligible because the discontinuous states constitute a zero-measure set. During the backpropagation in the optimization, the gradient values can be directly computed after QR decomposition of the constraint matrix. In their pipeline, there are several techniques that can be employed in other differentiable simulations.

[3]Their data and code is available at: https://gamma.umd.edu/researchdirections/virtualtryon/differentiablecloth

4.2.1 Derivatives of the Physics Solve

In modern simulation algorithms, implicit Euler is often used for stable integration results. Thus the mass matrix $\mathbf{M}$ often includes the Jacobian of the forces. It is denoted as $\hat{\mathbf{M}}$ in order to mark the difference. A linear solve will be needed to compute the acceleration since it is time-consuming to compute $\hat{\mathbf{M}}^{-1}$. They use implicit differentiation to compute the gradients of the linear solve. Given an equation $\hat{\mathbf{M}}\mathbf{a} = \mathbf{f}$ with a solution $\mathbf{z}$ and the propagated gradient $\frac{\partial \mathcal{L}}{\partial \mathbf{a}}|_{\mathbf{a}=\mathbf{z}}$, where $\mathcal{L}$ is the task-specific loss function, they derived the implicit differentiation form to derive the gradients. We refer the readers to the original paper [14] for more details.

4.2.2 Derivatives of the Collision Response

A general approach using LCP to integrating collision constraints into physics simulation has been proposed. However, constructing a static LCP is often impractical in cloth simulation due to high dimensionality. Collisions and contacts happen at each step are very sparse compared to the complete set. Therefore, they use a dynamic approach that incorporates collision detection and response.

Collision handling in their implementation is based on impact zone optimization. It finds all colliding instances using continuous collision detection and sets up the constraints for all collisions. In order to introduce minimum change to the original mesh state, a QP problem is developed to solve for the constraints. Since the signed distance function is linear in $\mathbf{x}$, the optimization takes a quadratic form, as shown originally in [14]:

$$\underset{\mathbf{z}}{\text{minimize}} \quad \frac{1}{2}(\mathbf{z}-\mathbf{x})^\top \mathbf{W}(\mathbf{z}-\mathbf{x}) \tag{1}$$

$$\text{subject to} \quad \mathbf{G}\mathbf{z} + \mathbf{h} \leq \mathbf{0} \tag{2}$$

where $\mathbf{W}$ is a constant diagonal weight matrix related to the mass of each vertex, and $\mathbf{G}$ and $\mathbf{h}$ are constraint parameters. They further denote the number of variables and constraints by n and m, *i.e.* $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{h} \in \mathbb{R}^m$, and $\mathbf{G} \in \mathbb{R}^{m \times n}$. Note that this optimization is a function with inputs $\mathbf{x}$, $\mathbf{G}$, and $\mathbf{h}$, and output $\mathbf{z}$. The goal here is to derive $\frac{\partial \mathcal{L}}{\partial \mathbf{x}}$, $\frac{\partial \mathcal{L}}{\partial \mathbf{G}}$, and $\frac{\partial \mathcal{L}}{\partial \mathbf{h}}$ given $\frac{\partial \mathcal{L}}{\partial \mathbf{z}}$, where $\mathcal{L}$ refers to the loss function.

Method	Runtime (sec/step/iter)	Density error (%)	Linear stretching stiffness error (%)	Bending stiffness error (%)	Simulation error (%)
Baseline	-	68 ± 46	160 ± 119	$\mathbf{70 \pm 42}$	12 ± 3.0
L-BFGS	2.89 ± 0.02	4.2 ± 5.6	72 ± 90	$\mathbf{70 \pm 43}$	4.9 ± 3.3
Liang *et al*. [14]	$\mathbf{2.03 \pm 0.06}$	$\mathbf{1.8 \pm 2.0}$	$\mathbf{45 \pm 41}$	77 ± 36	$\mathbf{1.6 \pm 1.4}$

Tab. 1 Results on the material parameter estimation task from [14]. Their proposed method runs faster than L-BFGS. Values of the material parameters are the Frobenius norms of the difference normalized by the Frobenius norm of the target. Smaller error percentage is better. Values of the simulated result are the average pairwise vertex distance normalized by the size of the cloth. The gradient-based method yields much smaller errors than the baselines.

When computing the gradient using implicit differentiation, the dimensionality of the linear system can be very high. Their key observation here is that $n \gg m > \text{rank}(\mathbf{G})$, since one contact often involves 4 vertices (thus 12 variables) and some contacts may be linearly dependent (*e.g*. multiple adjacent collision pairs). They minimize the size of the linear equation based on the QR decomposition of $\mathbf{G}$, which is the key to accelerate backpropagation of high dimensional QP problems.

One of their experiments shows its ability on material parameter optimization, the aim of which is to learn the material parameters of cloth from observation. The scene features a piece of cloth hanging under gravity and subjected to a constant wind force. The material model consists of three parts: density d, stretching stiffness $\mathbf{S}$, and bending stiffness $\mathbf{B}$. The stretching stiffness quantifies how large the reaction force will be when the cloth is stretched out; the bending stiffness models how easily the cloth can be bent and folded.

Tab. 1 shows the estimation result. They achieve a much smaller error in most measurements in comparison to the baselines. The table shows that the linear part of the stiffness matrix is optimized well. With the computed gradient using their model, one can effectively optimize the unknown parameters that dominate the cloth movement to fit the observed data.

As a follow-up work, Qiao *et al*. extend the differentiable simulation pipeline to couple with rigid body dynamics. They formulate the dynamics using generalized coordinates:

$$\frac{d}{dt}\begin{pmatrix}\mathbf{q}\\ \dot{\mathbf{q}}\end{pmatrix} = \begin{pmatrix}\dot{\mathbf{q}}\\ \ddot{\mathbf{q}}\end{pmatrix} = \begin{pmatrix}\dot{\mathbf{q}}\\ \mathbf{M}^{-1}\boldsymbol{f}(\mathbf{q},\dot{\mathbf{q}})\end{pmatrix}, \quad (3)$$

and update the optimization formulation for collision response accordingly (see [19] for details):

$$\begin{aligned} \underset{\mathbf{q}'}{\text{minimize}} \quad & \frac{1}{2}(\mathbf{q}-\mathbf{q}')^\top \hat{\mathbf{M}}(\mathbf{q}-\mathbf{q}') \\ \text{subject to} \quad & \mathbf{G}\boldsymbol{f}(\mathbf{q}') + \mathbf{h} \le \mathbf{0}. \end{aligned} \quad (4)$$

Due to the inclusion of rigid bodies, the constraints used in the optimization is no longer linear. When computing gradients, they linearize the constraints around the neighborhood as an approximation to enable the QR decomposition for acceleration as previously mentioned.

5 Garment Modeling and Design

Realistic apparel model generation has become increasingly popular, due to the rapid change of the fashion trend and the growing need for garment model in different applications such as virtual try-on. It is already the case even for state-of-the-art interactive apparel design systems [16]. For the application requirements, it is important to have a general cloth model that can represent a diverse set of garments. However, there are many challenges in automatic garment model generation. First, garments usually have different types of topology, especially for fashion apparel, that makes it difficult to design a universal generation pipeline. Moreover, it is often not straightforward for the general garments design to be retargeted onto another body shape, making it difficult for customization.

Some previous work has addressed this problem to some extent. Huang *et al*. [10] proposed an realistic 3D garment generation algorithm based on front and back image sketches, but it cannot retarget the generated garments to other body shape easily. Wang *et al*. [23] proposed an algorithm which can do retargeting conveniently, but have limited topology like T-shirt or skirt. As such, there is no recent work that addresses these two problems at the same time.

We introduce a learning-based parametric generative model to overcome the above difficulties, given garment sewing patterns and human body shapes as input. One possible approach would be to compute a *displacement image* on the UV space of the human body as an unified representation of the garment mesh. Different topology and sizes of the garment are represented by different values in the image. The 2D displacement image, as the representation of the 3D garment mesh data, can then be fed into conditional Generative Adversarial Network (cGAN) for latent space learning. The 2D representation for the garment mesh can transfer the irregular 3D mesh data to regular image data where a traditional CNN can easily learn. It can also extract the relative geometry information with respect to the human body, enabling garment retargeting to a different human body.

6 Conclusion

Although Virtual Reality and digital try-on demonstrate excellent potential and are rapidly developing, there remain open problems to address before the online try-on systems can be widely adopted. We listed three major challenges, all of which can be addressed or further improved using machine learning algorithms. For garment material prediction, state-of-the-art methods are still limited in that the training data is highly constrained: the scenario contains only a piece of cloth floating in the wind. To improve its applicability on daily tasks, it is necessary to focus on solving the problem on a more diverse set of inputs. Predicting the material from a garment on a fixed human body could be a good start, before generalizing to arbitrary human motions and predicting multiple garments on the same body. In the area of human shape estimation, it would be interesting to learn how external constraints could improve the estimation accuracy. For example, the shape and size of the garment are hard constraints that the predicted body should conform. While optimization-based methods can integrate these constraints fairly easily, it remains illusive for learning-based approaches. One possibility is to jointly estimate body and garment together and introduce intersection loss in between. This approach would require a new solution to the open problem of unified deep garment representation, if we do not want to train one model for every garment type, which could be even more challenging. We believe that substantial breakthroughs in digital try-on are achievable with more investigation towards these directions.

Acknowledgements

This research is supported in part by Iribe Professorship and National Science Foundation.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019.

[2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018.

[3] A. O. Bălan and M. J. Black. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision*, pages 15–29. Springer, 2008.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[5] F. de Avila Belbute-Peres, K. A. Smith, K. Allen, J. Tenenbaum, and J. Z. Kolter. End-to-end differentiable physics for learning and control. In *Advances in Neural Information Processing Systems*, 2018.

[6] J. Degrave, M. Hermans, J. Dambre, and F. wyffels. A differentiable physics engine for deep learning in robotics. *Frontiers in Neurorobotics*, 13, 2019.

[7] E. Dibra, H. Jain, C. Öztireli, R. Ziegler, and M. Gross. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 108–117. IEEE, 2016.

[8] Y. Hu, L. Anderson, T. Li, Q. Sun, N. Carr, J. Ragan-Kelley, and F. Durand. DiffTaichi: Differentiable programming for physical simulation. In *ICLR*, 2020.

[9] Y. Hu, J. Liu, A. Spielberg, J. B. Tenenbaum, W. T. Freeman, J. Wu, D. Rus, and W. Matusik. ChainQueen: A real-time differentiable physical simulator for soft robotics. In *International Conference on Robotics and Automation (ICRA)*, 2019.

[10] P. Huang, J. Yao, and H. Zhao. Automatic realistic 3d garment generation based on two images. *2016 International Conference on Virtual Reality and Visualization (ICVRV)*, 2016.

[11] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[12] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2252–2261, 2019.

[13] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 3, 2017.

[14] J. Liang, M. Lin, and V. Koltun. Differentiable cloth simulation for inverse problems. In *Advances in Neural Information Processing Systems*, pages 771–780, 2019.

[15] J. Liang and M. C. Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4352–4362, 2019.

[16] K. Liu, X. Zeng, P. Bruniaux, X. Tao, X. Yao, V. Li, and J. Wang. 3d interactive garment pattern-making technology. *Computer-Aided Design*, 104:113–124, 2018.

[17] 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):1–16, 2015.

[18] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei,

H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017.

[19] Y.-L. Qiao, J. Liang, V. Koltun, and M. Lin. Scalable differentiable physics for learning and control. In *ICML*, 2020.

[20] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2314, 2019.

[21] D. Smith, M. Loper, X. Hu, P. Mavroidis, and J. Romero. Facsimile: Fast and accurate scans from an image in less than a second. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5330–5339, 2019.

[22] 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)*, pages 20–36, 2018.

[23] T. Y. Wang, D. Ceylan, J. Popović, and N. J. Mitra. Learning a shared shape space for multimodal garment design. In *SIGGRAPH Asia 2018 Technical Papers*, page 203. ACM, 2018.

[24] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

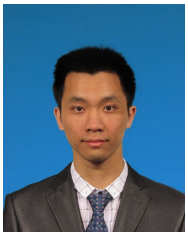
[25] Y. Xu, S.-C. Zhu, and T. Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7760–7770, 2019.

[26] S. Yang, J. Liang, and M. C. Lin. Learning-based cloth material recovery from video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4383–4393, 2017.

[27] S. Yang, Z. Pan, T. Amert, K. Wang, L. Yu, T. Berg, and M. C. Lin. Physics-inspired garment recovery from a single-view image. *ACM Transactions on Graphics (TOG)*, 37(5):1–14, 2018.

[28] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7739–7749, 2019.

[29] Z.-H. Zheng, H.-T. Zhang, F.-L. Zhang, and T.-J. Mu. Image-based clothes changing system. *Computational Visual Media*, 3(4):337–347, 2017.



Junbang Liang is a 4th year Ph.D. student in University of Maryland, College Park. He received his B.E. degree from Tsinghua University in 2016, and his M.S. degree from University of North Carolina in 2018. His research interests are physics-based cloth simulation, computer vision and machine learning.

Ming C. Lin is a Distinguished University Professor and Elizabeth Stevinson Iribe Chair of Computer Science at the University of Maryland College Park and John R. and Louise S. Parker Distinguished Professor Emerita of Computer Science at the University of North Carolina (UNC), Chapel Hill. She obtained her B.S., M.S., and Ph.D. in Electrical Engineering and Computer Science from the University of California, Berkeley. She is a Fellow of ACM, IEEE, and Eurographics, and a member of ACM SIGGRAPH Academy.

Declaration

Availability of data and materials. The data and the code of aforementioned state-of-the-art is publicly available:

- Data and code of [15]: `https://gamma.umd.edu/researchdirections/virtualtryon/humanmultiview`
- Data and code of [14]: `https://gamma.umd.edu/researchdirections/virtualtryon/differentiablecloth`
- Data and code of [26]: `http://gamma.cs.unc.edu/VideoCloth/`

Competing interests. The authors declare no competing interests.

Funding. This research is supported in part by Iribe Professorship and National Science Foundation.

Authors' contributions. We list a set of three open problems towards a complete and easy-to-use try-on system that can be enabled by recent advances in machine learning. For each of them, we define the problem, introduce state-of-the-art approaches, and provide future directions. A digital try-on system enabled by machine learning techniques can further enhance the consumer's E-shopping experience and provide notable economic benefits to the society.

Acknowledgments. This research is supported in part by Iribe Professorship and National Science Foundation.