# Virtual Draping

*** Aditti Ramsisaria**
ECE, Robotics
aramsisa@andrew.cmu.edu

*** Caroline Pang**
ECE, HCI
carolinep@andrew.cmu.edu

*** Claire Lin**
ECE, AI & ML
shaoyuli@andrew.cmu.edu

## Abstract

We present Virtual Draping (VDP)', a method to drape a garment from a virtual wardrobe over a 3D body mesh reconstructed from a single image using the SMPL format [1]. We estimate a full 3D mesh from just 2D joint information using the off-the-shelf SMPLify-X [2] model, and use a self-supervised method to learn dynamic deformations of garments from physics-based losses based on SNUG [3]. Our model allows the prediction of the SMPL parameters from the 2D image (or video frames for dynamic data) and renders a garment from a virtual wardrobe onto the 3D body mesh. The garment deformation model is capable of adapting robustly to body pose, shape, and translational velocity, as well as material specifications for the garment itself. Our key contribution is to put together a pipeline for virtual dressing in 3D space from a single image 2D image, using an off-the-shelf model in conjunction with a lightweight self-supervised model improving efficiency during inference as well as training. As in the SNUG model, we leverage an optimization-based scheme using physics-dependent loss terms that eliminates the need for annotated ground-truth data. We also extend the SNUG model to account for different materials, including 100% cotton, silk fiber, and sheep wool and present a virtual wardrobe of different garments in these materials that can be used for dressing in 3D space using sparse 2D data.

## 1 Introduction

Modeling draping 3D clothing on the human body has widespread applications in AR/VR content generation, as well as e-commerce, virtual try-on, gaming, etc. Due to the inability of buyers to try on clothes physically, retailers lose up to $600 billion each year due to sales returns on e-commerce websites. Additionally, online representation of how clothing fits is often not inclusive and does not offer any references for potential buyers of different physical characteristics such as limb differences, body shape, size, etc.

3D reconstruction of garments can help a person infer a particular garment's fitness by looking at artifacts like folds and wrinkles to see how the garment interacts with their body. The first step in addressing this problem is to model in 3D the pose and shape of the human body. Most methods assume known correspondence between 2D joints and the 3D skeleton [10, 11]. Such models have weak, or non-existent, models of human shape. Capturing differences in body shape can help reduce ambiguity and make it possible to model interpenetrations. The SMPLify-X [2] (CNN-based approach) model allows the estimation of a full 3D mesh of a human body from a single 2D image using bottom-up estimation followed by top-down verification. It directly optimizes pose and shape, so that the projected joints of the 3D model are close to the 2D joints estimated by the CNN.

Modeling garments digitally traditionally used physics-based simulations [5] which required high computational costs and could not be deployed for use in real-world systems. Although recent learning-based methods attempt reconstruction of people with clothing [6, 7, 8, 9], they use supervised approaches which can closely approximate the results of physical-based solutions, but require minimizing vertex-level differences between predicted garment meshes and ground-truth data. Additionally, most of these models require parameters defined in 3D space to predict garment deformations based on human pose and shape specifications and are unable to model differences based on the material of the cloth. To mitigate the need for large annotated datasets, exploring self-supervised strategies for dynamic 3D clothing (such as SNUG [3]) have proven to be very successful. We propose a pipeline that leverages some of these techniques for a 3D garment draping algorithm that adapts robustly to changes in pose, shape, garment style, and material specifications on a body inferred from a single image. We combine the more lightweight physics-based approaches like SNUG [3] with the capabilities and modularity of a generative network like SMPLify-X for a robust and realistic virtual try-on experience that can be rendered from any viewing direction. Demo videos of the full pipeline and code for 3D body reconstruction can be found at this link.

## 2 Related Work

### 2.1 Inferring the Body

There have been many approaches to solving the problem of reconstructing an accurate 3D body mesh from images. Many methods use an intermediate 2D representation such as joint locations [13] or part segmentation [14].

SMPLify [13] is one such method that utilizes a "bottom-up" "top-down" approach to reconstruct the body in SMPL format from a single image. It uses a CNN-based algorithm called Deep-cut to extract 2D joint locations, then fits a SMPL model to the joints by minimizing the error between the projected 3D joints and detected 2D joints. Further, they train a regressor which helps reduce interpenetrations of the resulting 3D mesh.

There have also been other approaches that use multiple images to reconstruct the human body. For example, Multi-Garment Net [6] takes in multiple frames from a video, extracts joint information and part segmentation from these frames, then predicts a SMPL body and separate garment template using a CNN. The resulting body and garment mesh are both modeled based on the SMPL representation which allows the pose and shape to be easily manipulable.

### 2.2 Garment Draping

Modeling how 3D garments drape on the human body using machine learning is also an active area of research. There are many works that use supervised techniques to learn garment deformations with respect to different attributes such as pose, shape, garment style, and garment size [6, 7, 8, 9, 12].

One notable model is TailorNet [12] which predicts deformations as a function of pose, shape, and garment style using a large dataset generated from physics-based simulations. Their approach involves separately predicting low-frequency and high-frequency deformations using MLPs for a particular pose, shape, and garment style, then combining these in a final, posed garment model.

Another example is SizerNet [9], a neural network that can predict garment deformations as a function of garment size, but not pose. This was trained using the SIZER dataset which contains real 3D scans of people dressed in garments of varying sizes.

These models trained using supervised techniques have shown promising results. However, their reliance on ground truth data limits their ability to predict deformations according to other garment variables such as material properties. On the other hand, a self-supervised approach as seen in SNUG [3] which learns garment deformations by optimizing physics-based loss terms can be more adaptable to different garment specifications.

## 3 Methods

Figure 1 describes the overview of our pipeline. We take a single input image and use OpenPose [4] to generate 2D joints using 135 predicted keypoints. We feed the original image and the 2D joints

into the SMPLify-X [2] model, which then reconstructs the 3D mesh of the human body and returns body shape, pose, and translation in the SMPL [1] format. We then use these parameters along with material-specific information to deform a garment template mesh and drape it over the 3D mesh.

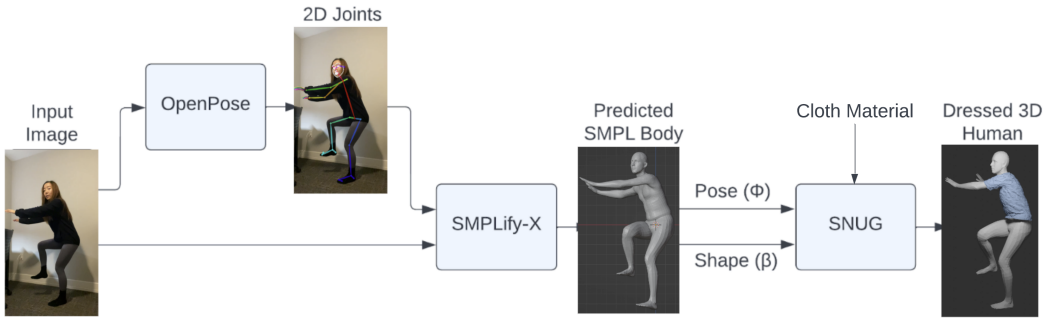


Figure 1: Overview of Our Pipeline

### 3.1 Data Preprocessing with OpenPose

Given a color input image with an entire body of a person captured, we called OpenPose [4] to extract the 2D locations of the anatomical joints of the person in the image. OpenPose is a multi-stage, feedforward CNN which extracts feature maps then uses a greedy bipartite matching algorithm to get poses of each person in an image. The architecture is designed to jointly learn part locations and their association via two branches of the same sequential prediction process.

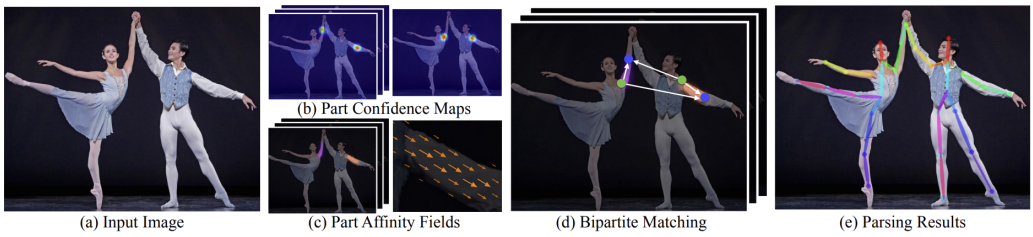


Figure 2: Overview of OpenPose Pipeline taken from [4]

OpenPose takes the entire image as the input for a two-branch CNN to jointly predict confidence maps for body part detection and part affinity fields for parts association. The parsing step performs a set of bipartite matchings to associate body parts candidates. It then assembles them into full-body poses for all people in the image.

### 3.2 3D Reconstruction from 2D Joints with SMPLify-X

We used a model called SMPLify-X [2] to predict a 3D body in SMPL format from a single input image and corresponding 2D joint information. SMPLify-X uses the same general approach as SMPLify [13], but makes various improvements that allow for the detection of additional features such as hands and facial expressions, better accuracy of prediction, and faster computation time, making it the optimal choice for our pipeline.

SMPLify-X seeks to minimize the following objective function which consists of a data term $E_J$ which represents the difference between 2D joints (obtained from OpenPose) and a projection of the 3D joints, a set of prior terms which penalize unnatural poses and shapes, and $\lambda_C E_C$ which penalizes interpenetrations.

$$E(\beta, \theta, \psi) = E_J + \lambda_{\theta_b} E_{\theta_b} + \lambda_{\theta_f} E_{\theta_f} + \lambda_{\theta_h} E_{\theta_h} + \lambda_\alpha E_\alpha + \lambda_\beta E_\beta + \lambda_\epsilon E_\epsilon + \lambda_C E_C \quad (1)$$

It improves upon SMPLify by using a more robust human body prior $E_{\theta_b}$ and a newly defined interpenetration penalty $E_C$. Their novel pose prior, VPoser, uses a variational autoencoder that regularizes the encodings of the pose parameters to be a normal distribution. Their new interpenetration penalty is able to detect collisions more accurately by using a more detailed geometric approximation of the body to check for interpenetrations. Overall, these updated loss terms both help prevent the model from predicting the body in a pose or shape that is unnatural or physically impossible.

Although SMPLify-X is capable of detecting and representing detailed facial expressions and hand poses in the unified SMPL-X format, we only use parameters from the basic SMPL model in our pipeline. We extract the pose($\theta$), shape($\beta$), and global translation($t$) parameters from the output of SMPLify-X, and use these as the base body template to drape the garments onto.

### 3.3 Garment Draping using SNUG (Self-Supervised Neural Dynamic Garments)

We implemented our own version of SNUG for garment deformations. SNUG is a self-supervised model that computes physics-based deformations in the garment template based on material specifications and motion sequences. The advantage of a lightweight model like SNUG is that it removes the need for ground truth samples and realizes physics-based deformation models as an optimization problem rather than by implicit integrators. Leveraging these loss functions allows the model to model interactive garments with dynamic deformations and fine wrinkles adaptable to pose, shape, and material specifications.

Our model is also built on the SMPL representation of a human body, in which $M(\cdot)$ represents the human body made of 6890 mesh vertices as a function of pose($\theta$), shape($\beta$), global translation($t$). The body motion $\phi$ contains the current body pose $\theta$ as well as the global velocity of the root joint:

$$M(\beta, \phi) = W(T(\beta, \phi, D), J(\beta), \theta, W_G) \tag{2}$$

$$T(\beta, \phi) = T + R(\beta, \phi) \tag{3}$$

SNUG recasts solutions to the equations of motion discretized with backward Euler as an optimization problem, and uses the objective function for this minimization as the loss function for a self-supervised training approach.

$$x^{t+1} = x \arg\min \frac{1}{2\Delta t^2}(x - x\hat{})^T M(x - x\hat{}) + \Phi \tag{4}$$

Where $x$ and $v$ are position and velocity of garment nodes respectively, $x\hat{} = x^t + \Delta t v^t$ is a tentative (explicit) position update, and $\Phi$ is the potential energy due to internal and external forces $f$ of the system, $M$ is the mass matrix.

Loss terms:

$$L = L_{inertia} + L_{static} \tag{5}$$

Where $L_{inertia}$ models the inertia of the garment and and is defined as:

$$L_{inertia} = \frac{1}{2\Delta t^2}(x - x\hat{})^T M(x - x\hat{}) \tag{6}$$

$L_{static}$ models the potential energy term in Eq. 10 as:

$$L_{static} = L_{strain} + L_{bending} + L_{gravity} + L_{collision} \tag{7}$$

$L_{strain}$: Corresponds to the membrane strain term which models the response of the material to in-plane deformation based on a first-order deformation metric.

$L_{bending}$: Models the energy due to the angle of two adjacent faces as a function of the dihedral angle of the faces and the bending stiffness.

$L_{gravity}$: Models the effect of gravity in the learned deformations as an additional loss term based on the gravitational potential energy.

$L_{collision}$: This is the collision penalty enforcing the garment to follow the underlying body motion based on the distance to the underlying body, and collision stiffness. It also takes into account a safety term to prevent the garment from overlapping with the body surface.

The regressor is implemented using 4 Gated Recurrent Units (GRU), each with an output of size 256, and tanh as the activation function. Our model was trained on the 10 sequences from the AMASS dataset consisting of 9000 total frames, split into subsequences of 3 frames each, and uses randomly sampled body shapes at each epoch.

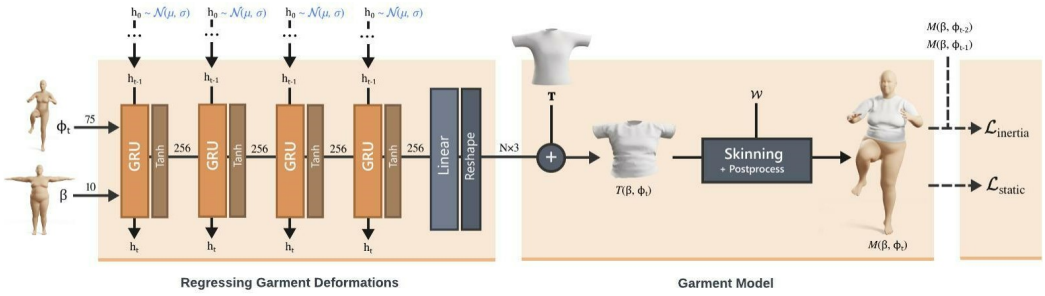


Figure 3: SNUG Model architecture as taken from [3]

Due to RAM considerations, we used a batch size of 8, and trained for 5 epochs using a learning rate of 0.001. The mass matrix specifications used were for 100% cotton fabric, silk fibers, and sheep wool. We trained models for 2 garments - namely t-shirts and pants - for each of the three materials. The material parameters used for the same are described in Table 1:

| Material | Thickness (m) | Bulk Density (kg/m$^3$) | Young's Modulus (GPa) | Poisson's Ratio |
|---|---|---|---|---|
| 100% Cotton | 4e-5 | 426 | 7 | 0.485 |
| Silk Fibers | 6e-5 | 510 | 14 | 0.475 (hyperelastic) |
| Sheep Wool | 2e-5 | 600 | 3.1 | 0.475 (hyperelastic) |

Table 1: Material Specifications

For our ablation studies, we also modified the network architecture by increasing the number of GRU layers and trained with the same hyper-parameters to observe the effect on the loss terms for a quantitative evaluation of the model.

## 4 Experiments and Results

We are able to successfully run through the entire pipeline, inputting a single 2D image and obtaining a predicted 3D body with a garment of choice draping on top of the body. We evaluate the performance of our model both qualitatively and quantitatively.

### 4.1 Qualitative Analysis

For qualitative evaluation, we chose to perform visual inspection as fitting of clothes can be assessed well by the eyes (in fact, this is what we do when we try on a piece of clothes). We observe that OpenPose precisely extracts the 2D positions of human joints in the input 2D image. We also observe that the 3D reconstructions of SMPLify-X model accurately capture the pose and body shape of the figures from the 2D images, as seen in Figure 4 (a) and (b). We conclude that OpenPose and SMPLify-X are able to work seamlessly together in our project domain, converting 2D human poses

to 3D. To evaluate the performance of our garment draping model, we ran ablation studies to study the effects of various parameters.

### 4.1.1 Evaluating against Pre-Trained SNUG Model

In this study, we evaluated our implementation of the original model architecture with the same material parameters as specified in the SNUG paper for 100% cotton to evaluate our baseline.

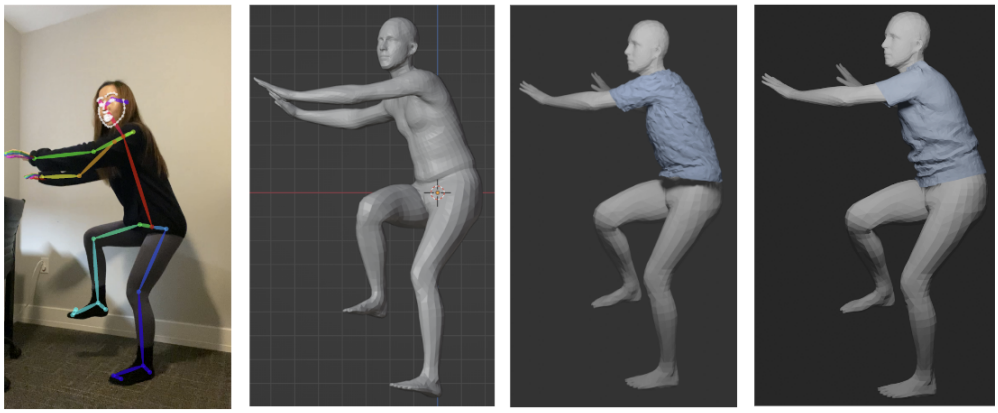


Figure 4: Pipeline Results for Original SNUG Architecture

Figure 4 (c) shows the results of our first model with four layers of GRUs, in comparison with Figure 4 (d), the pre-trained SNUG model provided in the SNUG codebase. Our implementation of the model produced folds in realistic positions across a variety of dynamic poses, similar to the pre-trained SNUG model provided by the authors.

With our implementation of the four-layer SNUG model, we believe there is room for improvement in terms of adapting to different body shapes and creating realistic folds and wrinkles. We notice that our model tends to add many excess small deformations due to a lack of convergence in the membrane strain energy loss. We hypothesize that this may be due to a lack of computational resources when training, since the original SNUG implementation used a batch size of 16 on a 32GB system, with close to 6900 frames for over 10 epochs.

Additionally, when testing with different body shapes, we noticed that there were interpenetrations between the garment and body vertices during inference. This is likely due to the model not generalizing to different body shape parameters in our implementation of linear blend skinning, causing posed vertices to overlap with each other. Despite the fact that our loss penalizes collisions between the garment and the body in the training samples, we found noticeable collisions in test motions.

### 4.1.2 Evaluating New Model Architecture

We further implemented a new SNUG variation with an increased number of GRU layers and compared it to our baseline implementation results. The first model with the original architecture produced the expected results with some interpenetrations between body and garment vertices as seen in Figure 5 (a). Adding an additional GRU layer improved the definition in the deformations, but led to more interpenetrations as illustrated in Figure 5 (b). We hypothesize that due to the increased complexity of the network, we were able to model more fine-grained deformations, however, to train a more complex network we need more training data and need to train for longer. We concluded that the added layer of GRU in the SNUG model training alone did not improve qualitative performance.

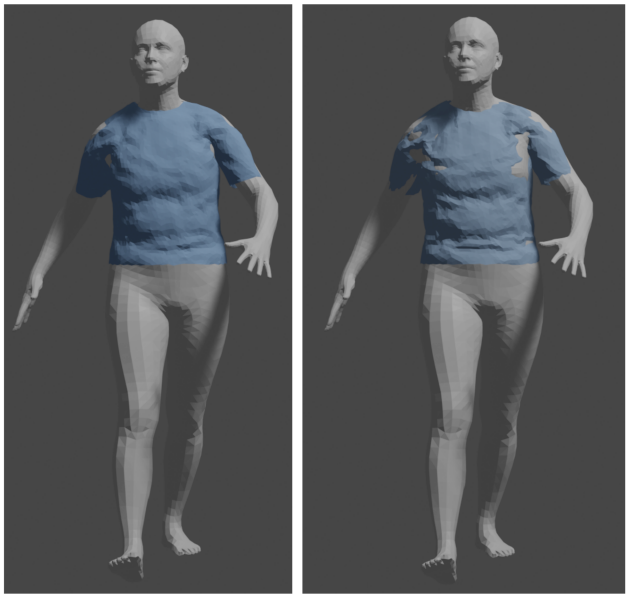


Figure 5: Qualitative Comparison of Models for Cotton T-Shirt

#### 4.1.3 Evaluating Material Variations

In this study, we evaluated the ability of our implementation (with the original network architecture) to adapt to different materials, namely silk, cotton and wool as illustrated in Figure 6. We notice that our model is able to successfully adapt to different material specifications, with physical property-dependent deformations. Deformations in the silk pants (left) are much smoother, while the wool pants (right) look bulkier with more defined deformations.

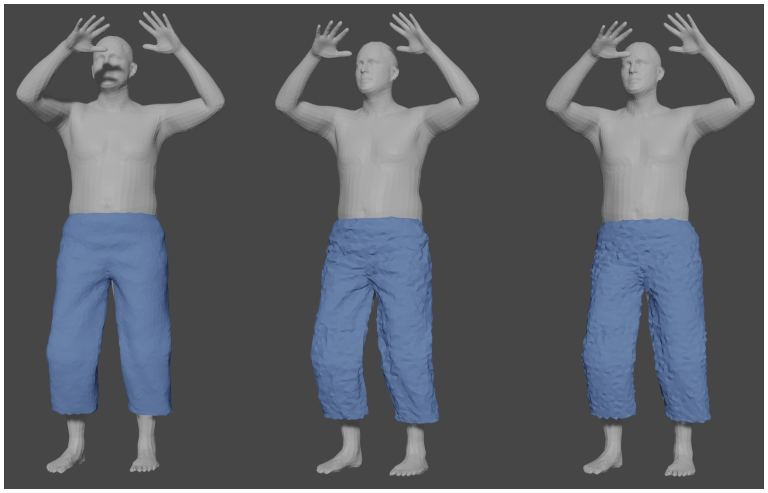

Figure 6: Material Comparisons: Silk (left), Cotton (center), Wool (right)

### 4.2 Quantitative Analysis

For quantitative evaluation, given that we are using OpenPose, and SMPLify-X which are well-established and well-researched models, we focused our efforts on the performance of the SNUG model where we implemented the training module and model architecture. We quantitatively evaluated

our results by examining the physics-based total loss term $L$ in the two versions of models we trained. The loss term of the first model we trained is shown on the left in Figure 7. The model is built of four layers of Gated Recurrent Units (GRU). The total loss converges close to 0.083 for the cotton t-shirt model.

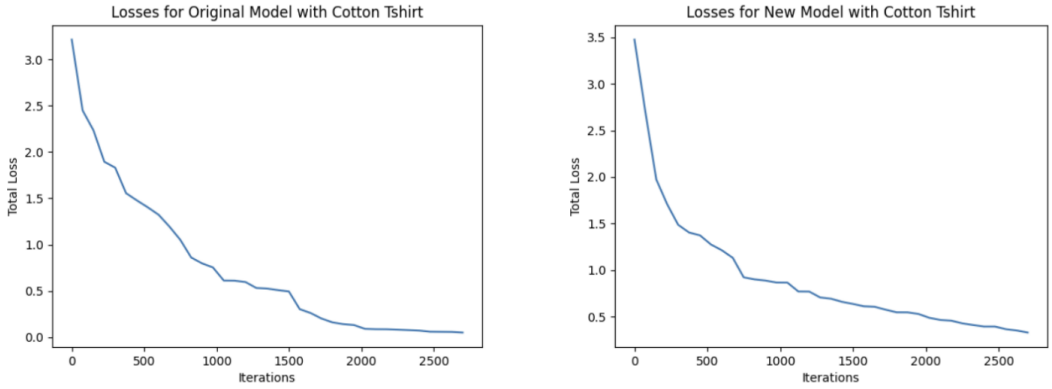


Figure 7: Losses over Iterations for 4-layer model (left) and 5-layer model (right)

In the second SNUG model that we trained with a network architecture of five layers of Gated Recurrent Units (GRU), the loss converged to 0.257. The loss term of the second model we trained is shown in Figure 7 (right). We hypothesize that this happens despite the decrease in membrane-strain loss because the collision penalties increase, which is consistent with the qualitative results.

## 5 Limitations and Conclusion

We were able to successfully combine three models - OpenPose, SMPLify-X, and SNUG - and build a pipeline that takes in a single 2D image/video frame of a human and produces the 3D reconstruction for the same body with a garment of choice draped on the body. In our implementation of the SNUG training module, we were able to successfully train the model to learn garment deformations and adapt to body poses and materials. Our model also produced folds in realistic positions across a variety of dynamic poses. With the ability to visualize clothing fit in 3D by simply inputting a single 2D image, our project can be adapted to e-commerce platforms and online clothing stores to help increase accuracy and confidence when shoppers select clothing sizes and styles.

A limitation of our model is the lack of computational resources; we hypothesize that increasing the batch size further will significantly improve our results, leading to more fine-grained approximations of deformations and fewer interpenetrations. Our model optimizes thousands of frames simultaneously during training at once, which makes our approach more prone to converge to simpler local minima. Another limitation of our model is the assumption that material properties stay constant during motion. Material properties are likely to change upon stretching e.g. spider silk fibers have Poisson's ratio between 0 to 1.52 during stretching as it is a hyperelastic isotropic material. Our model assumes a constant value of material parameters, which can lead to slight inaccuracies in deformations.

Regarding future work, we believe that exploring the possibility of enforcing collision penalties as a constraint on the network rather than a post-processing step would be valuable in achieving better deformations. We would also like to extend the model to include additional synthetic materials such as nylon and polyester. Additionally, we would like to explore the full potential of the SMPLify-X model in using expression and hand reconstructions for more realistic results that can potentially be used in real-time applications of virtual draping. Finally, we would like to incorporate lighting and color accuracy into the pipeline, forming a more holistic online clothing shopping experience for the customers and real-time relighting models for AR/VR applications.

## 6 Acknowledgements

We would like to thank Professor Marios Savvides and teaching assistants Fangyi Chen, Han Zhang, and Sai Mitheran Jagadesh Kumar for their support throughout the project. We would also like to thank SNUG author Dan Casas for his guidance over the course of the semester.

## References

[1] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, Michael J. Black. SMPL: A Skinned Multi-person Linear Model. ACM Transactions on Graphics Vol. 34, No. 6, Article No. 248, 2015.

[2] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman , Dimitrios Tzionas, and Michael J. Black. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019

[3] Igor Santesteban, Miguel A Otaduy, and Dan Casas. SNUG: Self-Supervised Neural Dynamic Garments. In Proc. of Computer Vision and Pattern Recognition (CVPR), 2022.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime Multi-person 2d Pose Estimation using Part Affinity Fields. In Proc. of Computer Vision and Pattern Recognition (CVPR), 2017.

[5] Andrew Nealen, Matthias Muller, Richard Keiser, Eddy Boxerman, and Mark Carlson. Physically Based Deformable Models in Computer Graphics. Computer Graphics Forum, 2006.

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

[7] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to Dress 3D People in Generative Clothing. In Proc. of Computer Vision and Pattern Recognition (CVPR), 2020.

[8] Igor Santesteban, Miguel A. Otaduy, and Dan Casas. Learning-Based Animation of Clothing for Virtual Try-On. Computer Graphics Forum (Proc. Eurographics), 2019

[9] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. SIZER: A Dataset and Model for Parsing 3D Clothing and Learning Size Sensitive 3D Clothing. In Proc. of European Conference on Computer Vision (ECCV), 2020.

[10] Fan, X., Zheng, K., Zhou, Y., Wang, S.: Pose locality constrained representation for 3D human pose reconstruction. In Proc. of European Conference on Computer Vision (ECCV), 2014.

[11] Zhou, X., Zhu, M., Leonardos, S., Derpanis, K., Daniilidis, K.: Sparse representation for 3D shape estimation: A convex relaxation approach. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), 2015.

[12] Chaitanya Patel, Zhouyingcheng Liao, Gerard Pons-Moll. TailorNet: Predicting Clothing in 3D as a Function of Human Pose, Shape, and Garment Style. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR, 2020.

[13] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, Michael J. Black. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In European Conference on Computer Vision (ECCV), 2016.

[14] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In 3DV, 2018.