Template-Based 3D Cloth Shape Estimation from Single View

Yuwei Zeng
A0267614Y

Abstract

Accurate deformable object state estimation can the bottleneck for deformable object manipulation. In this project we proposed a hybrid method that combines data-driven deformation prediction with optimization based shape estimation via differentiable rendering by learning the initialization then refine via the model-based optimization. It recovers richer shape details compared to purely learning based method which constantly generates over-smooth pattern, while significantly prevents the optimization from being stuck in local minimum. The method effectively reduces the vertex position error from 0.58 to 0.13. Qualitatively, it also demonstrates the ability to predict shapes with significantly highly resemblance compared to both approaches.

1 Introduction

Deformable object manipulation is one of the most essential skills in human life and for assistive robot to master. However, it remains a great challenge. One of the bottlenecks falls into accurate shape estimation for policy decision making, as deformable objects possesses infinite degrees of freedom and the state constantly contains severe self-occlusion [6]. Meanwhile, with the progress in model-based learning, it reduces complex tasks learning from hours to minutes when provided with an accurate state[14, 28]. Constantly we turned the question of good policy learning into how can we learn such accurate state?

This paper studies the problem of 3D shape estimation for cloth in the format of deformation prediction from a template mesh given single-view observation. The template-based approach is motivated by recent works on template-based dressing synthesis which are able to model and even animate variate clothing deformation realistically under different human configurations.[4, 20, 17] In the meantime, the choice of using mesh format is based on the fixed topology information can be viewed as a densely-annotated landmarks which suits the use case on deformable object manipulation, especially on reconfiguration tasks that require landmarks on a category level. The number and types of landmarks may also vary depending on the goal and the precision requirements that the dense annotation is a flexible format to connect or reuse across different tasks.

To reconstruct the 3D shape from 2D visual observation(s) is an inverse rendering problem, which aims to estimate physical attributes of the scene such as geometry, reflectance, lighting and camera poses from image(s).[13, 15] With recent successes of variate techniques developed on differentiable renderer development, it is a promising tool to tackle inverse rendering tasks that converts reconstruction into an optimization problem guided by pixel-level perceptual loss.[15, 5, 30] However common to other global optimization problems, direct optimization with differentiable rendering is known to suffer from ragged task landscapes and constantly stuck in local minimum that fails to faithfully reconstruct the target shape. [1]

This project proposes a hybrid approach by answering the question "Can we learn a good initialization for such optimization with a data-driven method?". An end-to-end neural network is designed to

Preprint. Under review.

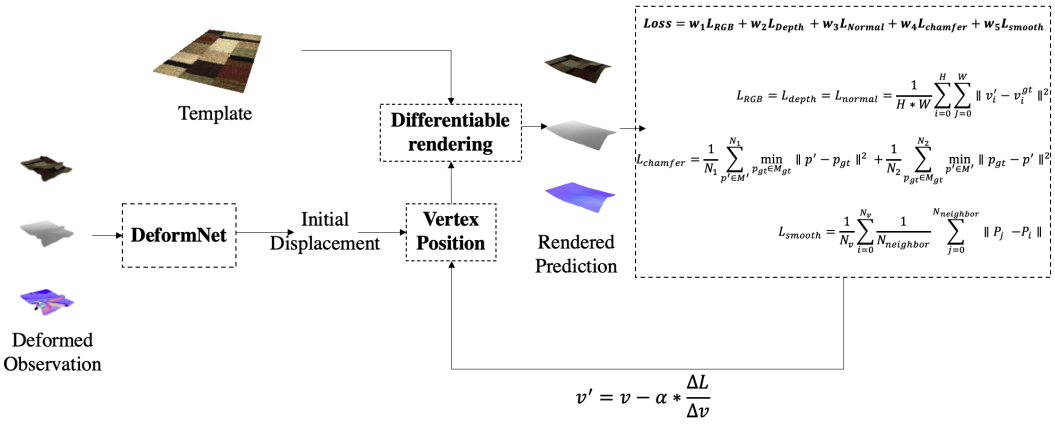

Figure 1: **Overall workflow**

predict vertex displacement from a template based on single-view RGB, depth and normal images. An automatic cloth perturbation and rendering is developed for deformed object generation in a physics-based simulator and provide the training data. By training the network with geometric losses and in a data-driven way, the network learns to predict more physically plausible deformation and nearer to the global optimal point. The prediction is passed as the initialization to shape optimization with differentiable rendering for further optimization guided by perceptual loss.

Both quantitative and qualitative analyses are conducted to validate this work. It shows the purely data-driven method tends to learn low-frequency features and generates over-smooth reconstruction, while the purely optimization-based method tends to get stuck in local minimum that generates shapes with poor resemblance to the target ones. Combining these two methods complements each other and improves the reconstruction quality significantly both in terms of the mean vertex position error and visual resemblance.

In summary, the contributions of this paper include:

- A hybrid approach that combines the data-driven method with the optimization-based method for deformed cloth shape estimation. This is done by providing the prediction from the data-driven method as the initialization to optimization later, so it is less affected by local minimum and continues to optimize fine details with a fully differentiable renderer.
- An encoder-decoder-based network for deformation prediction with automatic data generation from simulation. It includes a few modifications to both network architecture and loss terms to address the high-frequency feature learning issues.

2 Related Work

3D Reconstruction from Images Multiple lines of works have tackled on shape learning from multiple view, a sequence of images and even single image. In human and cloth shape estimation tasks such as BCNet[11] and Multi-Garment Net[4], it maps semantically segmented image to latent code corresponding to body shapes, 3D poses and garment style, then predict un-posed garment from it and refine the retarget garment with different consistency losses. In the field of rigid object reconstruction, multiple works have illustrate the efficacy of jointing optimization of shape and texture via differentiable rendering.[18, 30, 5] Among them, one sub-field that is particular relevant to this project is Shape from Template (SfT)[8, 2], majority of them involve hand-crafted geometric constraints and solve it through convex optimization.[23, 22, 29, 2] Some recent works also learns such surface vertices with neural network with a constrainted latent space or Gaussian Processes that implicitly satisfying a set of quadratic equality constraints[24]. The loss design of this method is largely inspired by multiple works from this line of work.

Clothing and Deformation Modeling Deformation modeling is the key topic for dressing synthesis in the field of digital human. Clothing are in generally modelled as displacement from SMPL body model[16]. Due to the nature of garments, it is typically modelled using template based approach

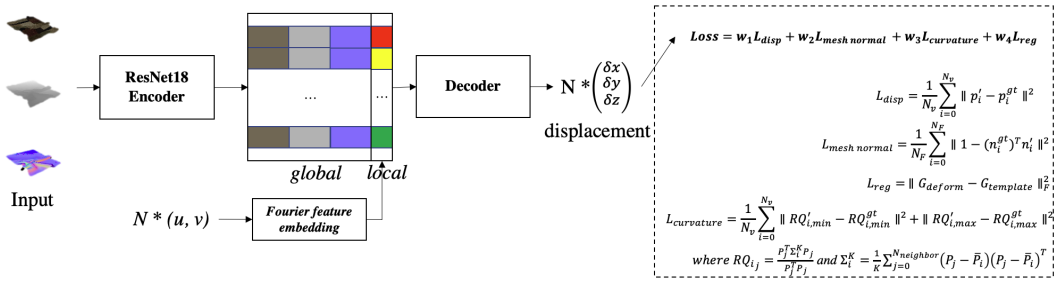

Figure 2: *DeformNet* model architecture and workflow

and parameterize the displacement according to body shape, pose, garment style and sometimes cloth dynamic parameters[4, 20, 9]. One key challenges for deformation modeling is high-frequency wrinkle modeling. Multiple works proposed to tackle as a high resolution texture mapping or normal mapping task[12, 27]. Others have modelled fine wrinkles procedurally to approximate wrinkling phenomena on skin and cloth[20, 19]. However those works are either heavily rely on human pose and shape parameterization which is absent in this case, or fine details are generated with a more accurate visual mapping but the underlying mesh shape is unchanged. In parallel, in N-Cloth [7] the author tooks away the deformation from poses assumption and expanded into a general deformation mechanism. It learns in a data-driven manner with an encoder-decoder architecture. This project is inspired by this line of work.

3 Approach

As illustrated in Fig 1, the pipeline consists of two main components. The input images are firstly passed to data-driven displacement prediction network as DeformNet to produce the initial vertex positions. It is further passed to differentiable rendering for shape optimization. In this section, we firstly introduce the architecture and losses for DeformNet. Then we will introduce the losses and optimization scheme under differentiable rendering

3.1 DeformNet

Assuming there is no scale change, the DeformNet takes input as RGB, depth and normal images from a single view of the segmented target object. It predicts vertex displacements from their canonical positions in template mesh. An encoder-decoder architecture is chosen here to generate 3D vertex displacements from image inputs. Three modalities of latent features are concatenated together to form a global feature of the deformed object and further passed to the decoder for displacement prediction for each vertex.

While the task is a well-formed supervised learning task that can be fully guided with ground truth vertex displacement loss, numerous existing works have demonstrated that neural networks tend to learn low-frequency information only. It is reflected on garment generation tasks as over-smooth shapes being generated and challenging to recover high-frequency features such as wrinkles, which is utterly important for this task. Such over-smooth shapes are also observed when DeformNet is plainly trained with ground truth displacements only. To address this, validated as a simple yet effective technique to address high-frequency feature learning on low-dimensional regression tasks[26], a Fourier feature transformation of the UV coordinates for each vertex is added to extract the local feature and concatenated with the global feature before passing to the MLP based decoder. In addition, two more geometric loss terms emphasizing the wrinkle features on mesh normals and curvatures are added.

Last, a regularization term is added to enforce Riemannian metric preservation between the deformed and template surface. It is to ensure the deformation is length preserving that no surface extends or shrinks from the deformation so that shape estimation will be unambiguous from the template.[3]

Vertex Displacement Error For each vertex position $p = p_{template} + \delta_{distplacement}$, denoting the prediction displacement as $p^{'}$, ground truth as p^{gt}, and the total number of vertices as N_v, the vertex

displacement error is computed according to the vertex position distance:

$$L_{disp} = \frac{1}{N_v} \sum_{i \in N_v} ||p_i^{'} - p_i^{gt}||^2 \tag{1}$$

Mesh Normal Loss Given a triangle face $[p_0, p_1, p_2]$ from a mesh, the face normal is computed as $\vec{n} = \overrightarrow{p_1 - p_0} \times \overrightarrow{p_2 - p_1}$. Denoting the total number of faces as N_F, the mesh normal loss is defined as:

$$L_{mesh\ normal} = \frac{1}{N_F} \sum_{i \in N_F} ||1 - (n_i^{gt})^T n_i^{'}||^2 \tag{2}$$

Mesh Curvature Loss The Rayleigh quotient (RQ) curvature is adopted here compared to alternative representations such as Gaussian curvature or eigen curvature due to its superior numerical stability and differentiability on both flat surface and avoiding unstable value decomposition by transferring it into an optimization process. More detailed analysis and visualization can be found in the original paper[9].

With the center vertex denoting as p_i, neighbor vertex as p_j, mean neighbor position as $\hat{p_i}$, the number of neighbour vertices as N_K, The RQ of a local neighbor region is defined as:

$$RQ(\Sigma_i^{N_k}, p_j) = \frac{p_j^T \Sigma_i^{N_k} p_j}{p_j^T p_j} \tag{3}$$

where covariance matrix for vertex i is:

$$\Sigma_i^{N_k} = \frac{1}{N_k} \sum_{j \in N_k} (p_j - \hat{p_i})(p_j - \hat{p_i})^T \tag{4}$$

The curvature loss is computed by measuring the curvature statistics consistency between both meshes. Following [9], it is further relaxed to capture the consistency between the minimum local RQ and maximum local RQ as:

$$L_{curvature} = \frac{1}{N_v} \sum_{i \in N_v} (RQ_{min}^{p_i^{'}} - RQ_{min}^{p_i^{gt}})^2 + (RQ_{max}^{p_i^{'}} - RQ_{max}^{p_i^{gt}})^2 \tag{5}$$

Isometry Surface Regularizer With G denoted as the 2x2 discrete Riemannian metric of the mesh surface, the regularization loss is computed as:

$$L_{reg} = \frac{1}{4} \sum_{i=0}^{4} ||G_i^{deform} - G_i^{template}|| \tag{6}$$

In summary, the total geometric loss for DeformNet is a weighted aggregation of the four loss terms:

$$L_{total} = w_1 L_{disp} + w_2 L_{mesh\ normal} + w_3 L_{curvature} + w_4 L_{reg} \tag{7}$$

3.2 Shape Optimization with Differentiable Rendering

Given the set of single-view target images, the differentiable rendering pipeline takes input vertex positions and renders under the same camera and lighting condition. The rendered outcomes are fully differentiable with respect to the input geometry such as the vertex positions used here. By comparing the perceptual differences between the rendered predicted cloth and the target cloth, it back-propagates the gradient of the loss with respect to vertex positions and updates the shape prediction. This update is done iteratively until prediction and losses converge. The perceptual losses include the following terms:

Pixel Consistency Loss RGB, depth and normal images are compared pixel by pixel and averaged across the image that:

$$L_{pixel} = \frac{1}{H \times W} \sum_{i=0}^{H} \sum_{j=0}^{W} ||(v_{ij}^{'} - v_{ij}^{gt}||^2 \tag{8}$$

Chamfer Loss The Chamfer distance is introduced here to enforce similar shapes between predicted and perceived shape from RGB-D images with known camera intrinsics and extrinsics.

$$L_{chamfer} = \frac{1}{N_1} \sum_{p^{'} \in M^{'}}^{N_1} \min_{p^{gt} \in M^{gt}} ||p^{'} - p^{gt}||^2 + \frac{1}{N_2} \sum_{p^{gt} \in M^{gt}}^{N_2} \min_{p^{'} \in M^{gt}} ||p^{gt} - p^{'}||^2 \tag{9}$$

Mesh Smoothing Solely optimizing the shape with respect to pixel and chamfer losses yields non-smooth shapes with sharp edges. Laplacian smoothing[25] is added to improve such issue with:

$$L_{smooth} = \frac{1}{N_v} \sum_{i \in N_v} \frac{1}{N_k} \sum_{j \in N_k} ||p_j - p_i|| \tag{10}$$

Similarly, the total loss is a weighted aggregation of all 5 terms:

$$L_{total} = w_1 L_{rgb} + w_2 L_{depth} + w_3 L_{normals} + w_4 L_{chamfer} + w_5 L_{smooth} \tag{11}$$

4 Experiments

The project is confined on the table-top deformation setting. The hybrid approach is evaluated quantitatively and qualitatively and compare it to purely optimization-based and purely data-driven outcome. The template here is a 4×6 single-layer cloth due to time limitation and intensive computation involved on self-collision for double-layer cloth on data generation, but double sided rendering is used here to differentiate it as 3D shape estimation from single surface reconstruction task. DiffCloth[14] is used as the physics based cloth simulator. For DeformNet, the encoder uses ResNet-18[10] architecture, and decoder is a 5-layer ReLU activated MLP layers. The Fourier feature uses 10 frequency resolutions. The differentiable renderer used here is the optimized version of SoftRasterizer[15] implemented in PyTorch3D[21]. 2000 update iterations are used for the optimization after observing the convergence behaviour.

For loss weights, the weights were adjusted so all the losses scale to approximately the same range. For DeformNet, it uses w_{disp} = 1, $w_{mesh\ normal}$ = 0.3, $w_{curvature}$ = 1000, w_{reg} = 100. For shape optimization with differentiable rendering, w_{rgb} = 0.2, w_{depth} = 1, $w_{normals}$ = 0.1, $w_{chamfer}$ = 1, w_{smooth} = 0.01.

4.1 Training Data Generation

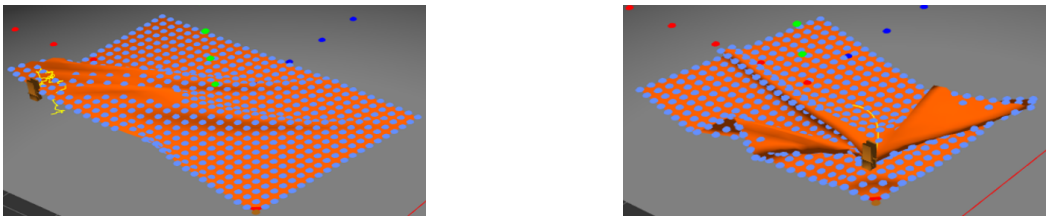

Figure 3: Cloth perturbation with *(left)* random trajectory; *(right)* targeted trajectory parameterized by the Bezier curve. The trajectory is visualized in yellow.

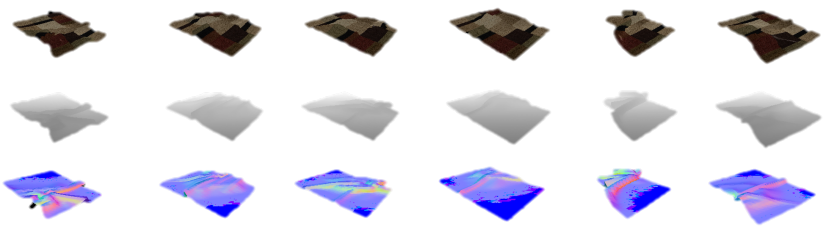

Figure 4: Rendered RGB, depth and normals images of the generated deformed cloth

An automation script are implemented to generate the training data for DeformNet. Two types of control trajectory on random vertex selected were implemented as visualized in Fig 3: (1) Random

steps; (2) Targeted trajectory parameterized by three points in Bezier Curve to mimic more realistic cloth manipulation. The trajectory is determined by three points: starting point which is the vertex being controlled; ending point as a vertex randomly selected with distance within [1, 3.5], and a middle high point that has plane coordinates between starting and ending points, and a random lift height in [0.1, 0.3]. The final deformed object meshes are saved and rendered. In total 10,000 pairs of RGB, depth and normals images with ground-truth displacement were generated as shown in Fig 4.

4.2 Quantitative and Qualitative Analysis

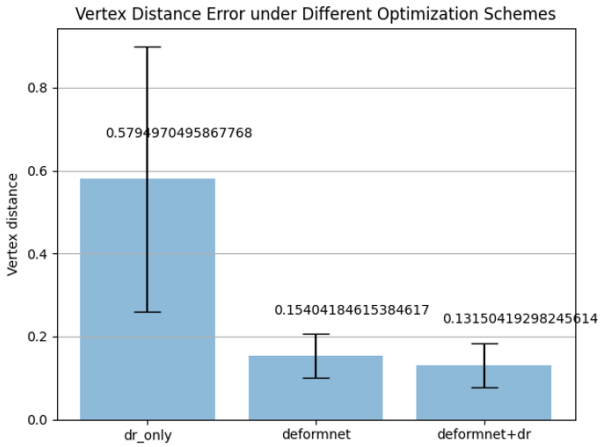

Figure 5: Mean vertex position error.

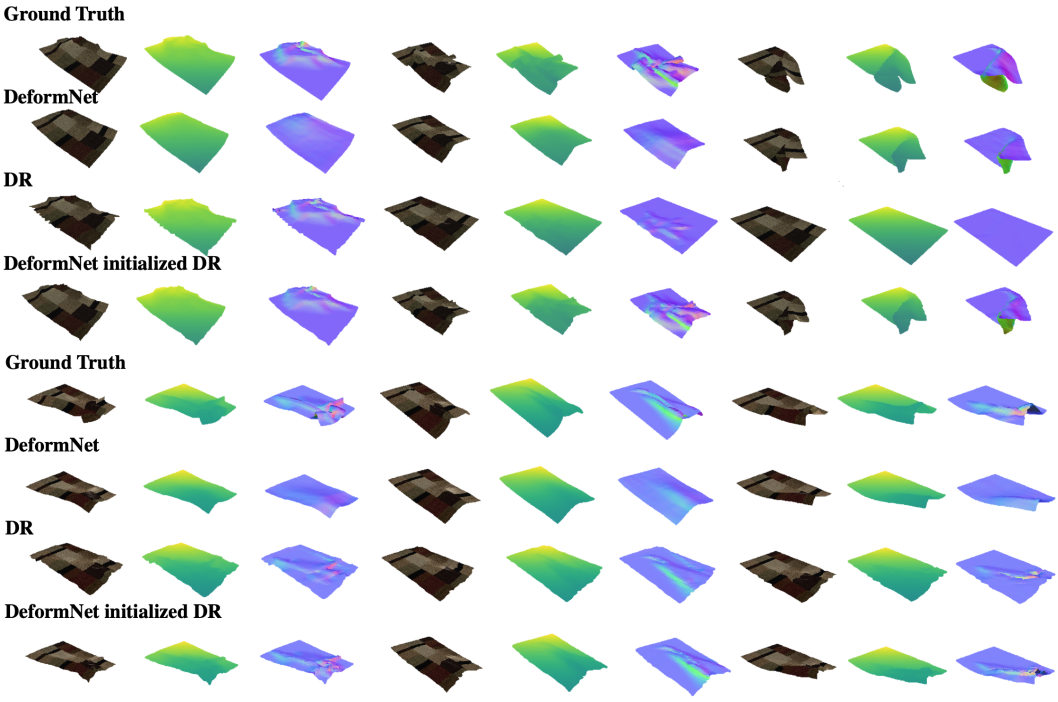

Figure 6: Rendering of the predicted shapes (rgb, colorised depth, normals) vs ground truth

The mean and standard deviation of vertex position prediction error on 1,000 unseen deformed cloth pieces is plotted in Fig 5. During the experiment, it is found the pure optimization method

constantly stops in an early or intermediate configuration with loss stopping descending. Such a phenomenon is especially noticeable when there is a sharp flip or large overlaying region on the deformed pieces. This is likely due to the ambiguity from the inverse rendering problem that multiple solutions exist with occlusion, and it is more significant with larger occlusion region. Meanwhile, a purely optimization-based method tends to generate more rugged cloth pieces such as sharp edges along the curves and borders likely due to the lack of physics prior or constraints. Though this can be improved with a larger smoothing coefficient it does not solve this problem fundamentally and may lead to less accurate prediction. It is also found that DeformNet still lacks the ability of accurate wrinkle learning and tends to generate over-smooth shapes. Though the error statistics also reduced to a relatively low level, as a wilder and smoother wrinkle may yield the same error stats compared to a narrower and sharper wrinkle region due to mean operation, it shows significant visual quality differences as illustrated in Fig 6. By combining these two, the initialization from the DeformNet which embeds more physically plausible deformation from its training data effectively enables the optimization to move towards a more globally optimal solution. It generates shapes that have much higher resemblance and fine deformation details with smaller vertex errors.

4.3 Ablation Study

Though it suggests accurate high frequency features learning is not fully solved and more works are needed, the modifications on DeformNet to address this issues are studied here to exam the efficacy. It is evaluated on image depth and normal losses. Adding the mesh normal, curvature related losses are found to improve this significantly.

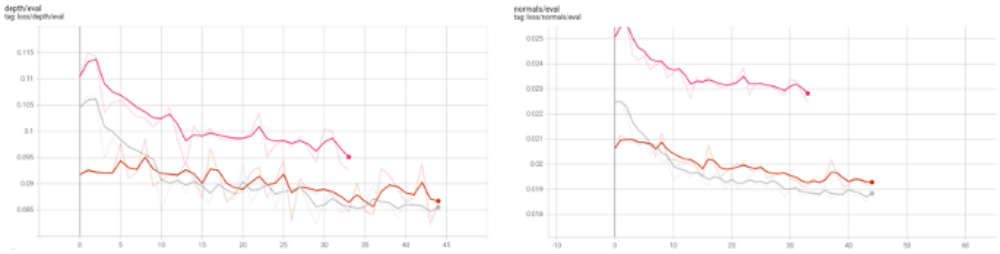

Figure 7: (left) Image depth , (right) image normals losses for (orange) using offset loss only; (pink) adding normal and curvature error; (gray) adding normal, curvature error and uv Fourier feature embedding.

5 Conclusion

This project presents a hybrid method for template-based 3d shape estimation for deformed cloth from single-view observation as RGB, depth and normal images. Optimization-based method via differentiable rendering has the potential to reconstruction sharp features from pixel consistency but constantly stuck at local minimum and tends to generate physical implausible rugged shapes from our experiment. We trained a data-driven deformation prediction from physics-based simulation data and use the prediction as initialization to guide the optimization. It significantly alleviate the local optimal issue and generates both quantitatively and qualitatively better results.

Various techniques on addressing high-frequency feature learning were also explored during this project and more work are expected on this for the future work. One potential idea is to generate frequency or deformation related weighted mask for the loss calculation as one cause for such over-smooth generation is the loss is averaged across all points. The loss is ambiguous with respect to the wrinkle shape in this way and requires ways to remove such ambiguity. Besides, though this work contains small translation and rotation offsets, it assumes same scale which constrains its generalizability. An pose and scale fitting scheme might be considered to add for different object usage. Last but not least, during the project, a preliminary test was implemented to remove the RGB from DeformNet as a fixed texture affects the prediction generalizability which limits such template-based method. The result from the preliminary test shows DeformNet performance was only mildly affected after removing the RGB features. Formal works are expected to validate this and enable a better generalizability of this approach.

References

[1] Rika Antonova, Jingyun Yang, Krishna Murthy Jatavallabhula, and Jeannette Bohg. Rethinking optimization with differentiable simulation from a global perspective. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 276–286. PMLR, 14–18 Dec 2023. URL `https://proceedings.mlr.press/v205/antonova23a.html`.

[2] Oriol Barbany, Adrià Colomé, and Carme Torras. Deformable surface reconstruction via riemannian metric preservation. *arXiv preprint arXiv:2212.11596*, 2022.

[3] Adrien Bartoli, Yan Gérard, François Chadebecq, and Toby Collins. On template-based reconstruction from a single view: Analytical solutions and proofs of well-posedness for developable, isometric and conformal surfaces. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2026–2033. IEEE, 2012.

[4] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5420–5430, 2019.

[5] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. *Advances in neural information processing systems*, 32, 2019.

[6] Cheng Chi, Benjamin Burchfiel, Eric Cousineau, Siyuan Feng, and Shuran Song. Iterative residual policy for goal-conditioned dynamic manipulation of deformable objects. In *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[7] Y D. Li, Min Tang, Yun Yang, Zi Huang, R F. Tong, Shuang Cai Yang, Yao Li, and Dinesh Manocha. N-cloth: Predicting 3d cloth deformation with mesh-based networks. In *Computer Graphics Forum*, volume 41, pages 547–558. Wiley Online Library, 2022.

[8] Mathias Gallardo, Daniel Pizarro, Adrien Bartoli, and Toby Collins. Shape-from-template in flatland. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2847–2854, 2015.

[9] Erhan Gundogdu, Victor Constantin, Shaifali Parashar, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. Garnet++: Improving fast and accurate static 3d cloth draping by curvature loss. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (1):181–195, 2020.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. 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*, pages 18–35. Springer, 2020.

[12] Zorah Lahner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European conference on computer vision (ECCV)*, pages 667–684, 2018.

[13] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 37(6): 222:1–222:11, 2018.

[14] Yifei Li, Tao Du, Kui Wu, Jie Xu, and Wojciech Matusik. Diffcloth: Differentiable cloth simulation with dry frictional contact. *ACM Transactions on Graphics (TOG)*, 42(1):1–20, 2022.

[15] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.

[16] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[17] Meysam Madadi, Hugo Bertiche, Wafa Bouzouita, Isabelle Guyon, and Sergio Escalera. Learning cloth dynamics: 3d+ texture garment reconstruction benchmark. In *NeurIPS (Competition and Demos)*, pages 57–76, 2020.

[18] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020.

[19] Xiaoyu Pan, Jiaming Mai, Xinwei Jiang, Dongxue Tang, Jingxiang Li, Tianjia Shao, Kun Zhou, Xiaogang Jin, and Dinesh Manocha. Predicting loose-fitting garment deformations using bone-driven motion networks. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

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

[21] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020.

[22] Mathieu Salzmann and Pascal Fua. Reconstructing sharply folding surfaces: A convex formulation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1054–1061. IEEE, 2009.

[23] Mathieu Salzmann, Richard Hartley, and Pascal Fua. Convex optimization for deformable surface 3-d tracking. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[24] Mathieu Salzmann, Raquel Urtasun, and Pascal Fua. Local deformation models for monocular 3d shape recovery. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[25] Olga Sorkine. Laplacian mesh processing. *Eurographics (State of the Art Reports)*, 4(4), 2005.

[26] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 2020.

[27] Lokender Tiwari and Brojeshwar Bhowmick. Deepdraper: Fast and accurate 3d garment draping over a 3d human body. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1416–1426, 2021.

[28] Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 201–221. Springer, 2022.

[29] Aydin Varol, Mathieu Salzmann, Pascal Fua, and Raquel Urtasun. A constrained latent variable model. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2248–2255. Ieee, 2012.

[30] Jingbo Zhang, Ziyu Wan, and Jing Liao. Adaptive joint optimization for 3d reconstruction with differentiable rendering. *IEEE Transactions on Visualization and Computer Graphics*, 2022.