

Learning High-Fidelity Garment Deformation via Skinning-Free Image Transfer

Supplementary Material

A. Dataset

The licenses for VTO [7], TailorNet [6] and CLOTH3D [1] datasets can be found in the below urls:

VTO: <https://github.com/isantesteban/vto-dataset>.

TailorNet: https://github.com/zycliao/TailorNet_dataset.

CLOTH3D: <https://chalearnlap.cvc.uab.cat/dataset/38/description/>.

B. Image Rendering

We illustrate the rendering process as shown in Figure 1. For a deformed mesh, we convert its vertex positions or normals to RGB colors, while projecting the garment template mesh with its vertices to render corresponding images in both front and back views. Note that the image silhouette remains the same for all deformed meshes as we always project the template vertices. From the rendered images, we observe that the position images mostly contain low-frequency information, *e.g.* areas of colors representing the posed garment shape, while the normal images mostly contain high-frequency details such as edges for wrinkles. Motivated this observation, we propose to leverage both modalities to model garment deformation, which facilitates to generate accurately deformed garments.

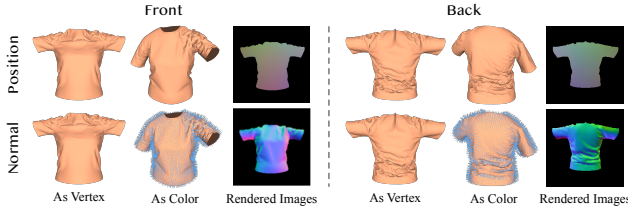


Figure 1. **Illustration of image rendering.** We project template mesh with its *vertices* and convert geometric attributes into *colors* to render images from both front and back views.

C. Network Architecture

We use the DINO [2] encoder model `dino-vitb16` pre-trained on the ImageNet [3]. During training, we fine-tune the last two layers 10 and 11, as well as the final layernorm module. The encoded image features $\bar{\mathbf{F}}_g^s, \mathbf{F}_b^s$ are in the shape $\mathbb{R}^{257 \times 768}$, where we include the [CLS] token to distinguish garment and body inputs. In each transformer block in the pose-conditioned feature refinement module, we refine the input garment feature $\mathbf{F}_g^{(0)}$, ignoring

the superscript of view s for simplicity, as:

$$\begin{aligned}\mathbf{F}_g^{(1)} &= \mathbf{F}_g^{(0)} + \text{Cross}(\text{Nm}(\mathbf{F}_g^{(0)}), \mathbf{F}_b^s, \mathbf{F}_b^s) \\ \mathbf{F}_g^{(2)} &= \mathbf{F}_g^{(1)} + \text{Self}(\text{Nm}(\mathbf{F}_g^{(1)}), \text{Nm}(\mathbf{F}_g^{(1)}), \text{Nm}(\mathbf{F}_g^{(1)})) \\ \mathbf{F}_g^{(3)} &= \mathbf{F}_g^{(2)} + \text{MLP}(\text{Nm}(\mathbf{F}_g^{(2)})) ,\end{aligned}\quad (1)$$

where $\mathbf{F}_g^{(1)}, \mathbf{F}_g^{(2)}$, and $\mathbf{F}_g^{(3)}$ are intermediate output features, $\text{Cross}(\cdot)$ and $\text{Self}(\cdot)$ denote 4-heads of cross and self attention respectively, and $\text{Nm}(\cdot)$ denotes the layer norm. In the MLP, we use linear layers of [3072, 768] and GELU activation. Finally, we use $L = 4$ residual blocks of 2D convolutions to construct the image decoder. In each block, we decode the input feature $\hat{\mathbf{F}}_g^{(0)}$ as:

$$\begin{aligned}\hat{\mathbf{F}}_g^{(1)} &= \text{Conv}_1(\hat{\mathbf{F}}_g^{(0)}) + \text{Conv}_2(\text{NL}(\text{Conv}_3(\text{NL}(\hat{\mathbf{F}}_g^{(0)})))) \\ \hat{\mathbf{F}}_g^{(2)} &= \text{TransConv}(\hat{\mathbf{F}}_g^{(1)}) ,\end{aligned}\quad (2)$$

where $\text{Conv}_i(\cdot)$ denotes the 2D convolution layer. For $i = \{1, 3\}$, the convolution halves the feature dimension, while for $i = 2$, the output dimension remains the same as the input. $\text{NL}(\cdot)$ denotes the non-linear Swish activation function, and $\text{TransConv}(\cdot)$ denotes the transposed 2D convolution that doubles the spatial resolution of the features.

D. Inference Time Comparison

We compare inference time on the "T-shirt" and "Dress" garments in the VTO dataset, which contain 4K and 12K triangles respectively. While we use test-time optimization for multimodal fusion, our method is significantly faster than simulation-based method [5] and comparable with physics-based methods [4], thus maintaining its practical applicability. Since we directly obtain the posed garment shape from pixel values of transferred position images as initialization, which is close to the optimal results and helps to improve convergence speed. In addition, the normal optimization is relatively simple and well-conditioned. These two designs ensure the efficiency of the fusion process and allows the optimization to converge in only 100 steps.

Table 1. **Inference time comparison.** Our method is more efficient than simulation [5] and physics [4] based methods.

Time (s)	[5]	[4]	Ours	[6]	[8]	[7]
T-shirt	3.891	0.127	0.115	0.028	0.003	0.005
Dress	5.680	0.167	0.153	0.040	0.004	0.008

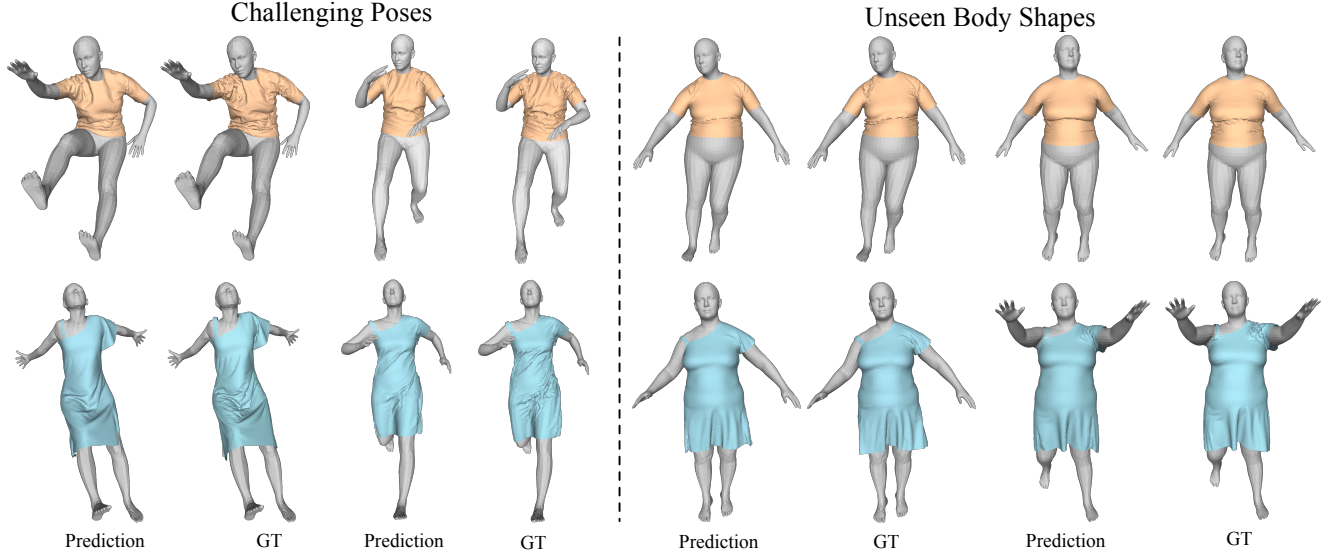


Figure 2. **Generalization to challenging poses and unseen body shapes.** Our method can consistently produce accurate and detailed wrinkles on challenging poses and unseen body shapes.

E. More Qualitative Results

In this section, we show more results with challenging poses and body shapes in Figure 2. Since high frequency wrinkle details are effectively decoupled, our method can consistently produce high-fidelity results on challenging poses. Furthermore, we show that our method can generalize to unseen body shapes when jointly trained with multiple shapes.

G. More Quantitative Comparison

In the main paper, we qualitatively show generalization results on CLOTH3D [1] dataset. We further include quantitative comparison against [6] and [7] in Table 2. Specifically, we jointly train on 50 dress garments and reserve 10 garments for testing. For both [6] and [7], we use their official code to re-train on this dataset (we use multi-garment version for [6]). Results show that our method noticeably improves the deformation accuracy and consistently generalizes well on unseen garment, demonstrating the efficacy of the proposed image-based skinning-free approach.

Table 2. **More Results on CLOTH3D.**

Methods	RMSE↓	Hausdorff↓	STED ↓
Santesteban <i>et al.</i> [7]	29.01	94.30	0.0863
TailorNet [6]	25.33	86.14	0.0751
Ours	19.54	73.22	0.0532

F. Failure Cases

Failure Case. To trade for model efficiency, we only use front and back views to render the images, with the assumption that common garment templates are flat under these two

views and are reasonably thin. For non-visible side views, the vertex positions are constrained by the edge length and normal consistency losses \mathcal{L}'_e and \mathcal{L}_{rn} , respectively. However, due to the lack of direct supervision on these vertices, their deformation may not align with the ground truth data and contain undesired artifacts, as shown in Figure 3. Nevertheless, we show the results are still more accurate than baseline work [8], even though it works on 3D networks with access to deformation of all vertices. We encourage future works to explore more robust approaches to estimate side view deformations, in particular leverage learning-based methods.

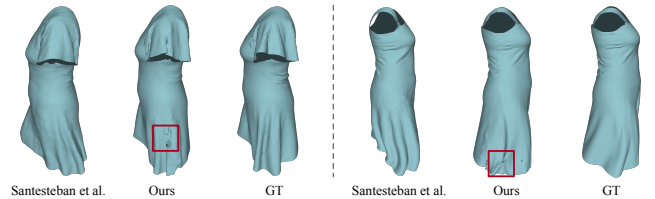


Figure 3. **Illustration of Failure Cases.** Failure to constrain non-visible side view vertices can produce incorrect deformations with undesired artifacts.

References

- [1] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: clothed 3d humans. In *European Conference on Computer Vision*, pages 344–359. Springer, 2020. 1, 2
- [2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Pro-*

ceedings of the IEEE/CVF international conference on computer vision, pages 9650–9660, 2021. [1](#)

- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. [1](#)
- [4] Artur Grigorev, Michael J Black, and Otmar Hilliges. Hood: Hierarchical graphs for generalized modelling of clothing dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2023. [1](#)
- [5] Rahul Narain, Armin Samii, Tobias Pfaff, and J O’Brien. Arcsim: Adaptive refining and coarsening simulator. *University of California–Berkley, Berkley, CA, accessed Oct, 1:2016*, 2014. [1](#)
- [6] 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. [1](#), [2](#)
- [7] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum*, pages 355–366. Wiley Online Library, 2019. [1](#), [2](#)
- [8] Igor Santesteban, Nils Thuerey, Miguel A Otaduy, and Dan Casas. Self-supervised collision handling via generative 3d garment models for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11763–11773, 2021. [1](#), [2](#)