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

Rong Wang Wei Mao Changsheng Lu Hongdong Li
The Australian National University
{rong.wang, changsheng.lu,hongdong.li}@anu.edu.au wei.mao.research@gmail.com

Abstract

We present a novel method for generating 3D garment deformations from underlying body poses, which is key to a wide range of applications, including virtual try-on and extended reality. To simplify the cloth dynamics, existing methods mostly rely on linear blend skinning to obtain low-frequency posed garment shape and only regress high-frequency wrinkles. However, due to the lack of explicit skinning supervision, such skinning-based approach often produces misaligned shapes when posing the garment, consequently corrupts the high-frequency signals and fails to recover high-fidelity wrinkles. To tackle this issue, we propose a ***skinning-free*** *approach by independently estimating posed (i)* ***vertex position*** *for low-frequency posed garment shape, and (ii)* ***vertex normal*** *for high-frequency local wrinkle details. In this way, each frequency modality can be effectively decoupled and directly supervised by the geometry of the deformed garment. To further improve the visual quality of deformation, we propose to encode both vertex attributes as* ***rendered*** *texture images, so that 3D garment deformation can be equivalently achieved via 2D image transfer. This enables us to leverage powerful pre-trained image models to recover fine-grained visual details in wrinkles, while maintaining superior scalability for garments of diverse topologies without relying on manual UV partition. Finally, we propose a multimodal fusion to incorporate constraints from both frequency modalities and robustly recover deformed 3D garments from transferred images. Extensive experiments show that our method significantly improves animation quality on various garment types and recovers finer wrinkles than state-of-the-art methods.*

1. Introduction

Generating high-quality 3D garment deformation given posed human bodies facilitates a wide range of applications, such as digital humans, virtual try-on, and extended reality. Traditional works [3, 33] mostly rely on simulation to generate plausible results. However, such physics-based methods are time-consuming and require manual fine-tuning simulator settings for each garment, which is laborious and does not scale to diverse garment types [25, 46].

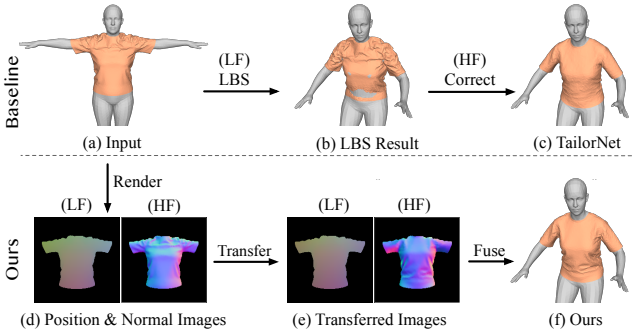

Figure 1. **Illustration of our method.** Given input garment and body meshes (a), previous work [31] relies on LBS to generate low-frequency (LF) posed garment shape. However, inaccurate skinning in LBS can produce artifacts and misaligned garment position (b), which corrupts high-frequency (HF) signals and hinders the wrinkle regression (c). In contrast, we decompose frequency modalities using two geometric attributes: vertex positions and normals, which are rendered as 2D texture images (d) and then transferred on pixel intensities (e) to represent garment deformation. After fusing from both modalities, we generate deformed garment with more accurate wrinkles (f).

Recently, learning-based methods [26, 30, 31, 35, 37, 48, 50] have received increasing attention thanks to the efficiency and scalability of deep networks. As garment deformation consists of both high-frequency wrinkles and low-frequency posed garment shape, it is challenging for neural networks to directly regress a single deformation field due to its spectral bias [34, 49]. To this end, previous works mostly adopt a simplified two-stage decomposition by firstly regressing wrinkles relative to the un-posed garment template, then using linear blend skinning (LBS) to obtain the low-frequency posed garment shape. However, due to the lack of explicit supervision on garment skinning, they either assume tight garments and directly skin from the closest body vertex [31, 35–37], or skin loose garments with virtual joints [30, 50]. Such unsupervised skinning can produce misaligned garment positions and undesired artifacts, which corrupts high-frequency signals. Con-

sequently, these skinning-based methods can fail to recover high-fidelity wrinkles details, as illustrated in Figure 1.

To tackle this issue, we present a novel *skinning-free* method that decomposes high-low frequency modalities instead with two geometric attributes. Specifically, we propose to directly estimate posed garment **vertex positions** instead of relying on garment skinning. As networks tend to prioritize learning low-frequency signals [34], we observe the model learns well with the overall posed shape but can generate over-smoothed garment geometry. To recover fine-grained wrinkles, we further estimate **vertex normals** that better capture local surface bending arising in wrinkles. In contrast to skinning-based methods, our method effectively decouples frequency components and enables *explicit supervision* for both modalities, which avoids noisy skinning and produces high-quality wrinkles.

Motivated by the capability of visual perception in large vision models [7, 10, 29], we propose to project 3D garments onto 2D image space aiming at improving perceptual quality of deformation results. Specifically, unlike previous works that mostly rely on manual UV partition for semantically-meaningful image representation [16, 47], we instead render both vertex attributes as multi-view texture images encoded in the garment *canonical shape*, which efficiently scales to large collection of garments with diverse topologies and minimizes garment-body occlusion during deformation. In this way, we effectively model 3D garment deformation as a 2D image transfer task, as illustrated in Figure 1. Finally, we fuse priors from transferred images of both frequency modalities to recover deformed 3D garments, which robustly generates plausible deformation for invisible areas introduced by rendering projection.

Our contributions can be summarized as follows. (*i*) We propose a novel skinning-free pipeline for garment deformation with effective high-low frequency modalities decomposition, which avoids noisy garment skinning and facilitates fine-grained wrinkle regression. (*ii*) We model 3D garment deformation via 2D image transfer, leveraging powerful vision models and scalable image representation to recover high-fidelity visual details for wrinkles. Extensive experiments show that our method noticeably improves deformation and perceptual quality over state-of-the-art learning-based methods.

2. Related Works

Physics-based Methods. To generate physically plausible garment animation, physics-based methods either rely on time-consuming simulators [6, 20, 33, 45], or optimize through physics-inspired losses [5, 11, 37]. To ensure realism and accuracy, simulator parameters need to be fine-tuned for each garment instance, which can be laborious. Several works propose to estimate these parameters through differentiable simulation [18, 21] or neural networks [9, 44], however, the estimation needs to be performed in a controlled setting with known external factors, which limits their applications. The challenge in data preprocessing thus restricts such method from scaling to diverse garment types.

Learning-based Methods. In contrast, learning-based methods [8, 26, 30, 31, 35, 41, 48, 50] have been developed to achieve superior efficiency and scalability. Pioneered by [19], most works follow to estimate pose space deformation (PSD), namely they adopt LBS to obtain low-frequency posed garment shape, while predicting high-frequency wrinkles in the canonical garment space. Specifically, [35] directly regresses local vertex displacements using recurrent neural networks. [31] proposes to first use mixture models to construct bases of high-frequency deformations, then combine them with narrowed bandwidth kernels. [48] leverages generative models to encode the feasible high-frequency latent space. Similar to our approach, [16, 47] uses normal maps to model fine wrinkles. However, they require manually built UV maps and rely on LBS to generate initial normals, which we show in the ablation study that are sub-optimal. While the above works tackle tight garments and directly access body skinning weights, [30, 50] further extend to loose garments by predicting virtual garment joints to which garments are skinned. However, the prediction of virtual joints can not be explicitly supervised, which can lead to incorrect joint transformations. In summary, existing learning-based methods mostly suffer from noisy skinning that can not be directly supervised. Consequently, the skinning artifacts need to be jointly refined during wrinkle regression, which prevents them from estimating accurate wrinkles. In contrast, we present a *skinning-free* approach, which effectively avoids noisy skinning and facilitates to generate more accurate wrinkles.

Image-based 3D Representation. In view of large-scale image datasets and effective pretrained image models, recent works propose to represent 3D geometry in the image space. Most existing works leverage UV mapping as the image representation, which have been widely applied in human pose estimation [12], avatar generation [22, 27], and scan registration [13]. However, they mostly rely on manual UV unwrapping to produce semantically meaningful islands, which requires expert knowledge and is laborious, thus does not scale to large-scale collections. Alternatively, [22, 23, 42, 43] propose to render multi-view images to automatically establish vertex-to-pixel correspondence. Specifically, [42, 43] learn to generate 3D clothed humans by integrating from estimated normal images. [23] encodes 3D character animation via ultra dense pose images. [22] designs Gaussian maps rendered from template human meshes to encode parameters for Gaussian splatting. Unlike all above works that consider only a single image source, we observe that accurate garment deformation requires effectively fusing *multiple* image domains with

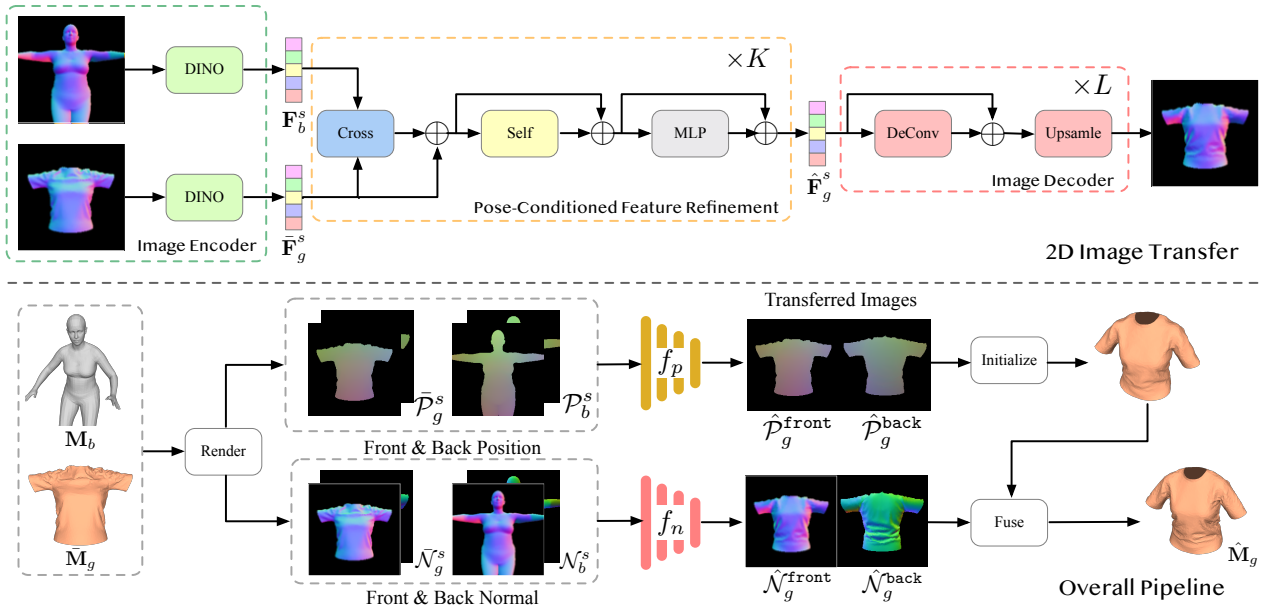

Figure 2. **Overview of our method.** Given the input garment template $\bar{\mathbf{M}}_g$ and posed body mesh $\mathbf{M}_b$, we first render position and normal images for the garment $\{\bar{\mathcal{P}}_g^s, \bar{\mathcal{N}}_g^s\}$ and body $\{\mathcal{P}_b^s, \mathcal{N}_b^s\}$ from each view s, aiming to project the 3D garment onto the image space. Next, we transfer position images in $f_p(\cdot)$ and normal images in $f_n(\cdot)$, where the two networks have the same architecture as shown in the top row (taking front normal images as an example). Finally, we initialize the posed garment mesh from transferred position images $\hat{\mathcal{P}}_g^s$ and recover missing wrinkle details by fusing from normal images $\hat{\mathcal{N}}_g^s$ to obtain the deformed garment $\hat{\mathbf{M}}_g$. "$\oplus$" denotes residual connection.

mixed frequency modalities, which is achieved via a novel pipeline as will be introduced below.

3. Method

To avoid skinning artifacts and produce more accurate wrinkles, we present a novel skinning-free pipeline, as shown in Figure 2. Given a garment template mesh $\bar{\mathbf{M}}_g = \{\bar{\mathbf{V}}_g, F_g\}$ where $\bar{\mathbf{V}}_g \in \mathbb{R}^{N\times 3}$ denotes vertex positions and $F_g \in \mathbb{Z}_+^{F\times 3}$ denotes triangle faces, we aim to estimate its deformed mesh $\hat{\mathbf{M}}_g = \{\hat{\mathbf{V}}_g, F_g\}$ conditioned on the posed body mesh $\mathbf{M}_b = \{\mathbf{V}_b, F_b\}$. Instead of directly regressing 3D vertex displacements, we propose to model garment deformation in 2D image space. Specifically, we first render RGB images $\bar{\mathcal{P}}_g^s$ and $\bar{\mathcal{N}}_g^s \in \mathbb{R}^{H\times W\times 3}$ that respectively encode vertex positions and normals of the garment template from each view $s \in \{\text{front}, \text{back}\}$. Similarly for the posed body mesh we render corresponding images $\{\mathcal{P}_b^s, \mathcal{N}_b^s\}$. We then develop an image transfer network to generate transferred images $\hat{\mathcal{P}}_g^s$ and $\hat{\mathcal{N}}_g^s$, which describe posed garment shape and wrinkles respectively (Section 3.2). Finally, we propose a multimodal fusion process to leverage priors of both modalities and optimize 3D deformed garment mesh from transferred images (Section 3.3).

3.1. Image Rendering

We represent 3D garment and body meshes as 2D images rendered from *multiple views*, where the intensities of pixels encode the garment geometry, *e.g.* positions or normals of vertices. Specifically, to generate such an image for a deformed mesh, we use its vertex positions and normals to color the corresponding vertices on the *template mesh*, then render the results from both front and back views. In this way, for different deformations of the same mesh, the rendered pixel values will be different while the image silhouette *remains the same*, as we always project the template mesh onto the image space. Taking the garment as an example, given vertex positions $\mathbf{V}_g$ for a deformed mesh, we compute the corresponding vertex normals $\mathbf{N}_g \in \mathbb{R}^{N\times 3}$ and linearly rescale their values to fit RGB colors, *i.e.* within the range $[0, 1]$. We then render the images from each view with a perspective camera of known transformation matrix as:

$$\mathcal{P}_g^s = f_r^s(\text{RGB}(\mathbf{V}_g); \bar{\mathbf{M}}_g) \quad (1)$$

$$\mathcal{N}_g^s = f_r^s(\text{RGB}(\mathbf{N}_g); \bar{\mathbf{M}}_g)\ , \quad (2)$$

where $\text{RGB}(\cdot)$ represents the linear rescaling function that maps positions or normals to RGB values, and $f_r^s(\cdot; \bar{\mathbf{M}}_g)$ represents the renderer function from view s with the constant template $\bar{\mathbf{M}}_g$. Similarly, we can obtain images for the posed body as $\mathcal{P}_b^s$ and $\mathcal{N}_b^s$, where we use the same cameras as for the garment to capture only relevant body areas.

Discussion. Our rendering configuration has two advantages compared to alternatives: 1) We can automatically establish vertex-to-pixel correspondence through perspective projection, thus do not require manually built UV parameterization [16] or sew patterns [21, 32]. The rendered images also retain the canonical garment shape, which facili-

tates the image feature extraction; 2) Instead of directly projecting the deformed mesh, we project the template mesh throughout rendering, *which will not be influenced by self-occlusions during garment deformation.* Moreover, we use front and back views to efficiently capture most visible garment vertices, which also provide sufficient constraints to infer non-visible vertices at side views. After rendering all images, we can estimate 3D garment deformation via 2D image transfer as illustrated next section.

3.2. 2D Image Transfer

As the garment geometry can be fully represented by the position and normal images, we formulate 3D garment deformation as an image transfer task, *i.e.* we aim to transfer from the initial images $\{\bar{\mathcal{P}}_g^s, \bar{\mathcal{N}}_g^s\}$ representing the garment template to the target images $\{\mathcal{P}_g^s, \mathcal{N}_g^s\}$ representing the deformed garment, conditioned on the posed body images $\{\mathcal{P}_b^s, \mathcal{N}_b^s\}$ as:

$$\hat{\mathcal{P}}_g^s = f_p(\bar{\mathcal{P}}_g^s, \mathcal{P}_b^s), \qquad \hat{\mathcal{N}}_g^s = f_n(\bar{\mathcal{N}}_g^s, \mathcal{N}_b^s) \ , \tag{3}$$

where $\hat{\mathcal{P}}_g^s$ and $\hat{\mathcal{N}}_g^s$ are the estimated posed position and normal images, respectively. The position transfer network $f_p(\cdot)$ and normal transfer network $f_n(\cdot)$ have the same architecture (as in Figure 2). Each network contains three consecutive modules: (*i*) an image feature encoder that extracts visual features of garment and body geometry, (*ii*) a pose-conditioned feature refinement module that injects pose condition and models fine-grained body-garment interaction, and (*iii*) an image decoder that decodes the transferred images. We will introduce each module in below.

Image Feature Encoder. We forward each image input to a pre-trained vision transformer DINO [7] to encode patch-wise tokens of image features $\bar{\mathbf{F}}_g^s, \mathbf{F}_b^s \in \mathbb{R}^{M \times D}$ for garment and body respectively, where M represents the number of tokens and D represents the feature dimension. Compared with other encoders like ImageNet-pretrained ResNet [14], DINO can effectively encode detailed structural and visual information through attention on salient image contents, which is beneficial for generating fine wrinkles. We further show its efficacy in Section 4.3.

Pose-Conditioned Feature Refinement. We refine image features to introduce pose priors in K transformer blocks. In each block, motivated by [40], we model body-garment interaction by first computing the multi-head cross-attention [39] between the garment feature $\bar{\mathbf{F}}_g^s$ and the body feature $\mathbf{F}_b^s$ to generate the pose-conditioned garment feature. In contrast to the traditional skinning process that computes the skinning weights with respect to sparse joints, we learn to model the *dense* correlation between image patches, which can capture fine-grained body-garment interaction. Furthermore, we follow the vanilla transformer structure [39] and continue to forward the feature into a self-attention layer followed by a multi-layer perceptron (MLP) to generate the refined garment feature $\hat{\mathbf{F}}_g^s$.

Image Decoder. Finally, $\hat{\mathbf{F}}_g^s$ is rearranged spatially to form a 3D tensor corresponding to the 2D image feature map and forwarded to an image decoder to generate transferred images. The image decoder consists of residual blocks of 2D convolution layers, followed by transposed convolution layers to upsample the spatial resolution.

Training Objectives. We train each network using the masked L1 loss as:

$$\mathcal{L}_p = \textstyle\sum_s ||\bar{\mathcal{S}}_g^s \odot \hat{\mathcal{P}}_g^s - \bar{\mathcal{S}}_g^s \odot \mathcal{P}_g^s||_1 \tag{4}$$

$$\mathcal{L}_n = \textstyle\sum_s ||\bar{\mathcal{S}}_g^s \odot \hat{\mathcal{N}}_g^s - \bar{\mathcal{S}}_g^s \odot \mathcal{N}_g^s||_1 \ , \tag{5}$$

where $\mathcal{P}_g^s$ and $\mathcal{N}_g^s$ represent ground truth position and normal images, $\bar{\mathcal{S}}_g^s$ represents the silhouette of the garment template to mask for valid pixels, and $\odot$ represents pixel-wise multiplication. Note that we independently model each modality regardless of their consistency constraints, as we observe that mixing frequency components during image transfer leads to inferior accuracy (as compared in Section 4.3). Alternatively, we opt to fuse position-normal correlation via explicitly optimization.

3.3. 3D Multimodal Fusion

While we can obtain vertex positions of the deformed garment solely from the position images $\hat{\mathcal{P}}_g^s$ of both views, we observe two major issues of such an approach: (*i*) since the high frequency wrinkle details are often reflected by relatively small position changes, it is hard for the position transfer network $f_p(\cdot)$ to capture such subtleties, thus leading to an over-smoothed mesh, (*ii*) although images from front and back views cover most of the garment, the positions of non-visible vertices at side views can not be directly obtained from these images. In this section, we propose a multimodal fusion process to address both issues. The key idea is to incorporate high-frequency wrinkle details recorded by the normal images to refine the over-smoothed mesh initialized from the position images, while using the edge and surface priors to constrain the non-visible vertices. Specifically, we aim to optimize the deformed vertex positions $\mathbf{V}_g^\star$ that aligns with both image observations as:

$$\mathbf{V}_g^\star = \arg\min_{\hat{\mathbf{V}}_g} \textstyle\sum_s (||f_r^s(\hat{\mathbf{V}}_g) - \hat{\mathcal{P}}_g^s|| + ||f_r^s(\hat{\mathbf{N}}_g) - \hat{\mathcal{N}}_g^s||) \ , \tag{6}$$

where $f_r^s(\cdot)$ is the renderer function in Eq.(2) that omits constants and $\hat{\mathbf{N}}_g$ are vertex normals computed from vertex positions $\hat{\mathbf{V}}_g$. The optimization includes two stages where garment mesh is firstly initialized from position images and then refined with normal images to recover fine wrinkles.

Vertex Position Initialization. We initialize vertex positions from position images based on their visibility under

the perspective projection. For visible vertices, we simply interpolate their corresponding pixel values in the transferred position images to initalize their positions. For non-visible vertices especially at side views, we initialize them by linearly interpolating from the closest front and back visible vertex pairs. To correct the linear interpolation and ensure a smooth boundary between two types of vertices, we smooth the results by minimizing the edge length loss $\mathcal{L}_e$ as:

$$\mathcal{L}_e = \frac{1}{|\mathcal{E}|}\sum\nolimits_{\{i,j\}\in\mathcal{E}}(\|\hat{\mathbf{V}}_g[i]-\hat{\mathbf{V}}_g[j]\|-\|\bar{\mathbf{V}}_g[i]-\bar{\mathbf{V}}_g[j]\|)^2 \ , \quad (7)$$

where $\mathcal{E}$ represents the index set of all edges defined by the garment faces F_g, $\hat{\mathbf{V}}_g$ and $\bar{\mathbf{V}}_g$ represent estimated and template mesh vertices, respectively. Moreover, we impose a regularization loss $\mathcal{L}_{rv}$ to penalize the L_2 distance on displacements of visible vertices, in order to align with the position images. The overall loss for this stage can be summarized as $\mathcal{L}_e + \lambda_{rv}\mathcal{L}_{rv}$, with loss weight λ_{rv}.

Vertex Normal Fusion. To amend high-frequency wrinkle details on the position-initialized vertices, we fuse normal predictions onto them by minimizing the normal rendering loss $\mathcal{L}_r$ defined as the second term in Eq.(6), and then smooth the results using a normal consistency loss $\mathcal{L}_{rn}$ as:

$$\mathcal{L}_{rn} = \frac{1}{|\mathcal{E}|}\sum\nolimits_{\{i,j\}\in\mathcal{E}}(1 - \hat{\mathbf{N}}_g[i]^T\hat{\mathbf{N}}_g[j]) \ . \quad (8)$$

Similar to the initialization stage, we impose the edge length loss $\mathcal{L}_e$ to penalize irregular rim contours and include the vertex displacement regularization $\mathcal{L}_{rv}$ on all vertices. Finally, to penalize garment-body collision, we impose a collision loss by penalize the penetration distance as:

$$\mathcal{L}_c = \frac{1}{N}\sum\nolimits_i \max(0, -\text{SDF}(\hat{\mathbf{V}}_g[i], \mathbf{M}_b)) \ , \quad (9)$$

where SDF$(\cdot)$ represents the vertex-to-mesh signed distance. The overall optimization objectives for normal fusion can be summarized as:

$$\mathcal{L} = \mathcal{L}_r + \lambda_{rn}\mathcal{L}_{rn} + \lambda_e\mathcal{L}_e + \lambda_{rv}\mathcal{L}_{rv} + \lambda_c\mathcal{L}_c \ , \quad (10)$$

where $\lambda_{rn}, \lambda_e, \lambda_{rv}, \lambda_c$ balance the weights of losses.

4. Experiments

4.1. Experiment Setup

Datasets: We evaluate our method and baselines on three benchmarks: 1) **VTO** dataset [35]containing two types of garments "t-shirt" and "dress" Each garment is draped onto an SMPL [24] human body using ground truth deformations simulated in the ARCSim [28] simulator. We follow [30] to use 4 clips (`01_01`, `111_02`, `55_27` and `91_36`) of medium body shape ($\boldsymbol{\beta} = 0$) and unseen poses for testing and the remaining 49 clips for training. 2) **TailorNet** dataset [31]. Since VTO only contains upper garments, we further adopt TailorNet to test on lower garments "pants" and "skirt". We follow [30] to use the medium body shape and garment style ($\boldsymbol{\beta} = 0$, $\boldsymbol{\gamma} = 0$) split, and adopt 2 clips (`005`, `010`) of unseen poses for testing and the remaining 16 clips for training. 3) **CLOTH3D** dataset [4]. For the first two datasets, we test on a single body and garment shape *only to ensure a fair comparison* with the commonly used evaluation standard in [30, 50]. However, our method is not limited to this set up. To verify this, we further demonstrates results on unseen dress garments from CLOTH3D for generalization evaluation (Section 4.4).

Metrics. Following [30, 50], we evaluate all methods on three metrics: Root Mean Squared Error (RMSE), Hausdorff distance [2], and spatio-temporal edge difference (STED) [38] between predicted and ground truth meshes. Specifically, RMSE and Hausdorff distance assess the prediction accuracy in, while STED evaluates the perceptual quality of deformation by measuring relative edge differences in each test clip.

Implementation Details. We render all images at 256×256 pixels using differentiable renderer from Nvdiffrast [17] and normalize deformed garment and posed body vertices with the global rotation and translation from the human pose. For the image transfer network, we fine-tune the last two layers of the DINO encoder, along with $K = 4$ transformer blocks for the feature refinement module. We train the model using the Adam [15] optimizer for 100K iterations, and set the learning rate to 1×10^{-4}. For multimodal fusion, we use the same optimizer with a learning rate of 1×10^{-3} and optimize for 100 steps in both the initialization and normal fusion stages. For the loss weights, we set $\lambda_{rv} = 0.02$, $\lambda_e = \lambda_c = 100$, and $\lambda_{rn} = 0.001$ on t-shirt and 0.01 on other garments based on their scales. We include the detailed model architecture and inference time comparison in supplement materials.

4.2. Results on Garment Deformation

Following [30, 50], we report all metrics by training and testing on each garment instance to ensure a fair comparison. In Table 1, we present the results on the VTO dataset, where metrics for baselines [30, 31, 35, 50] are borrowed from [30, 50] where we followed the same test configuration. For [36], we use its official weights and evaluate the results on our test split. We observe that our method achieves the best accuracy against all skinning-based methods [30, 31, 35, 50] thanks to the proposed skinning-free pipeline that avoids artifacts of noisy garment skinning. In particular, we outperform [30, 50] on loose garments, without the need to estimate additional virtual joints or anchors to facilitate garment skinning. Moreover, thanks to the capability of perceptual learning in image models [1], our method achieves improved perceptual quality (lower STED

Table 1. **Quantitative comparison on the VTO Dataset.** Best results are highlighted in **bold**. Our method achieves superior deformation accuracy and perceptual quality compared to state-of-the-art skinning based methods [30, 31, 35, 36, 50] on both tight and loose garments.

Methods	Dress			T-shirt		
	RMSE ↓	Hausdorff ↓	STED ↓	RMSE ↓	Hausdorff ↓	STED ↓
TailorNet [31]	22.95	76.80	0.0757	9.90	27.02	0.0418
Santesteban [35]	20.96	87.01	0.0745	10.25	29.56	0.0449
Santesteban [36]	21.07	87.98	0.0620	9.97	25.64	0.0335
VirtualBones [30]	19.91	83.39	0.0722	10.52	31.51	0.0452
AnchorDEF [50]	16.05	74.20	0.0493	6.25	26.31	0.0262
Ours	**13.40**	**61.73**	**0.0407**	**4.66**	**20.89**	**0.0205**

Table 2. **Quantitative comparison on the TailorNet Dataset.** Best results are highlighted in **bold** and and inapplicable results are marked with "-". Our method consistently generates more accurate results than baseline methods on lower garments.

Methods	Pants			Skirt		
	RMSE ↓	Hausdorff ↓	STED ↓	RMSE ↓	Hausdorff ↓	STED ↓
TailorNet [31]	4.84	14.46	0.0127	7.76	16.28	0.0162
Santesteban [35]	4.91	14.87	0.0129	-	-	-
VirtualBones [30]	4.76	18.75	0.0166	-	-	-
Ours	**4.03**	**13.55**	**0.0114**	**5.38**	**14.06**	**0.0150**

values) on deformed garments.

We further report quantitative comparison results in Table 2. For pants, we borrow the metrics results reported by [30] which has the same set-up with us. Since [30, 35] do not release the training code or pretrained weights for the skirt split, we are unable to evaluate their results and thus only compare with [31] and retrain with their official code on the single garment set-up. We observe that our method consistently outperforms the skinning-based method [31] when applied to lower garments.

Qualitative results. As shown in in Figures 3 and 4, our method can generate 3D deformed garments with finer wrinkle details and more accurate fold patterns. In comparison, the skinning-based baselines [31, 36] can not recover accurate wrinkles in challenging cases due to the corrupted high-frequency signals, thus generating over-smoothed geometries. For completeness, we further include a comparison with physics-based methods [11], which requires fine-tuning physical parameters for each garment to ensure accurate deformation. However, the fine-tuning process is non-trivial with only inputs of garment geometry, thus are not directly comparable to our method. We show that directly using their default physical parameters can lead to unrealistic behaviors, *e.g.* over-stretched skirts, as these physical parameters often have a global effect on the overall deformation. In contrast, learning-based methods like us ensures a more robust results by only estimating local vertex displacements. Moreover, we show in Figure 5 that accurate estimation produces results with reduced penetration, which align well with the underlying body motions. Finally, we show in the supplementary video that accurate results also lead to temporal consistency, although we use only a single pose as condition. We attribute this to the continuous nature of the neural networks on continuous inputs, which is similarly observed in previous works [31] with the same setup.

Table 3. **Effects of the skin-free approach.** We show that mixing linear blend skinning with image transfer (IT) leads to inferior performance than our full skinning-free approach.

Methods	Dress			T-shirt		
	RMSE ↓	Hausdorff ↓	STED ↓	RMSE ↓	Hausdorff ↓	STED ↓
IT + LBS	18.50	78.25	0.0625	6.85	25.30	0.0295
LBS + IT	15.75	68.40	0.0493	6.01	24.15	0.0242
Ours (Skin-free)	**13.40**	**61.73**	**0.0407**	**4.66**	**20.89**	**0.0205**

4.3. Ablation Study

Skinning-free Approach. To show the efficacy of the skinning-free approach, we compare in Table 3 with two LBS-based variants (both using skinning weights from nearest body vertices) equipped with our image transfer modules: (*i*) we supervise with GT transferred images in the canonical space, which are generated via inverse LBS (IT + LBS), and (*ii*) we refine from input images of LBS re-posed garments (LBS + IT), analog to [16, 47]. We observe that both variants produce inferior results, as inaccurate skinning can produce noisy artifacts, which corrupt high-frequency signals in either GT or input images and thus hindering learning correct wrinkle patterns.

Study on Image Transfer (IT). To verify the effects of key modules in the image transfer network, we compare several alternatives to the current designs: (*i*) replacing body-garment cross-attention with simply adding the body and garment features (w/o Body Attn.) (*ii*) replacing the DINO encoder with ResNet-50 (ResNet Encoder), and (*iii*) adding cross-attention between two modalities (w/ Corss Modal), as illustrated in Figure 6. We find both (*i*) and (*ii*) lead to smoothed garment geometry, showing the efficacy of the body-garment cross-attention for modeling fine-grained body-garment interaction, as well as the DINO encoder for extracting detailed garment structural information. Moreover, we empirically observe that mixing frequency signals like (*iii*) results in inferior accuracy, thus we choose to separately tackle each modality.

In addition, we compare with the image representation of automatically generated UV maps via xatlas in Figure 6 and Table 4. Since such images contain a large number of islands and therefore destroys garment shape priors, which is not conducive to the pretrained encoders and thus does not benefit wrinkle estimation. While manual UV can potentially solve this issue, it does not scales to a large collection of garments. Finally, we quantitatively compare in Table 4 with two more variants, which uses 3D MLP and graph convolution network (GCN) to directly regress 3D vertex positions and normals. Due to limited capacity of simple

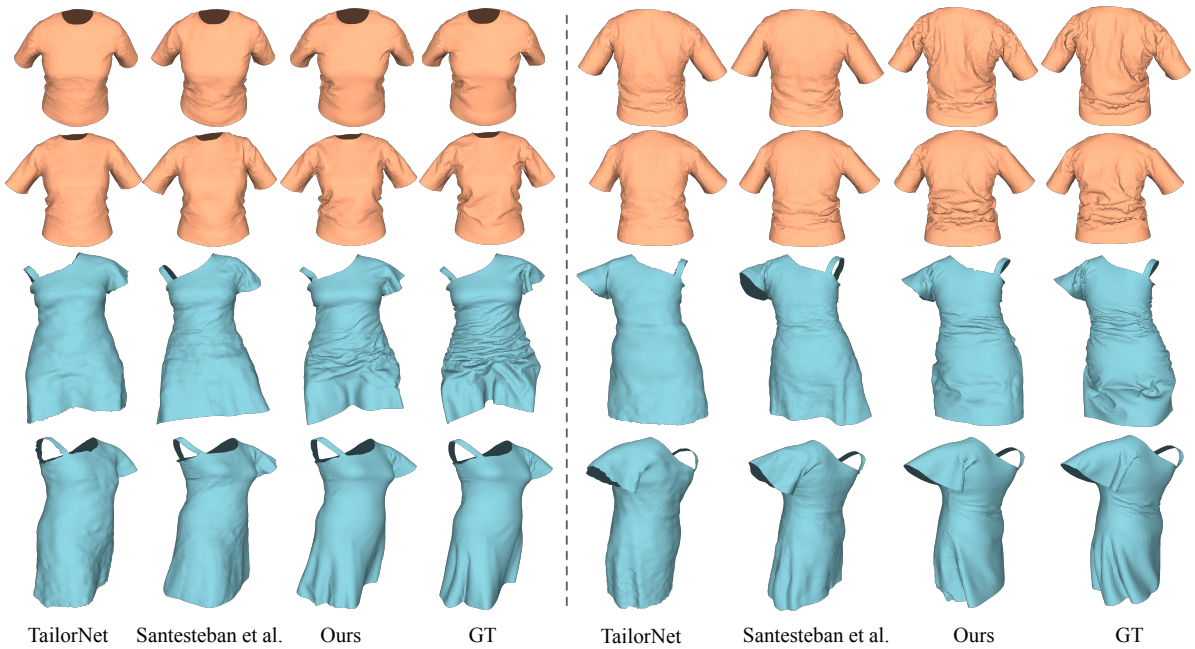

Figure 3. **Results on VTO dataset.** We produce more accurate wrinkles and folds than skinning-based methods [31, 36].

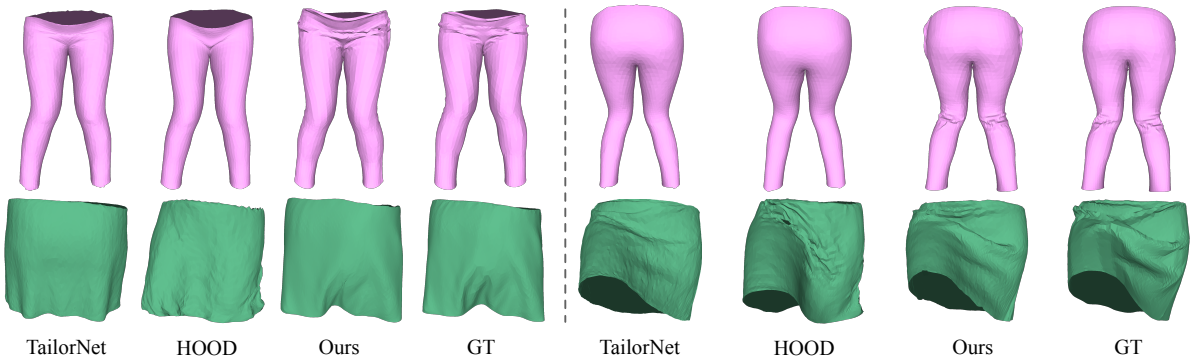

Figure 4. **Results on TailorNet dataset.** Our method consistently produces more accurate deformations on lower garments than [11, 31].

Figure 5. **Qualitative results with human motions.** Our method generates accurate and plausible garment deformations for a sequence of unseen human poses. Moreover, the deformed garments are temporally consistent. We show more results in the supplementary video.

3D networks, it produces inferior results than the proposed image transfer module.

Fusion Losses. In Table 5 and Figure 7, we show the effect of each loss during fusion optimization. The initial mesh (a) interpolated from the position images does not contain enough high-frequency wrinkle details, and non-visible vertices at side views can not be constrained, which leads to large RMSE error. By enforcing edge length consistency, we repair non-visible vertices as in (b). Jointly with (a) and (b), we obtain a reasonably good initialization that allows normal fusion to be feasibly achieved with few optimization steps. Moreover, we recover more accurate wrinkles after fusing from normal images (c). However, with only the normal loss, we observe artifacts at mesh rims due to under-constrained objectives. We thus further refine the results by penalizing irregular boundary edges and abrupt nor-

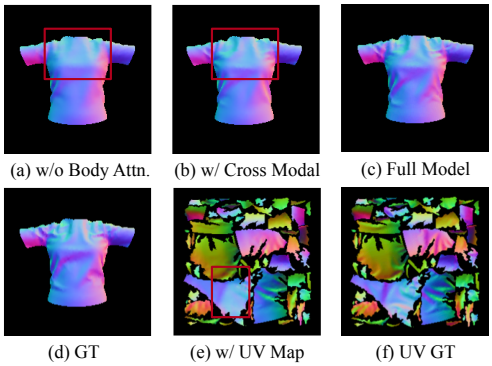

Figure 6. **Comparison of design variants for image transfer modules.** We show that design variants in the network architecture (a) and (b) produce smoother results, while the full model (c) can generate fine wrinkles that are closer to the GT (d). In addition, we show that using automatically generated UV maps (e) results in complex islands that are not beneficial for wrinkle estimation.

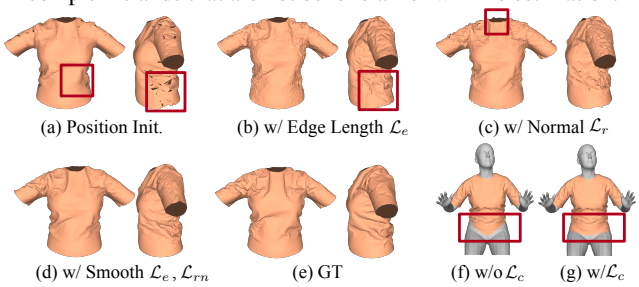

Figure 7. **Illustration of intermediate fusion results.** We show optimization results after adding corresponding losses. By fusing both modalities, we produce more accurate deformation with higher perceptual quality. Moreover, the edge length and normal consistency help to constrain non-visible vertices and resolve artifacts at mesh rims.

mal changes, which generate smoothed results as in (d). Finally, we show that the collision loss $\mathcal{L}_c$ effectively reduces body-garment collision as shown in (g).

Table 4. **Variants with IT.**

Methods	RMSE↓	Hausdorff↓	STED ↓
3D MLP	8.39	24.77	0.0310
3D GCN	7.95	24.25	0.0337
Auto-UV	6.43	23.38	0.0277
Full Model	**4.66**	**20.89**	**0.0205**

Table 5. **Study on fusion losses.**

Methods	RMSE↓	Hausdorff↓	STED ↓
Init. (no normal)	9.98	30.04	0.0771
Init. + $\mathcal{L}_e$	5.52	23.45	0.0525
Init. + $\mathcal{L}_e$ + $\mathcal{L}_r$	4.96	22.23	0.0238
Full Model	**4.66**	**20.89**	**0.0205**

4.4. Generalization Analysis

While baseline methods [30, 35, 36, 50] are often evaluated for single garment set-up, our method is not limited to this configuration and in fact is well-suited for training across multiple garments thanks to several designs: (*i*) the use of pretrained DINO encoder that is capable of extracting detailed semantic features for various garments, (*ii*) the image transfer approach that is agnostic to garment topologies, and (iii) the use of front and back view projections to establish image representations for garments, which will not be scalable for manual UV parameterization on a large collection of garments. To verify the generalizability of our method, we further jointly train on 50 dress garments on the CLOTH3D [4] dataset, and show the results in Figure 8. By training across multiple garments, our method can effectively generalize to unseen garment shapes, with the skinning-free method particularly beneficial for tackling loose garments. Furthermore, we show that our method can be extended to jointly train on multiple body shapes, which enables generalization to unseen body shapes and producing different plausible deformations.

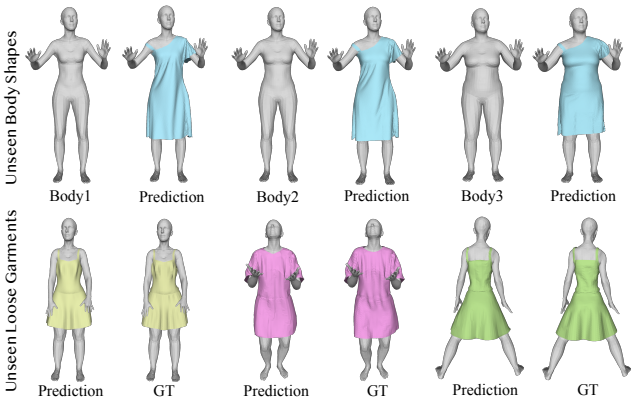

Figure 8. **Generalization results on unseen body shapes and garment types.** We show our method can be jointly trained on multiple body shapes and garment types to achieve effective generalization, demonstrating the scalability of the approach.

5. Discussion

Limitation. Although our method succeeds in generating fine-grained garment wrinkles, the geometric fidelity of the deformation are still bounded by the resolution of the rendered images, thus may lead to missing details such as deformation for tiny garment pieces and small wrinkles. In addition, we only consider pose-dependent deformations, while omitting complex dynamics such as interactions between multiple layered garments.

Conclusion. In this paper, we propose a novel skinning-free pipeline to generate high-fidelity 3D garment deformation via image transfer. We decompose garment deformation into decoupled frequency modalities represented by vertex positions and normals, and further project both modalities into the image space, which allows us to leverage powerful vision models to produce wrinkles of superior scalability and perceptual quality. Thanks to these designs, our method effectively produces finer wrinkle details over previously dominant skinning-based baselines.

Acknowledgement. This research is funded in part by an ARC (Australian Research Council) Discovery Grant of DP220100800. Prof. Hongdong Li holds concurrent appointments as a Full Professor with the ANU and as an Amazon Scholar (part-time). This paper describes works performed at ANU and is not associated with Amazon.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021. 5

[2] Hedy Attouch, Roberto Lucchetti, and Roger J-B Wets. The topology of the ρ-hausdorff distance. *Annali di Matematica pura ed applicata*, 160(1):303–320, 1991. 5

[3] David Baraff and Andrew Witkin. Large steps in cloth simulation. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 767–778. 2023. 1

[4] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: clothed 3d humans. In *European Conference on Computer Vision*, pages 344–359. Springer, 2020. 5, 8

[5] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Pbns: Physically based neural simulator for unsupervised garment pose space deformation. *arXiv preprint arXiv:2012.11310*, 2020. 2

[6] Sofien Bouaziz, Sebastian Martin, Tiantian Liu, Ladislav Kavan, and Mark Pauly. Projective dynamics: Fusing constraint projections for fast simulation. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 787–797. 2023. 2

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 4

[8] Ruochen Chen, Liming Chen, and Shaifali Parashar. Gaps: Geometry-aware, physics-based, self-supervised neural garment draping. In *2024 International Conference on 3D Vision (3DV)*, pages 116–125. IEEE, 2024. 2

[9] David Clyde, Joseph Teran, and Rasmus Tamstorf. Modeling and data-driven parameter estimation for woven fabrics. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 1–11, 2017. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[11] 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. 2, 6, 7

[12] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 2

[13] Jingfan Guo, Fabian Prada, Donglai Xiang, Javier Romero, Chenglei Wu, Hyun Soo Park, Takaaki Shiratori, and Shunsuke Saito. Diffusion shape prior for wrinkle-accurate cloth registration. *arXiv preprint arXiv:2311.05828*, 2023. 2

[14] 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. 4

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[16] 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. 2, 3, 6

[17] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 5

[18] Egor Larionov, Marie-Lena Eckert, Katja Wolff, and Tuur Stuyck. Estimating cloth elasticity parameters using position-based simulation of compliant constrained dynamics. *arXiv preprint arXiv:2212.08790*, 2022. 2

[19] J. P. Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, page 165–172, USA, 2000. ACM Press/Addison-Wesley Publishing Co. 2

[20] 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. 2

[21] Yifei Li, Hsiao-yu Chen, Egor Larionov, Nikolaos Sarafianos, Wojciech Matusik, and Tuur Stuyck. Diffavatar: Simulation-ready garment optimization with differentiable simulation. *arXiv preprint arXiv:2311.12194*, 2023. 2, 3

[22] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. *arXiv preprint arXiv:2311.16096*, 2023. 2

[23] Zuzeng Lin, Ailin Huang, and Zhewei Huang. Collaborative neural rendering using anime character sheets. *arXiv preprint arXiv:2207.05378*, 2022. 2

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 5

[25] Christiane Luible and Nadia Magnenat-Thalmann. The simulation of cloth using accurate physical parameters. *CGIM 2008, Insbruck, Austria*, 2008. 1

[26] Qianli Ma, Jinlong Yang, Michael J Black, and Siyu Tang. Neural point-based shape modeling of humans in challenging clothing. In *2022 International Conference on 3D Vision (3DV)*, pages 679–689. IEEE, 2022. 1, 2

[27] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 64–73, 2021. 2

[28] 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. 5

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez,

Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[30] 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. 1, 2, 5, 6, 8

[31] 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, 5, 6, 7

[32] Nico Pietroni, Corentin Dumery, Raphael Falque, Mark Liu, Teresa A Vidal-Calleja, and Olga Sorkine-Hornung. Computational pattern making from 3d garment models. *ACM Trans. Graph.*, 41(4):157–1, 2022. 3

[33] Xavier Provot et al. Deformation constraints in a mass-spring model to describe rigid cloth behaviour. In *Graphics interface*, pages 147–147. Canadian Information Processing Society, 1995. 1, 2

[34] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019. 1, 2

[35] 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, 5, 6, 8

[36] 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. 5, 6, 7, 8

[37] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Snug: Self-supervised neural dynamic garments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8140–8150, 2022. 1, 2

[38] Libor Vasa and Vaclav Skala. A perception correlated comparison method for dynamic meshes. *IEEE transactions on visualization and computer graphics*, 17(2):220–230, 2010. 5

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[40] Rong Wang, Wei Mao, and Hongdong Li. Interacting hand-object pose estimation via dense mutual attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5735–5745, 2023. 4

[41] Rong Wang, Fabian Prada, Ziyan Wang, Zhongshi Jiang, Chengxiang Yin, Junxuan Li, Shunsuke Saito, Igor Santesteban, Javier Romero, Rohan Joshi, et al. Fresa: Feedforward reconstruction of personalized skinned avatars from few images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 281–291, 2025. 2

[42] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 2

[43] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 512–523, 2023. 2

[44] Shan Yang, Junbang Liang, and Ming C Lin. Learning-based cloth material recovery from video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4383–4393, 2017. 2

[45] Xinyuan Yu, Siheng Zhao, Siyuan Luo, Gang Yang, and Lin Shao. Diffclothai: Differentiable cloth simulation with intersection-free frictional contact and differentiable two-way coupling with articulated rigid bodies. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 400–407. IEEE, 2023. 2

[46] Joy Xiaoji Zhang, Gene Wei-Chin Lin, Lukas Bode, Hsiao-yu Chen, Tuur Stuyck, and Egor Larionov. Estimating cloth elasticity parameters from homogenized yarn-level models. *arXiv preprint arXiv:2401.15169*, 2024. 1

[47] Meng Zhang, Tuanfeng Wang, Duygu Ceylan, and Niloy J Mitra. Deep detail enhancement for any garment. In *Computer Graphics Forum*, pages 399–411. Wiley Online Library, 2021. 2, 6

[48] Meng Zhang, Duygu Ceylan, and Niloy J Mitra. Motion guided deep dynamic 3d garments. *ACM Transactions on Graphics (TOG)*, 41(6):1–12, 2022. 1, 2

[49] Renrui Zhang, Liuhui Wang, Ziyu Guo, Yali Wang, Peng Gao, Hongsheng Li, and Jianbo Shi. Parameter is not all you need: Starting from non-parametric networks for 3d point cloud analysis. *arXiv preprint arXiv:2303.08134*, 2023. 1

[50] Fang Zhao, Zekun Li, Shaoli Huang, Junwu Weng, Tianfei Zhou, Guo-Sen Xie, Jue Wang, and Ying Shan. Learning anchor transformations for 3d garment animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 491–500, 2023. 1, 2, 5, 6, 8