DOI: 10.1111/cgf.15028

EUROGRAPHICS 2024 / A. Bermano and E. Kalogerakis
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 43 (2024), Number 2

Neural Garment Dynamics via Manifold-Aware Transformers

Peizhuo Li[1], Tuanfeng Y. Wang[2], Timur Levent Kesdogan[1], Duygu Ceylan[2], Olga Sorkine-Hornung[1]

[1]ETH Zurich, Switzerland [2]Adobe Research, United Kingdom

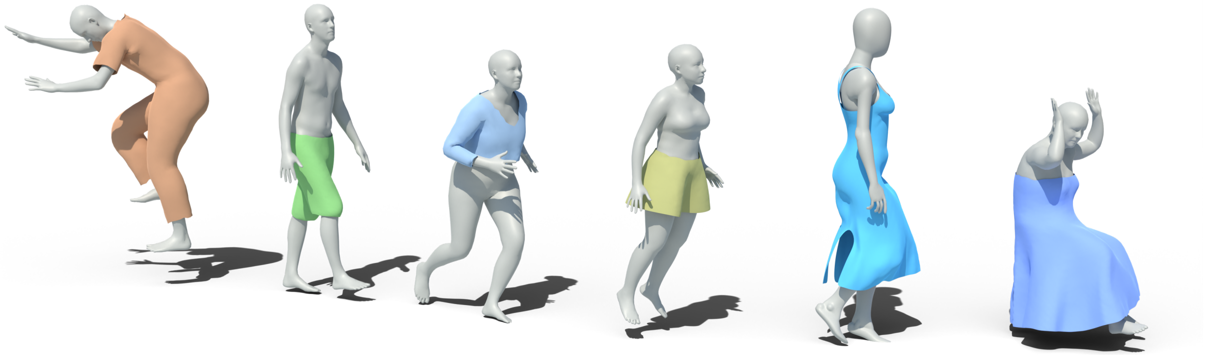

Figure 1: *We propose a neural garment dynamics inference network powered by manifold-aware transformers. Our approach can be directly applied to unseen garments, bodies, as well as motions that were not included in the training data.*

Abstract

Data driven and learning based solutions for modeling dynamic garments have significantly advanced, especially in the context of digital humans. However, existing approaches often focus on modeling garments with respect to a fixed parametric human body model and are limited to garment geometries that were seen during training. In this work, we take a different approach and model the dynamics of a garment by exploiting its local interactions with the underlying human body. Specifically, as the body moves, we detect local garment-body collisions, which drive the deformation of the garment. At the core of our approach is a mesh-agnostic garment representation and a manifold-aware transformer network design, which together enable our method to generalize to unseen garment and body geometries. We evaluate our approach on a wide variety of garment types and motion sequences and provide competitive qualitative and quantitative results with respect to the state of the art.

1. Introduction

Modeling the dynamics of garments as they interact with an underlying collider, such as a moving human body, is a core component for many graphics applications, e.g., animation [WCPM18], virtual try-on [STOC21], video editing [YAP*16], etc. There are two main directions to tackle this problem, i.e., physically-based simulation [NMK*06; NSO12], and learning-based neural approaches [PLP20; BME21]. Physically-based simulation provides a generic framework to produce plausible and accurate geometric details with realistic motion dynamics. However, acquiring and setting the physical parameters used in the simulation is not easy and often the simulation process is sensitive to initial conditions [ZWCM21]. To address such issues, learning-based approaches are getting popular in recent years. A typical learning-based workflow [WCPM18; PMJ*22] models the garment dynamics as a mapping from the encoded garment and collider (i.e., body) motion features to a latent code representing the garment shape in the next frame. The compact latent representation enables efficient inference and acts as a regularizer. However, such global approaches are difficult to generalize to unseen garment and collider geometries during training.

In this work, we introduce a novel and generalizable learning-based framework for predicting the garment dynamics by modeling the local interaction between the garment and the underlying collider. Specifically, we represent the garment deformation with a continuous deformation field [SCL*04; LSC*04] where we treat each face of the garment geometry as a sample of this field. We define a set of *garment and interaction features* for each face to encode the state of the garment relative to the underlying body as the body moves from the current to the next frame. Such features inherently encode how each local patch on the garment geometry interacts with the underlying body. We accumulate such features over the past few frames to encode the dynamic behavior and augment them with global context (i.e., the global velocity of the garment). Finally, we train a neural network to predict the deformation gradient of each garment face given this local and global context. Given per-face predictions, we perform a Poisson solve to obtain the final garment geometry in the next frame. Our network runs in an autoregressive manner by utilizing its past predictions when computing the garment and interaction features that are provided as input to the network in future frames.

Central to our approach is a transformer-based network architecture [VSP*17] that predicts the deformation of a target point (e.g., the centroid of a face) when given tokens that represent the aforementioned features sampled on a random set of points on the garment surface. The transformer architecture is capable of modeling long-range correlations between how different parts of the garment deform. While spatial proximity is a strong cue for similar deformation behavior, in certain cases, spatially close garment parts can have very different dynamic behavior. For example, imagine two sides of a skirt with a cut (see Figure 9) where nearby points on two sides of the cut behave differently. In order to handle such challenging cases, we empower our transformer to be *manifold aware*. Specifically, we utilize the geodesic distance matrix obtained from the rest state of the garment as part of the attention weights. This encourages the predicted deformation to preserve the geodesic distance in the resulting garment geometry.

Our approach generalizes to a variety of garment types and geometries. We evaluate our method on garment and body types and motion sequences unseen during training. Our approach produces plausible garment geometry with vivid dynamics and performs competitively with respect to the state-of-the-art learning-based approaches. In summary, our main contributions are:

- We present a generalizable learning-based approach that predicts plausible garment dynamics for unseen garment and body types.
- We present a novel *manifold-aware* transformer architecture that incorporates both spatial and topological information and is agnostic to the underlying meshing density, and can generalize to garments with unseen local connectivity changes such as cuts.

2. Related Work

2.1. Physics-based Simulation

Modeling the dynamics of the garment w.r.t. the underlying collider motion has been studied in computer graphics for more than 30 years [TPBF87; MW88]. Physics-based methods [Mül08; MHHR07; NSO12] tend to model the garment dynamics with real-world physics based on material properties and deform them according to laws of physics using time integration and collision response. The focus of this community includes material modeling [BTH*03; MBT*12], mechanical modeling [CK05; VMF09], as well as collision modeling [HVS*09; LKJ20], and more recently, converting the whole pipeline to a differentiable setup [LDW*22; LLK19] for inverse problems. Physics-based workflows often produce high-quality dynamics but suffer from high computational costs and the tediousness of tuning the material property for a desired effect.

2.2. Data-driven and Learning-based Methods

Prior to the bloom of deep learning, there have been several data-driven approaches to model garment deformations. Aguiar et al. [DSTH10] propose to learn a linear dynamic system on the PCA subspace of garment deformation driven by a pre-defined body to achieve real-time performance. Guan et al. [GRH*12] also explore statistical models to tackle how garments drape on different body shapes and poses. Luo et al. [LSW*18] use a lightweight neural network to transform a linear elasticity-based deformation into a non-linear deformation. Holden et al. [HDDN19] introduce neural networks that adapt subspace representations, making it possible to handle interactions between multiple objects. While being efficient, subspace-based methods are often limited to the training data and difficult to extrapolate to unseen settings.

In order to leverage the recent success of neural network architectures in the 2D domain, recent approaches have utilized 2D canonical representations (i.e., UV mapping) for encoding the garment deformation. DeepWrinkle [LCT18] predicts pose-dependent wrinkles represented as normal maps in the UV space. Zhang et al. [ZWCM21] propose to refine the details of coarse simulation using a similar representation. In addition to pose-dependent effects, Jin et al.[JZGF20] aim to model motion-dependent deformations. While such UV-based representations effectively utilize 2D convolutions, they are limited in encoding spatial neighborhoods across UV seam boundaries.

Predicting the 3D geometry of the garment directly is considered an alternative solution. Gundogdu et al. [GCS*19], Habermann et al. [HLX*21] fuse body and garment geometry features to predict pose-dependent effects in the canonical pose, and deform the result with linear blend skinning (LBS). Patel et al. [PLP20] also incorporate the dynamics and the change of garment styles, but model the garment as a height field over the body surface which is limited to tight-fitting garments. Zhang et al. [ZCM22] handle loose garments by first learning a plausible deformation latent space but still training garment-specific networks. Pan et al. [PMJ*22] utilize a virtual skeleton with additional virtual joints to better capture the low-frequency of loose garments with respect to the body motion. Similar to D. Li et al. [DTY*22], high-frequency displacements are then added with a graph neural network. While showing impressive progress, many of these methods, however, are specific to a training garment.

The time-consuming generation of training data with physically-based simulation often acts as a bottleneck for data-driven approaches. Hence, several unsupervised methods have been proposed recently. Bertiche et al. [BME21] use physically-inspired loss terms in combination with LBS deformation to predict pose-dependent deformations. A similar approach is also used by De Luigi et al. [DLG*22]. Santesteban et al [SOC22] extend the idea to and introduce strain and inertia-based losses. [BME22] train a network to predict the garment status such that it minimizes a combination of energy terms. We provide comparisons to some of these methods in the experiments section.

While learning-based methods have shown remarkable advances in recent years, generalization, i.e., generalizing to unseen garment types, still remains a challenge. In most recent concurrent approaches, GarSim [TB23] and HOOD [GTBH23] tackle this chal-

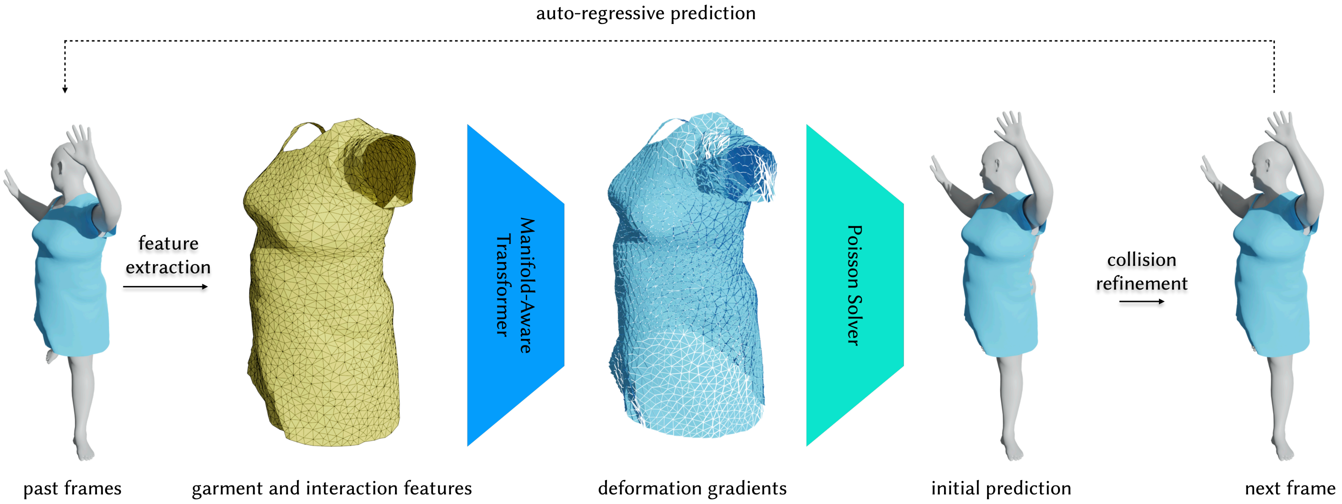

Figure 2: *Our framework overview. We extract the garment features and interaction features on the garment geometry from the past frames. Our manifold-aware transformer is then applied spatially to the input features and predicts the relative deformation gradients to the next frame. An initial prediction is obtained with a Poisson solver. After the collision refinement, we get the prediction for the next frame. We auto-regressively repeat this process until the desired number of frames is reached.*

lenge by utilizing a graph-convolution-based framework [PFSB20]. Due to the limited receptive field of graph convolutions, their methods are sensitive to the resolution of the input meshes and require specific designs to incorporate more global information. We provide qualitative comparisons in the supplementary material regarding these methods.

2.3. Transformers

Proposed by Vaswani et al. [VSP*17] for natural language processing, transformers have quickly been adapted to many areas, including 3D tasks [GCL*21; DB20; CZG*22; YSW*23]. We also adapt a transformer-based architecture and introduce a *manifold-aware* structure to capture both local mesh neighborhoods and long-distance dependencies effectively.

3. Overview

Given the motion of an underlying collider, such as a human body, our goal is to predict the deformation of the garment as it interacts with the body. Our key insight is that garment deformations can be predicted in a generalizable manner by modeling the local interaction between the garment and the body surface. We introduce a set of garment features (e.g., the deformation gradient, velocity, and relative distance between the garment and the body) that capture such local interactions. We form tokens from the features obtained from a set of triangle faces sampled on the garment surface mesh. We introduce a manifold-aware transformer network that utilizes such tokens obtained from a set of past frames to predict the deformation of the garment in the current frame. We empower the transformer network to be *manifold-aware* by encoding the geodesic information of the garment in addition to the local interaction features. Specifically, we replace part of the learned attention weights with the geodesic matrix obtained from the garment surface.

We represent the garment deformation using a Jacobian field [AGK*22]. We discretize the continuous Jacobian field with random samples obtained on the mesh representation of the garment. Given the predicted deformation gradient, we solve a Poisson equation to reconstruct the explicit garment surface. Our approach is agnostic to the connectivity and the density of the triangulation of the garment surface and can handle garments with various topologies. We provide the overall architecture of our framework in Figure 2 and next discuss the details of our approach.

4. Method

4.1. Pipeline

We represent the garment and the underlying body geometry with their corresponding vertex positions at a particular time t and the mesh triangulation. Specifically, let $\{\mathbf{V}^t, \mathbf{T}^g\}$ denote the garment with triangulation $\mathbf{T}^g$ and vertex positions $\mathbf{V}^t$ and let $\{\mathbf{U}^t, \mathbf{T}^b\}$ denote the underlying body with triangulation $\mathbf{T}^b$ and vertex positions $\mathbf{U}^t$. Given the previous states of the garment in the past n_{hist} frames, i.e., $\{\mathbf{V}^{t-n_{\text{hist}}+1}, \cdots, \mathbf{V}^t\}$, and the state of the body in the next frame, i.e., $\mathbf{U}^{t+1}$, our goal is to predict $\mathbf{V}^{t+1}$, i.e., the state of the garment in the next frame. We introduce a transformer-based network that takes as input a set of features $\mathbf{F}_i^t$ computed for each face on the garment surface utilizing the past and current states of the garment and the underlying body. The output is the relative deformation gradients Ψ^{t+1} of each face in the next frame, and global velocity q^{t+1} of the entire geometry. We then compute the absolute deformation Φ^{t+1} to reconstruct the garment mesh at time $t+1$ via the Possion equation [SP04; SCL*04]. To reduce the accumulation of error, we also predict the singular values Σ^{t+1} for Φ^{t+1}

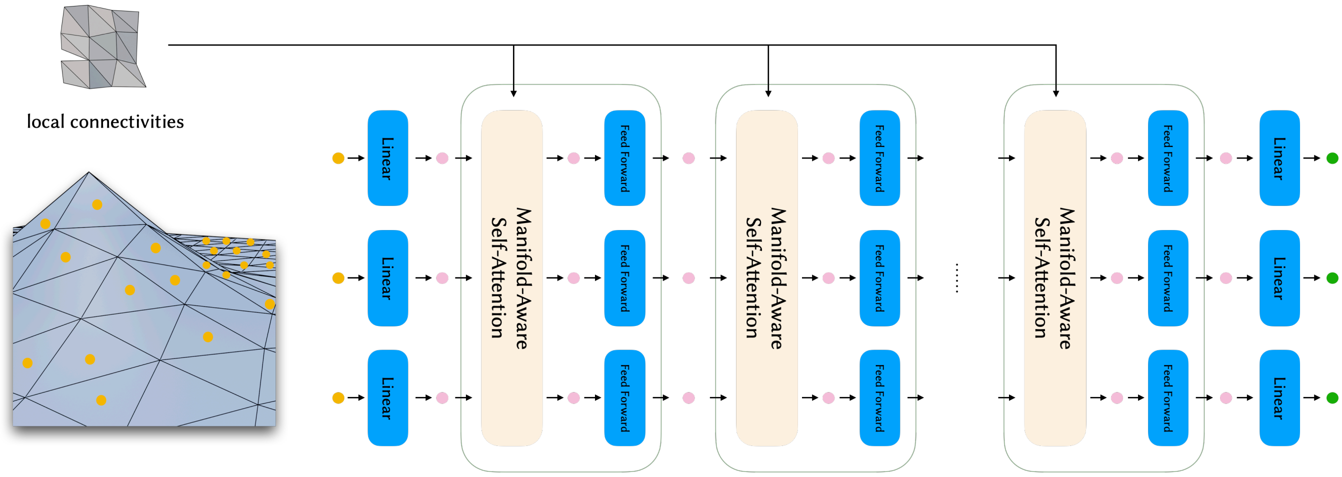

Figure 3: *Our manifold-aware transformer architecture. The input features are extracted from the faces of the input mesh. After being projected into embedding space by a linear transformation, they are fed into the transformer encoder consisting of n_l identical layers. Our manifold-aware self-attention layers explicitly involve local connectivities of the input geometry, making it possible to predict accurate dynamics caused by seams. The output of the transformer encoder is projected to the output features by another linear transformation. The output features are then used to predict the next frame of garment deformation.*

at the next frame and use it to regularize the stretching. Since the predicted cloth geometry is not guaranteed to be collision-free with the body geometry, we use a post-processing strategy adopted from DRAPE [GRH*12]. A pseudo-code of this process is shown in Algorithm 1. We refer the readers to the supplementary material for a detailed discussion on the deformation gradients and the post-process collision refinement method.

Algorithm 1 Prediction of frame $t+1$ from the past n_{hist} frames

```
Input: Garment vertex positions $\mathbf{V}^{t-n_{\text{hist}}:t}$, body vertex position $\mathbf{U}^{t-n_{\text{hist}}:t+1}$
Output: Vertex positions $\mathbf{V}^{t+1}$
procedure PREDICTFRAME($\mathbf{V}^{t-n_{\text{hist}}:t}, \mathbf{U}^{t-n_{\text{hist}}:t+1}$)
    $\mathbf{F}^t \leftarrow$ features extracted from garment $\mathbf{V}$ and body $\mathbf{U}$
    $\Psi^{t+1}, \Sigma^{t+1}, q^{t+1} \leftarrow$ prediction from the network with $\mathbf{F}^t$
    $\bar{\Phi}^{t+1} \leftarrow \Psi^{t+1}\Phi^t$
    $\Phi^{t+1} \leftarrow$ replace singular values of $\bar{\Phi}^{t+1}$ with $\Sigma^{t+1}$
    $\bar{\mathbf{V}}^{t+1} \leftarrow$ solve a Poisson problem with $\Phi^{t+1}$ and velocity $q^{t+1}$
    $\mathbf{V}^{t+1} \leftarrow$ collision refinement for $\bar{\mathbf{V}}^{t+1}$
end procedure
```

4.2. Garment and Interaction Features

Our network takes as input a set of features defined on the garment geometry. In order to capture the dynamics of the garment, we stack the features obtained from past n_{hist} frames together. For simplicity, we describe the features computed from a single frame and omit the frame index t in the following text unless otherwise specified.

In addition to the deformation gradient Φ_i, derived from the predictions of preceding frames, in an auto-regressive fashion into the present frame, we also define a set of features for a given triangle i to encode the current state of the garment as well as its interaction with the body. The following features encode the state of the garment geometry.

Orientation. The deformation gradient Φ_i records the deformation *relative* to the rest state. Hence, the network is not aware of the orientation of the surface. To mitigate this issue, we include the normal direction $\mathbf{n}_i$ in the world-coordinate of triangle i as part of the input feature.

Centroid. The input to the transformer network is permutation invariant. Similar to the positional encoding used in the original transformer, we include the centralized centroid coordinate $\mathbf{c}_i = c_i - z$ of each triangle to preserve the spatial order information. c_i is the centroid of triangle i and $z = 1/|T|\sum_{i\in T} c_i$ is the average centroid of the garment at a given frame.

Global velocity. The solution of the Poisson equation is not unique up to a translation. We thus incorporate the per-frame global velocity $q^t = z^t - z^{t-1}$ as part of the input.

In order to capture the interaction of the garment with the underlying body, we also define a set of *interaction* features as follows:

Signed distance. We encode the relative position of the garment with respect to the body. For every triangle i in the garment, we record its signed distance d_i to the body. In addition, we also record a direction $\vec{v}_i$ of the nearest point on the body to the centroid c_i of triangle i. This constitutes the singed distance feature $\mathbf{s}_i = \{d_i, \vec{v}_i\} \in \mathbb{R}^4$. It serves as the local coordinate of the collider, enabling us to encode collider deformation in the garment space.

Collider deformation. For the network to predict the deformation at frame $t+1$, we incorporate the deformation of the collider, i.e., the body, at frame $t+1$ as part of the input feature. To this end, for each face i in the garment in the current frame, we compute its nearest face on the body in the current frame. We define a *collider deformation feature*, $\mathbf{d}_i = \{d_i, \vec{q}_i\} \in \mathbb{R}^{12}$, where $d_i \in \mathbb{R}^{3\times 3}$ represents the relative deformation gradient to the next frame of its nearest body face and $\vec{q}_i$ is the velocity of the centroid of the nearest body face.

Note that, unlike commonly used SMPL [LMR*15] body and pose parameters, the proposed interaction features are not limited to a specific body model and can be directly applied to any other type of collider geometry. We conduct experiments showing the versatility of the proposed interaction features in Section 5.2.

4.3. Manifold-aware transformer networks

Network input and output. We denote the concatenation of the aforementioned per-triangle features as $\mathbf{f}_i = \{\Phi_i, \mathbf{n}_i, \mathbf{c}_i, \mathbf{s}_i, \mathbf{d}_i, \Sigma_i\}$. Furthermore, we collect the features of past n_{hist} frames together as $\mathbf{F}_i^t = \{\mathbf{f}_i^k\}_{k=t-n_{\text{hist}}+1}^t$. The input to our network is $\mathbf{F}^t$ along with the global velocity of the garment in the past $n_{\text{hist}} - 1$ frames denoted as $\mathbf{Q}^t = \{q^k\}_{k=t-n_{\text{hist}}+2}^t$. The network then predicts the relative deformation gradient Ψ_i^{t+1} and the singular value Σ_i^{t+1} for every face i and the global velocity q^{t+1} of the garment in the next frame.

Architecture. The overview of our network architecture is demonstrated in Figure 3. For memory and computational efficiency, we evenly split the faces into n_s disjoint subsets $\{\mathbf{T}_i\}_{i=1}^{n_s}$. For each of the split $\mathbf{T}_i \subset \mathbf{T}$, we gather the features $\{\mathbf{F}_j^t\}_{j\in\mathbf{T}_i}$ of every triangle in this split, concatenated with the global velocity feature $\mathbf{Q}^t$, as the input to the network. Note that the features $\{\mathbf{F}_j^t\}$ are calculated before the splitting, and no downsampling is involved for feature calculation. A linear transformation first maps the input features into an embedding space of dimension n_e. The embeddings of the features are then passed through n_l transformer [VSP*17] encoder layers, where the self-attention mechanism learns to capture the global context. Unlike graph convolution-based networks with only limited receptive fields, our transformer-based architecture is capable of learning long distance correlation. The output of the last encoder layer is passed through a linear layer to predict the relative deformation gradient Ψ^{t+1} and singular values Σ^{t+1}, as well as the global velocity q^{t+1} for the next frame. Note that this framework is triangulation-agnostic and is by nature not limited to a single garment. The network can be trained with multiple types of garments and colliders, and can be used to predict unseen clothes. We demonstrate the versatility of our network in Section 5.1.

Manifold-aware self-attention. For the cases with complex garments or challenging body poses, close-by spatial locations over the garment surface may have very different dynamic behavior, e.g., the sleeve opening for left/right arm of a shirt can be spatially close under a crossed arm pose while their dynamic behavior might be different as shown in Figure 8. To prevent such spurious correlations, we consider the connectivity of the garment surface. Specifically, we propose a manifold-aware self-attention mechanism. For n_{conn} heads, we use the pairwise geodesic distance between the centroids of the triangles as the attention score as below:

$$\mathbf{A}_{i\cdot} = \text{softmax}(-\mathbf{D}_{i\cdot}^{p_{\text{geo}}}), \tag{1}$$

where $\mathbf{D}_{ij}$ denote the matrix of pair-wise geodesic, p_{geo} is the exponential index to control the attention range. We visualize the geodesic attention weights for a triangle in Figure 4. For an in-depth study of the effect of manifold-aware self-attention, we refer the readers to Section 5.2.

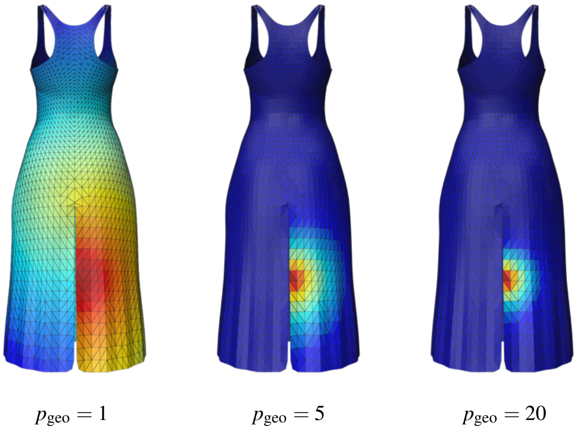

Figure 4: *Geodesic attention weights. We use p_{geo} to control the attention range.*

4.3.1. Singular value prediction

Our network predicts the relative deformation between two consecutive frames and auto-regressively predicts further frames based on existing predictions. This can lead to error accumulation of the absolute deformation gradient, and cause severe area distortion of the geometry. As singular values indicate the scaling of the deformation gradient along the three principal directions, to mitigate this issue, we also predict the singular values of the deformation gradient. During inference time, we accumulate the predicted relative deformation gradients to absolute deformation gradients. We perform SVD and replace the singular values with the predicted ones before using them to reconstruct the deformed mesh with Poisson equation. We refer the readers to Section 5.2 for the study on the effect of singular value prediction.

4.3.2. Loss functions and training

We adopt a fully-supervised training scheme to learn the deformation field from a set of physically-simulated training data. The training is supervised via the following loss terms:

Deformation gradient loss. The L1 norm is employed to measure the difference between the predicted relative deformation gradients Ψ_i^{t+1} and the ground truth $\tilde{\Psi}_i^{t+1}$:

$$\mathcal{L}_{\text{def}} = \frac{1}{|\mathbf{T}_i|} \sum_{j\in\mathbf{T}_i} \|\Psi_j^{t+1} - \tilde{\Psi}_j^{t+1}\|_1. \tag{2}$$

Singular value loss. Similarly, the L1 norm is utilized to measure the difference between the predicted singular values Σ_i^{t+1} and the ground truth $\tilde{\Sigma}_i^{t+1}$:

$$\mathcal{L}_{\text{sv}} = \frac{1}{|\mathbf{T}_i|} \sum_{j\in\mathbf{T}_i} \|\Sigma_j^{t+1} - \tilde{\Sigma}_j^{t+1}\|_1. \tag{3}$$

Global velocity loss. The L1 norm is applied to measure the

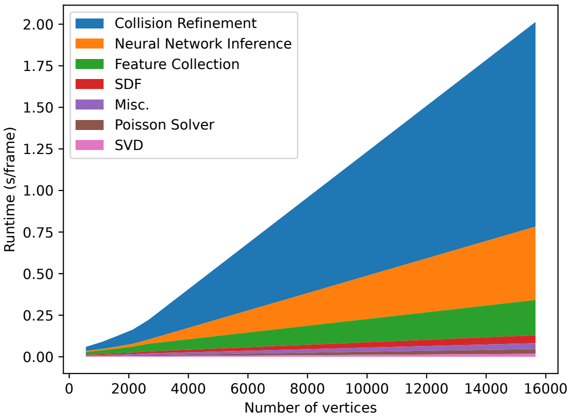

Figure 5: *Breakdown of inference performance.*

difference between the predicted global velocity q^t and the ground truth $\tilde{q}^t$:

$$\mathcal{L}_{\text{vel}} = \|q^t - \tilde{q}^t\|_1. \tag{4}$$

Our full loss used for training summarizes as:

$$\mathcal{L} = \mathcal{L}_{\text{def}} + \lambda_{\text{sv}}\mathcal{L}_{\text{sv}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}}. \tag{5}$$

During training, we supervise only one step of prediction. We random sample $n_{\text{hist}} + 1$ consecutive frames from a physically-simulated dataset and use the features of the first n_{hist} frames as input of the network, and supervise the prediction on frame $n_{\text{hist}} + 1$ using the loss functions in Equation (5). Besides, all input features are normalized to have zero mean and unit variance. To prevent error accumulation in our auto-regressive workflow, we add noise to the input features of the network during the training. Specifically, we add a random noise $\varepsilon \sim \mathcal{N}(0, \sigma_n)$ to the normalized input features $\mathbf{F}^t$. The noise is sampled independently for each feature dimension and each triangle. We find that this is a simple yet effective solution to stabilize the long-term generation. For a detailed description of the layers in our network and the specific values of the hyper-parameters, we refer to the supplementary material.

5. Experiments

We evaluate our approach on various garment and body types to demonstrate the generalization ability of our method. We also compare with other neural techniques and provide an ablation to evaluate the effectiveness of various components in our design. Please refer to the supplementary video for qualitative results.

5.1. Implementation details

Our framework is implemented in PyTorch [PGM*19], and the experiments are performed on an NVIDIA GeForce RTX 3090 GPU. We optimize the parameters of our network with the loss term in Equation (5) using the Adam optimizer [KB14]. It takes about 48 hours to train our network. We refer the readers for a detailed description of our network architecture and hyper-parameters to the supplementary material.

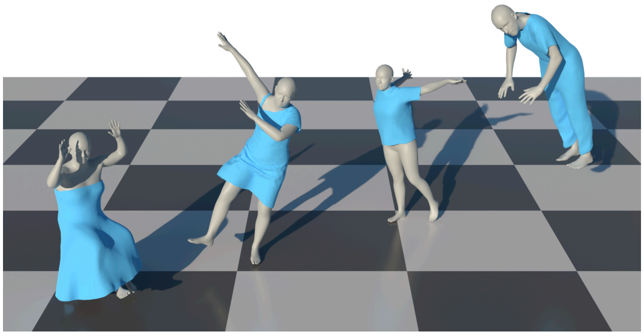

Figure 6: *A general model for different garments. Our model is capable of predicting the dynamics of different garments driven by various motions.*

Running time. We show a breakdown of the running time of each component in Figure 5. The most expensive operation is collision refinement, which involves solving a sparse linear system of size $3N \times 3N$, where N is the number of vertices. The feature collection contains the computation of the network input $\mathbf{f}_i$ excluding signed distance function (SDF), which is listed separately. Our efficient SDF calculation is implemented on GPU by combining minimum pairwise distance and winding numbers [JKS13]. The Poisson solver is implemented with a pre-computed Cholesky decomposition using CHOLMOD [CDHR08] and solving on the GPU [Nau11] using the implementation by Nicolet et al. [NJJ21].

Dataset. We train our model with the CLOTH3D dataset [BME20] which contains 7 categories of garments (shirt, shirt, top, trousers, skirt, jumpsuit, and dress). Within each category, garment shape is augmented using cutting and resizing. The garments are then simulated in 3D on a body that is animated with a motion sequence from the CMU human motion dataset [CMU19]. We randomly select 200 simulated sequences (around 50,000 frames) and train our model with 6 garment categories, excluding the "skirt" category. Since the dataset also contains several material configurations, we use the "cotton" configuration as our training data.

At test time, we evaluate our model with unseen augmentations (i.e., cutting and resizing) of the seen garment types and with garments in the unseen "skirt" category, driven by 30 unseen motion sequences.

Evaluation Metrics. We evaluate our method with respect to ground truth simulation results using the mean vertex error (in cm) and Chamfer Distance [WPZ*21]. We also measure the geodesic distortion by calculating the L1 error between the pairwise geodesic distance of the generated results and ground truth.

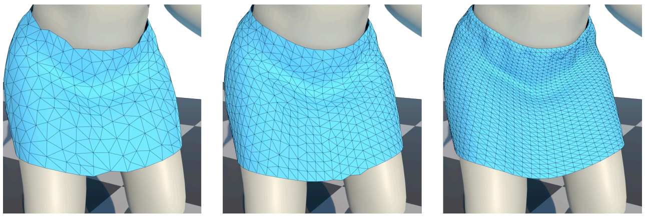

Figure 7: *Robustness to remeshing. Our network predicts consistent results for different meshing, thanks to our feature representation and network architecture.*

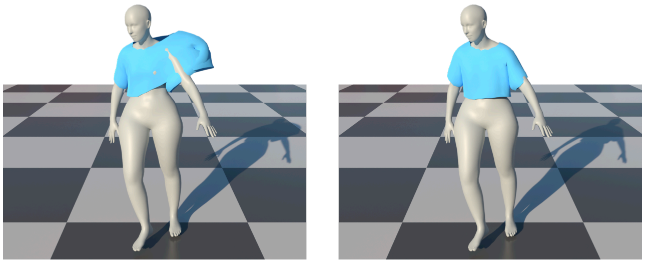

Figure 8: *Manifold awareness. The manifold aware component is able to resolve instances where spatial proximity happens to unconnected cloth segments. While without the manifold-aware component, it fails to distinguish the two segments.*

5.2. Results

As shown in Figure 6 and our supplementary video, our method can synthesize realistic results on different types of garments and correctly capture the subtle dynamics. In the following, we provide a more thorough evaluation.

Robustness to remeshing. The input and output features of our network are triangulation-agnostic. Furthermore, the manifold-aware self-attention module is also robust to changes in triangulation. We evaluate our model on the same garment with different meshing, specifically containing 1032, 2067, and 5168 faces, respectively. It can be seen in Figure 7 that our method generates consistent results across different mesh resolutions, which is not possible to handle with graph-convolution-based methods [GTBH23; TB23].

Effect of manifold-aware self-attention. In certain motion sequences, two different parts of the garment (e.g., the sleeves and the torso) can spatially come close together while their dynamics are still significantly different. As shown in Figure 8, our manifold-aware attention module can effectively handle such cases and generate plausible results.

Since our model explicitly encodes the geodesic information, it can generalize to garments with unseen seam cuts. Due to the lack of connectivity changes in the CLOTH3D dataset, we created a dataset from a given dress with 10 different seam cuts as the training set and 2 different seam cuts as the test set. It can be seen in Figure 9 that our model is able to synthesize plausible results reflecting the unseen seam cuts without re-training. Please refer to the accompanying video for a complete result.

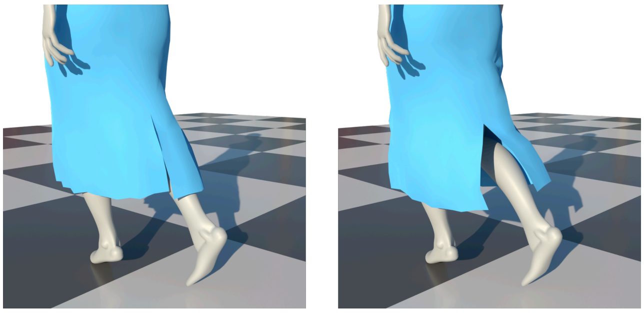

Figure 9: *Local connectivity changes, such as a cut as seen here, are incorporated into our network's prediction with the help of the manifold-aware transformer.*

Table 1: *Ablation study.*

	Mean vertex error (cm)	Geodesic distortion
w/o manifold-aware	4.54	2.83×10^{-2}
w/o singular prediction	12.2	9.33×10^{-2}
Full approach	3.19	2.05×10^{-2}

We conduct an ablation study over the manifold-aware component by using only learned attention weights. As shown in Table 1, the geodesic information not only leads to better visual performance, but also better preserves the geodesics.

Choice of p_{geo}. As can be seen in Figure 10, a too small or too large p_{geo} yields inferior results. We show that our model is robust to a large range of p_{geo} values in Table 2. When p_{geo} ranges between 1 and 50, the mean vertex error (MVE) remains small. We use $p_{\text{geo}} = 20$ in our other experiments.

Effect of singular value prediction. We evaluate the effectiveness of our singular value prediction term in Equation (3) by removing it from the loss function and directly using the predicted defor-

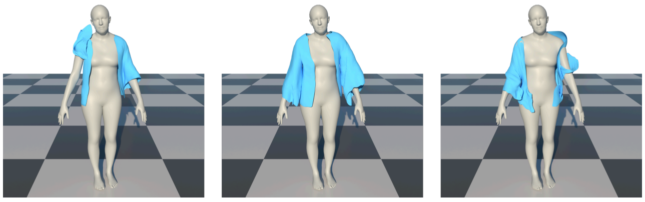

Figure 10: *Choice of p_{geo}. We show that with a suitable choice of p_{geo}, our method can handle the problem of spatial proximity, while a too small or a too large p_{geo} can lead to artifacts.*

Table 2: *Ablation study on the choices of* p_{geo}.

p_{geo}	0.01	0.1	1	10	20	50	100
MVE (cm)	4.34	3.79	3.23	3.31	3.19	3.20	4.72

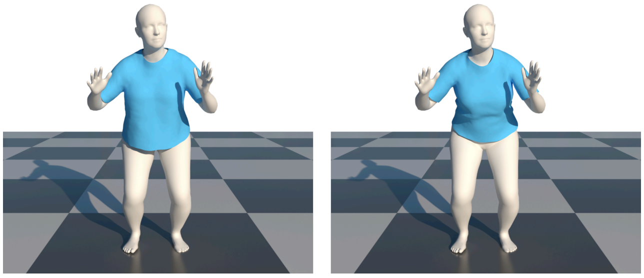
w/o singular value prediction Full approach

Figure 11: *Singular value prediction. With the help of predicting singular values of absolute deformation gradient, our method does not suffer from the over-stretching problem.*

mation gradient without replacing the singular values. As shown in Table 1 and Figure 11, the singular value prediction term helps prevent the accumulation of stretching and produces more accurate results.

Unseen body model. While our training data is driven by the SMPL [LMR*15] model, our model can be applicable to different human body geometries. As shown in Figure 12, we test our model with a mannequin, a significantly different body model from the SMPL and demonstrate plausible results. Please refer to the accompanying video for additional results.

5.3. Comparisons

Supervised methods. We compare out method to SSCH [STOC21] as a baseline supervised method. For a fair comparison, we retrain our model with the same VTO dataset this method is trained on.

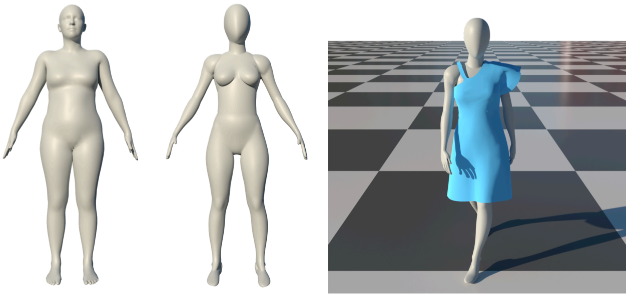
Training body Unseen body Our result

Figure 12: *Unseen body model. The mannequin wearing the dress is not seen by our model during training. However, it is still able to predict a plausible result.*

Table 3: *Quantitative comparison to the supervised method.*

	Mean vertex error (cm)	Chamfer distance
SSCH [STOC21]	2.93	3.84×10^{-4}
Ours	2.69	3.30×10^{-4}

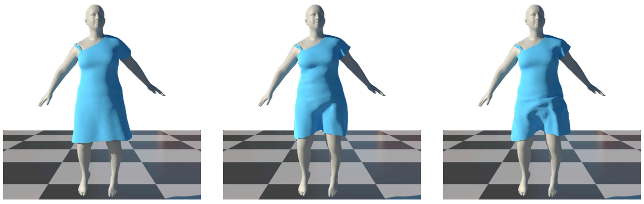
SSCH [STOC21] Ours Ground truth

Figure 13: *Comparison to SSCH. Our network is able to faithfully capture the dynamics of the garment.*

The results generated by our method faithfully reconstruct the dynamics of the garment as demonstrated in Figure 13. We also report superior quantitative results in Table 3.

Unsupervised methods. We compare our method to the state-of-the-art of unsupervised learning method SNUG [SOC22]. We use the version of our method trained on the VTO dataset that this method is also trained on. Note that SNUG requires new training for every new garment while our method provides a general model. It can be seen in Figure 14 that our model is able to generate plausible results and provides better dynamics, while the garments are not seen during training. Please refer to the accompanying video for a complete result.

Generalizable methods. Finally, we compare our method to the concurrent works [TB23; GTBH23] that tackle the generalization problem. As the code/date of GarSim [TB23] is not publicly available yet, we are limited to providing qualitative comparisons in the supplementary material. To compensate for the limited receptive field of graph convolution, HOOD [GTBH23] constructs a hierar-

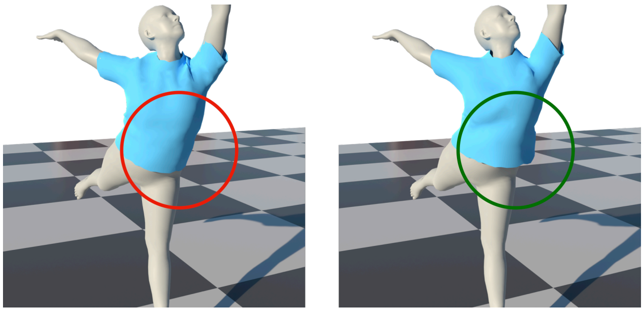
SNUG [SOC22] Ours

Figure 14: *Comparison to SNUG. The global self-attention of our model allows the cloth to deform naturally in front of the torso, yielding better visual quality, while the baseline fails to generate the same effect.*

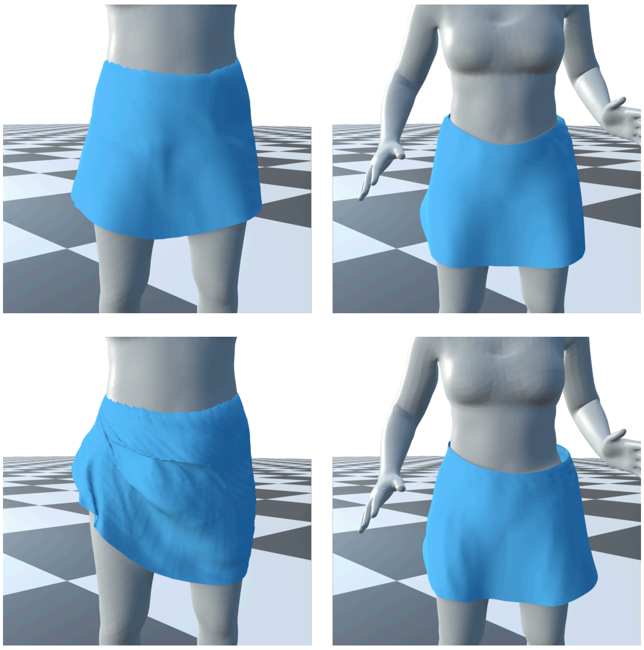

Figure 15: *Comparison to HOOD. The meshes shown in the first and second rows contain 5,169 and 31,008 faces, respectively. It can be seen that our method produces consistent results across different resolutions, whereas HOOD exhibits unnatural dynamics and artifacts when processing high-resolution input.*

chy of simplified meshes. As demonstrated in Figure 15, when provided with high-resolution input, HOOD struggles to capture correct global dynamics due to the slow propagation of these dynamics and generates undesired artifacts caused by simplified graph construction. Please refer to the accompanying video for detailed results.

6. Conclusion

In this paper, we introduce a learning-based generalizable solution for predicting garment dynamics with respect to an underlying body in motion. Previous learning-based approaches have been limited to garment-collider combinations present in the training dataset, often requiring refinement when applied to unseen cases. To address this limitation, we propose a deformation field-based garment representation combined with a transformer-based neural network. This combination enables us to handle different garment types and body models as colliders, contrary to existing methods, which are usually limited to a predefined parametric model (e.g., SMLP [LMR*15]). Additionally, we incorporate pairwise geodesic distances to weight the self-attention heads in our network, resulting in a manifold-aware transformer capable of capturing not only spatial but also topological correlations.

Our experiments demonstrate that our method can handle challenging scenarios involving complex garments and dynamic motions. We believe that our approach represents a novel direction toward modeling realistic and captivating clothing behavior for general garment dynamics and digital human modeling.

Limitations and Future work. To focus on our key insights, we have made several assumptions to simplify the problem setup. However, these assumptions also set limitations for our work and may stimulate potential future work.

Self-collision handling We do not explicitly handle the garment-to-garment collisions. Thanks to our manifold-aware transformer, garment self-interpenetrations do not affect our auto-regressive workflow and are visually hard to observe according to our experiments. However, this could still be a problem for some downstream applications and may eventually lead to physically incorrect garment modeling.

Material variation. Our training dataset is produced by the same set of parameters, so the fabric material variation is not formulated into our current model. Allowing the users to control the material property can be a bonus feature for many scenarios [WCPM18].

Fine details in predicted geometry. Although our method can handle global dynamics well, the results lack fine details such as wrinkles when compared with state-of-the-art methods, as can be seen in the collar region in Figure 14. We conjecture that this is due to the downsampling strategy and could be improved by reducing the downsampling rate and training over a longer time.

Unsupervised learning. Last but not least, recent advances in unsupervised garment dynamics learning based on physical properties [BME22; SOC22] show a promising direction with no data generation burdens. We believe having the terms from the physical constraints can potentially elaborate our approach in an unsupervised manner as well.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback. This work was supported in part by the ERC Consolidator Grant No. 101003104 (MYCLOTH), Adobe Research Internship Program, and Adobe University Gift Fund. Open access funding provided by Eidgenossische Technische Hochschule Zurich.

References

[AGK*22] AIGERMAN, NOAM, GUPTA, KUNAL, KIM, VLADIMIR G., et al. "Neural Jacobian Fields: Learning Intrinsic Mappings of Arbitrary Meshes". *ACM Trans. Graph.* 41.4 (July 2022). ISSN: 0730-0301. DOI: 10.1145/3528223.3530141. URL: https://doi.org/10.1145/3528223.3530141 3.

[BME20] BERTICHE, HUGO, MADADI, MEYSAM, and ESCALERA, SERGIO. "CLOTH3D: clothed 3d humans". *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer. 2020, 344–359 6.

[BME21] BERTICHE, HUGO, MADADI, MEYSAM, and ESCALERA, SERGIO. "PBNS: physically based neural simulation for unsupervised garment pose space deformation". *ACM Transactions on Graphics (TOG)* 40.6 (2021), 1–14 1, 2.

[BME22] BERTICHE, HUGO, MADADI, MEYSAM, and ESCALERA, SERGIO. "Neural Cloth Simulation". *ACM Transactions on Graphics (TOG)* 41.6 (2022), 1–14 2, 9.

[BTH*03] Bhat, Kiran S, Twigg, Christopher D, Hodgins, Jessica K, et al. "Estimating cloth simulation parameters from video". (2003) 2.

[CDHR08] Chen, Yanqing, Davis, Timothy A., Hager, William W., and Rajamanickam, Sivasankaran. "Algorithm 887: CHOLMOD, Supernodal Sparse Cholesky Factorization and Update/Downdate". *ACM Trans. Math. Softw.* 35.3 (Oct. 2008). ISSN: 0098-3500. DOI: 10.1145/1391989.1391995. URL: https://doi.org/10.1145/1391989.1391995 6.

[CK05] Choi, Kwang-Jin and Ko, Hyeong-Seok. "Stable but responsive cloth". *ACM SIGGRAPH 2005 Courses*. 2005, 1–es 2.

[CMU19] CMU. *CMU Graphics Lab Motion Capture Database*. May 2019. URL: http://mocap.cs.cmu.edu/ 6.

[CZG*22] Chandran, Prashanth, Zoss, Gaspard, Gross, Markus, et al. "Shape Transformers: Topology-Independent 3D Shape Models Using Transformers". *Computer Graphics Forum*. Vol. 41. 2. Wiley Online Library. 2022, 195–207 3.

[DB20] Dwivedi, Vijay Prakash and Bresson, Xavier. "A generalization of transformer networks to graphs". *arXiv preprint arXiv:2012.09699* (2020) 3.

[DLG*22] De Luigi, Luca, Li, Ren, Guillard, Benoît, et al. "DrapeNet: Generating Garments and Draping them with Self-Supervision". *arXiv preprint arXiv:2211.11277* (2022) 2.

[DSTH10] De Aguiar, Edilson, Sigal, Leonid, Treuille, Adrien, and Hodgins, Jessica K. "Stable spaces for real-time clothing". *ACM Transactions on Graphics (TOG)* 29.4 (2010), 1–9 2.

[DTY*22] D. Li, Y, Tang, Min, Yang, Yun, et al. "N-Cloth: Predicting 3D Cloth Deformation with Mesh-Based Networks". *Computer Graphics Forum*. Vol. 41. 2. Wiley Online Library. 2022, 547–558 2.

[GCL*21] Guo, Meng-Hao, Cai, Jun-Xiong, Liu, Zheng-Ning, et al. "Pct: Point cloud transformer". *Computational Visual Media* 7.2 (2021), 187–199 3.

[GCS*19] Gundogdu, Erhan, Constantin, Victor, Seifoddini, Amrollah, et al. "Garnet: A two-stream network for fast and accurate 3d cloth draping". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, 8739–8748 2.

[GRH*12] Guan, Peng, Reiss, Loretta, Hirshberg, David A, et al. "Drape: Dressing any person". *ACM Transactions on Graphics (ToG)* 31.4 (2012), 1–10 2, 4.

[GTBH23] Grigorev, Artur, Thomaszewski, Bernhard, Black, Michael J, and Hilliges, Otmar. "HOOD: Hierarchical Graphs for Generalized Modelling of Clothing Dynamics". 2023 2, 7–9.

[HDDN19] Holden, Daniel, Duong, Bang Chi, Datta, Sayantan, and Nowrouzezahrai, Derek. "Subspace neural physics: Fast data-driven interactive simulation". *Proceedings of the 18th annual ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 2019, 1–12 2.

[HLX*21] Habermann, Marc, Liu, Lingjie, Xu, Weipeng, et al. "Real-time deep dynamic characters". *ACM Transactions on Graphics (TOG)* 40.4 (2021), 1–16 2.

[HVS*09] Harmon, David, Vouga, Etienne, Smith, Breannan, et al. "Asynchronous contact mechanics". *ACM SIGGRAPH 2009 papers*. 2009, 1–12 2.

[JKS13] Jacobson, Alec, Kavan, Ladislav, and Sorkine-Hornung, Olga. "Robust inside-outside segmentation using generalized winding numbers". *ACM Transactions on Graphics (TOG)* 32.4 (2013), 1–12 6.

[JZGF20] Jin, Ning, Zhu, Yilin, Geng, Zhenglin, and Fedkiw, Ronald. "A pixel-based framework for data-driven clothing". *Computer Graphics Forum*. Vol. 39. 8. Wiley Online Library. 2020, 135–144 2.

[KB14] Kingma, Diederik P and Ba, Jimmy. "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980* (2014) 6.

[LCT18] Lahner, Zorah, Cremers, Daniel, and Tung, Tony. "Deepwrinkles: Accurate and realistic clothing modeling". *Proceedings of the European conference on computer vision (ECCV)*. 2018, 667–684 2.

[LDW*22] Li, Yifei, Du, Tao, Wu, Kui, et al. "DiffCloth: Differentiable cloth simulation with dry frictional contact". *ACM Transactions on Graphics (TOG)* 42.1 (2022), 1–20 2.

[LKJ20] Li, Minchen, Kaufman, Danny M, and Jiang, Chenfanfu. "Codimensional incremental potential contact". *arXiv preprint arXiv:2012.04457* (2020) 2.

[LLK19] Liang, Junbang, Lin, Ming, and Koltun, Vladlen. "Differentiable cloth simulation for inverse problems". *Advances in Neural Information Processing Systems* 32 (2019) 2.

[LMR*15] Loper, Matthew, Mahmood, Naureen, Romero, Javier, et al. "SMPL: A Skinned Multi-Person Linear Model". *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34.6 (Oct. 2015), 248:1–248:16 5, 8, 9.

[LSC*04] Lipman, Yaron, Sorkine, Olga, Cohen-Or, Daniel, et al. "Differential coordinates for interactive mesh editing". *Proceedings Shape Modeling Applications, 2004.* IEEE. 2004, 181–190 1.

[LSW*18] Luo, Ran, Shao, Tianjia, Wang, Huamin, et al. "NNWarp: Neural network-based nonlinear deformation". *IEEE transactions on visualization and computer graphics* 26.4 (2018), 1745–1759 2.

[MBT*12] Miguel, Eder, Bradley, Derek, Thomaszewski, Bernhard, et al. "Data-driven estimation of cloth simulation models". *Computer Graphics Forum*. Vol. 31. 2pt2. Wiley Online Library. 2012, 519–528 2.

[MHHR07] Müller, Matthias, Heidelberger, Bruno, Hennix, Marcus, and Ratcliff, John. "Position based dynamics". *Journal of Visual Communication and Image Representation* 18.2 (2007), 109–118 2.

[Mül08] Müller, Matthias. "Hierarchical position based dynamics". (2008) 2.

[MW88] Moore, Matthew and Wilhelms, Jane. "Collision detection and response for computer animation". *Proceedings of the 15th annual conference on Computer graphics and interactive techniques*. 1988, 289–298 2.

[Nau11] Naumov, Maxim. "Parallel solution of sparse triangular linear systems in the preconditioned iterative methods on the GPU". *NVIDIA Corp., Westford, MA, USA, Tech. Rep. NVR-2011* 1 (2011) 6.

[NJJ21] Nicolet, Baptiste, Jacobson, Alec, and Jakob, Wenzel. "Large steps in inverse rendering of geometry". *ACM Transactions on Graphics (TOG)* 40.6 (2021), 1–13 6.

[NMK*06] Nealen, Andrew, Müller, Matthias, Keiser, Richard, et al. "Physically based deformable models in computer graphics". *Computer graphics forum*. Vol. 25. 4. Wiley Online Library. 2006, 809–836 1.

[NSO12] Narain, Rahul, Samii, Armin, and O'brien, James F. "Adaptive anisotropic remeshing for cloth simulation". *ACM transactions on graphics (TOG)* 31.6 (2012), 1–10 1, 2.

[PFSB20] Pfaff, Tobias, Fortunato, Meire, Sanchez-Gonzalez, Alvaro, and Battaglia, Peter W. "Learning mesh-based simulation with graph networks". *arXiv preprint arXiv:2010.03409* (2020) 3.

[PGM*19] Paszke, Adam, Gross, Sam, Massa, Francisco, et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". *Advances in Neural Information Processing Systems 32*. Ed. by Wallach, H., Larochelle, H., Beygelzimer, A., et al. Curran Associates, Inc., 2019, 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf 6.

[PLP20] Patel, Chaitanya, Liao, Zhouyingcheng, and Pons-Moll, Gerard. "Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, 7365–7375 1, 2.

[PMJ*22] Pan, Xiaoyu, Mai, Jiaming, Jiang, Xinwei, et al. "Predicting loose-fitting garment deformations using bone-driven motion networks". *ACM SIGGRAPH 2022 Conference Proceedings*. 2022, 1–10 1, 2.

[SCL*04] Sorkine, Olga, Cohen-Or, Daniel, Lipman, Yaron, et al. "Laplacian surface editing". *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*. 2004, 175–184 1, 3.

[SOC22] Santesteban, Igor, Otaduy, Miguel A, and Casas, Dan. "Snug: Self-supervised neural dynamic garments". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 8140–8150 2, 8, 9.

[SP04] Sumner, Robert W and Popović, Jovan. "Deformation transfer for triangle meshes". *ACM Transactions on graphics (TOG)* 23.3 (2004), 399–405 3.

[STOC21] Santesteban, Igor, Thuerey, Nils, Otaduy, Miguel A, and Casas, Dan. "Self-supervised collision handling via generative 3d garment models for virtual try-on". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, 11763–11773 1, 8.

[TB23] Tiwari, Lokender and Bhowmick, Brojeshwar. "GarSim: Particle Based Neural Garment Simulator". *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, 4472–4481 2, 7, 8.

[TPBF87] Terzopoulos, Demetri, Platt, John, Barr, Alan, and Fleischer, Kurt. "Elastically deformable models". *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*. 1987, 205–214 2.

[VMF09] Volino, Pascal, Magnenat-Thalmann, Nadia, and Faure, Francois. "A simple approach to nonlinear tensile stiffness for accurate cloth simulation". *ACM Transactions on Graphics* 28.4 (2009), Article–No 2.

[VSP*17] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, et al. "Attention is all you need". *Advances in neural information processing systems* 30 (2017) 2, 3, 5.

[WCPM18] Wang, Tuanfeng Y, Ceylan, Duygu, Popović, Jovan, and Mitra, Niloy J. "Learning a shared shape space for multimodal garment design". *ACM Transactions on Graphics (TOG)* 37.6 (2018), 1–13 1, 9.

[WPZ*21] Wu, Tong, Pan, Liang, Zhang, Junzhe, et al. "Density-aware chamfer distance as a comprehensive metric for point cloud completion". *arXiv preprint arXiv:2111.12702* (2021) 6.

[YAP*16] Yang, Shan, Ambert, Tanya, Pan, Zherong, et al. "Detailed garment recovery from a single-view image". *arXiv preprint arXiv:1608.01250* (2016) 1.

[YSW*23] Ying, Hui, Shao, Tianjia, Wang, He, et al. "Adaptive Local Basis Functions for Shape Completion". *ACM SIGGRAPH 2023 Conference Proceedings*. 2023, 1–11 3.

[ZCM22] Zhang, Meng, Ceylan, Duygu, and Mitra, Niloy J. "Motion guided deep dynamic 3d garments". *ACM Transactions on Graphics (TOG)* 41.6 (2022), 1–12 2.

[ZWCM21] Zhang, Meng, Wang, Tuanfeng, Ceylan, Duygu, and Mitra, Niloy J. "Deep detail enhancement for any garment". *Computer Graphics Forum*. Vol. 40. 2. Wiley Online Library. 2021, 399–411 1, 2.