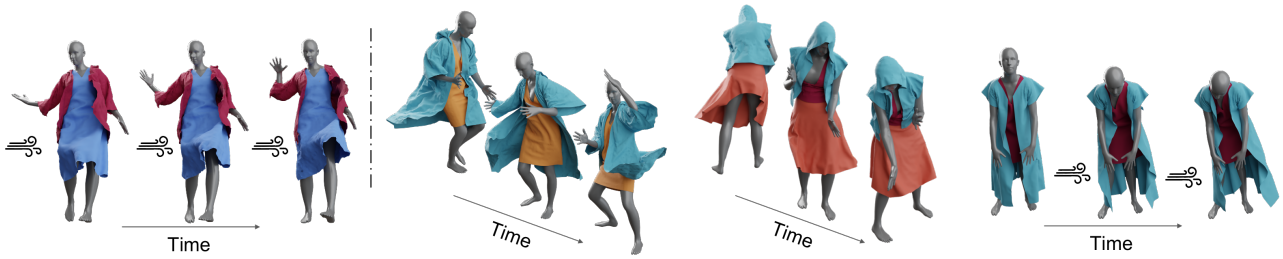
Towards Multi-Layered 3D Garments Animation

Yidi Shao[1] Chen Change Loy[1] Bo Dai[2]
[1]S-Lab for Advanced Intelligence, Nanyang Technological University
yidi001@e.ntu.edu.sg, ccloy@ntu.edu.sg
[2]Shanghai AI Laboratory
daibo@pjlab.org.cn



(a) Animations (b) Rich Dynamics (c) Triple Garments (d) Wind Blowing

Figure 1: We present LayersNet, featuring a novel Rotation Equivariant Transformation, devised to simulate garment animation with increased realism. Our LayersNet is able to animate multi-layered garments, responding to various external forces, such as human bodies and wind as shown in (a). LayersNet is powered by our proposed D-LAYERS, a novel large-scale 3D garment animation dataset involving realistic and challenging scenarios, as shown in (b)-(d).

Abstract

Mimicking realistic dynamics in 3D garment animations is a challenging task due to the complex nature of multi-layered garments and the variety of outer forces involved. Existing approaches mostly focus on single-layered garments driven by only human bodies and struggle to handle general scenarios. In this paper, we propose a novel data-driven method, called ***LayersNet****, to model garment-level animations as particle-wise interactions in a micro physics system. We improve simulation efficiency by representing garments as patch-level particles in a two-level structural hierarchy. Moreover, we introduce a novel* ***Rotation Equivalent Transformation*** *with* ***Rotation Invariant Attention*** *that leverage the rotation invariance and additivity of physics systems to better model outer forces. To verify the effectiveness of our approach and bridge the gap between experimental environments and real-world scenarios, we introduce a new challenging dataset,* ***D-LAYERS****, containing 700K frames of dynamics of 4,900 combinations of multi-layered garments driven by human bodies and randomly sampled wind. Our LayersNet achieves superior performance both quantitatively and qualitatively. Project page: www.mmlab-ntu.com/project/layersnet/index.html.*

1. Introduction

3D garment animation has been an active and important topic in computer graphics and machine learning, due to its great potential in various downstream tasks including virtual reality, virtual try-on, gaming and film production. However, realistic 3D garment animation remains an open research problem due to the intrinsic challenge of modeling garments dynamics.

Specifically, the dynamics of garments are jointly affected by both internal and external driving factors. For internal factors, while garments vary in topologies and materials, different topologies and materials result in drastically different dynamics. Moreover, in practice, humans usually wear multiple garments in a layered manner, and such multi-layered garments further complicate the problem. For example, the rigid outer layer of a jacket can press against a softer inner dress, while the inner layer of a rigid t-shirt tries to maintain its shape against outer softer clothing. As for external factors, in addition to the movement of human body, gravity, wind and friction also significantly influence the dynamics of garments in different ways. Given the complexity of 3D garment animation, previous approaches [40, 38, 20, 24] tend to simplify the problem, considering only single-layered garments with the movement of human body being the only external driving factor. Though being

effective in such a simplified setting, their applicability in real-life scenarios is significantly reduced. Moreover, they often resort to garment-specific designs, which further limits their generality across garments with different topologies and materials.

In this paper, we propose a novel data-driven method, **LayersNet**, for 3D garment animation, which is inspired by the observation that although different driving factors, garment topologies and materials lead to significantly varying garments behaviors in the macro view, at the micro level the dynamics of particles with same attributes share similarities. Therefore, LayersNet realizes a Transformer-based simulation system to capture system dynamics via *particle-wise interactions*, where garments, human body as well as other external factors are all represented by particles, making LayersNet agnostic to specific garment topology, the number of layered garments, and the set of considered external factors. In practice, we also adopt a two-level structural hierarchy in LayersNet, where garments are made of patches, and patches consist of vertices of a fixed configuration. Patches are thus treated as garments' basic particles, and LayersNet only needs to learn the interactions between patches, resulting in a significant reduction of computational complexity.

To further improve the effectiveness of LayersNet, we also propose novel **Rotation Invariant Attention** and **Rotation Equivalent Transformation**, which utilize the properties of rotation invariance and additivity of physics systems, to ease the modeling complexity of external factors. Specifically, the Rotation Invariant Attention ensures that LayersNet equivalently attends to interactions under different rotations and obeys the rotation invariance of the system. While external factors can influence garment particles in diverse directions, the behaviors of interaction forces remain consistent in local canonical spaces, which are under the directions of forces or the normals of obstacles' surfaces. For instance, the wind blows garments along the force directions, while the meshes of human skin consistently push other objects outside of the body. The proposed Rotation Equivalent Transformation thus transforms high-dimensional features to the local canonical space to reduce the redundant rotation information and capture interactions' semantics, followed by transforming features back to the global space for aggregations. In this way, it enables LayersNet to effectively exchange semantics across multiple complicated external factors.

To verify the effectiveness of LayersNet in more general cases and bridge the gap between experimental environments and real-world applications, we introduce a new challenging dataset called **D-LAYERS**, Dynamic muLtilAYerEd gaRmentS dataset. The dataset focuses on multi-layered garment animation driven by both human body and wind. Multi-layered garments in D-LAYERS are prepared as combinations of inner and outer clothes, each with different attribute values, such as bend stiffness and frictions. All garments on the same human body interact with each other, constrained by the physics laws and simultaneously affected by the wind with randomly sampled direction and strength. D-LAYERS contains 4,900 different combinations of multi-layered garments and 700k frames in total, with a maximum sequence length of 600 frames. Experiments on D-LAYERS demonstrate that LayersNet outperforms existing methods and is more generalizable in complex settings.

Our contributions can be summarized as follows: 1) We propose a Transformer-based simulation method, LayersNet, with a novel rotation equivalent transformation for 3D garment animation that uses rotation invariance and additivity of physics systems to uniformly capture and process interactions among garment parts, different garments, as well as garments against driving factors. 2) We further propose D-LAYERS, a large-scale and new dynamic dataset for 3D garment animation. The dataset and code are available at www.mmlab-ntu.com/project/layersnet/index.html .

2. Related Work

Data-driven Cloth Model. Most existing approaches aim to estimate a function that outputs garment deformations for any input by learning a parametric garment model to deform corresponding mesh templates. This is accomplished by modeling garments as functions of human pose [40], shape [38], pose-and-shape [3, 4, 35], motions [29], garment type [20, 24, 19], and extended anchors beside human joints [23]. These approaches rely heavily on SMPL-based human models and and blend weights to animate garments according to registered templates, limiting generalization due to task-specific design. To handle obstacles with arbitrary topologies, N-Cloth [15] predicts garments deformations given the states of initial garments and target obstacles. Other studies [31, 43] generate 3D garments based on UV maps. SMPLicit [8] generates garments by controlling clothes' shapes and styles, but intersection-free reconstruction is not guaranteed.

In contrast to existing methods, our LayersNet animates garments by inferring garments' future positions through interactions between garment particles and other driving factors. Since driving factors are also represented by particles, garment animation simulates particle-wise interactions, which is shape-independent and generalizable to unseen scenarios. A concurrent work [12] adopts a Graph Neural Network (GNN)-based simulation network to model garment dynamics, which uses interactions between adjacent vertices and distant vertices as edges, resulting in redundant computational overhead. On the contrary, our proposed LayersNet adopts a Transformer-based network and model garment dynamics with patch-wise interactions, so

that the computational complexity is significantly reduced. In addition, the proposed novel rotation equivalent transformation further improves the effectiveness of LayersNet.

Rotation Invariant Neural Network. Many existing approaches [32, 2, 9, 10, 11, 6, 17] adopt spherical harmonics to encode higher-order interactions and achieve SE(3)-equivariance. These approaches focus on extracting and propagating rotation-invariant features through different layers. In contrast, while LayersNet is motivated by the rotation equivalent property of physics systems, we aim to rotate high-dimensional features into local canonical space using the mapped rotation matrix from 3D space to eliminate rotation effects and model interactions involving outer forces. We then rotate the learned features back to the shared hidden space for aggregation.

3D Garment Datasets. Existing 3D garment datasets are generated either synthetically [26, 24, 29, 3] or from real-world scans [42, 44, 20, 34, 7]. Synthetic datasets such as 3DPeople![26], TailorNet [24], and Cloth3D [3], mostly contain single-layered 3D garment models, and while some datasets have multiple garments, there are very few overlapping areas among different cloth pieces [5]. Layered-Garment Net [1] proposes a static multi-layered garments dataset in seven static poses for 142 bodies to generate layers of outfits from a single image, but the garments mostly consist of skinning clothes that do not follow physics laws, and interpenetration is solved by simply forcing penetrated vertices out of inner garments. To our knowledge, D-LAYERS is the first dataset to include dynamic multi-layered 3D garments. The different layers of garments have distinct attributes and interact with each other, following the laws of physics. Furthermore, we introduce wind as an extra driving factor to animate the garments, adding complexity to their dynamics given similar human movements. Our dataset provides all the necessary 3D information, allowing for easy generalization to other tasks, such as reconstructions from a single image.

Physics Simulation by Neural Network. Learning-based methods for physics simulation can be applied to different kinds of representations, e.g., approaches for grid representation [33, 39], meshes [22, 27, 41, 25], and particles [14, 37, 28, 30]. Some methods adopt GNN [14, 28, 25]. Another approach [16] focuses on accelerating gradient computation for collision response, serving as a plug-in for neural networks. A recent method, TIE [30], applies Transformer with modified attention to recover the semantics of interactions. Our LayersNet is inspired by TIE in the notion of modeling particle-wise interactions, thus inheriting the appealing properties of being topology-independent and easy to generalize to unseen scenarios. Different from TIE, we establish a two-level hierarchy structure for garments, which are made of deformable patches. We further propose a rotation equivalent transformation to extract canonical semantics under different local coordinates in high dimensions to cope with complex outer forces.

3. Methodology

Figure 2 presents an overview of our proposed method, LayersNet, which aims to animate garments faithfully, regardless of their topology and driving factors. In our case, these factors include rigid human bodies and winds. To achieve this, we introduce a patch-based garment model, which enables us to simulate the garment animation in a particle-wise manner. The main novelty of LayersNet lies in our use of the properties of rotation invariance and additivity of physics systems. Specifically, we propose Rotation Invariant Attention and Rotation Equivalent Transformation to enable the communication and aggregation of semantics from different canonical spaces in a unified manner. In the following sections, we describe our particle simulation formulation for garment animation, the patch-based garment model, the Rotation Invariant Attention, and the Rotation Equivalent Transformation.

3.1. LayersNet

Problem Formulation. We denote each mesh at time t by $M^t = \{\boldsymbol{V}^t, \boldsymbol{E}^M, \boldsymbol{E}^W\}$, where $\boldsymbol{V}^t = \{\boldsymbol{x}_i^t, \dot{\boldsymbol{x}}_i^t, \ddot{\boldsymbol{x}}_i^t\}_{i=1}^N$ are the vertices' positions, velocities, and accelerations, and $\boldsymbol{E}^M$ denote the mesh edges. $\boldsymbol{E}^W$ are the world space edges [25], where we dynamically connect node i and node j if $|\boldsymbol{x}_i^t - \boldsymbol{x}_j^t| < R$, excluding node pairs already exist in the mesh. In a particle-based system, each mesh is represented by particles, which correspond to the vertices of the mesh. During simulation, particle i and particle j will interact with each other only if an edge e_{ij} connects them, where $e_{ij} \in \boldsymbol{E}^M \cup \boldsymbol{E}^W$. The interactions guided by $\boldsymbol{E}^M$ enable learning internal dynamics of mesh, while interactions indicated by $\boldsymbol{E}^W$ serve to compute external dynamics such as collisions.

We adopt abstract particles to represent the garments' attributes and the wind. Specifically, we use $\boldsymbol{a}_g$ to denote each garment's attribute, such as friction and stiffness, and $\boldsymbol{w}^t$ to denote the wind. Since the wind has constant strength in the whole 3D space, we use the quaternion rotation $\boldsymbol{\eta}^t$ and the strength s^t to represent the wind as $\boldsymbol{w}^t = \{\boldsymbol{\eta}^t, s^t\}$. In this way, given the human body and wind at $t+1$ as well as their previous h states, we predict the garments' states at time $t+1$ given the current states at t and corresponding previous meshes $\{M^{t-1}, \cdots, M^{t-h}\}$. In practice, we choose $h = 1$ in all experiments. Our approach can be described as:

$$\hat{\boldsymbol{V}}_g^{t+1} = \Gamma(\boldsymbol{a}_g, \{M_g^{t-i}, M_b^{t+1-i}, \boldsymbol{w}^{t+1-i}\}_{i=0}^h), \quad (1)$$

where M_g^t and M_b^{t+1} are the meshes of garments and human body, respectively, $\Gamma(\cdot)$ is the simulator and runs recursively during predictions, and $\hat{\boldsymbol{V}}_g^{t+1}$ is the garment's new vertices'

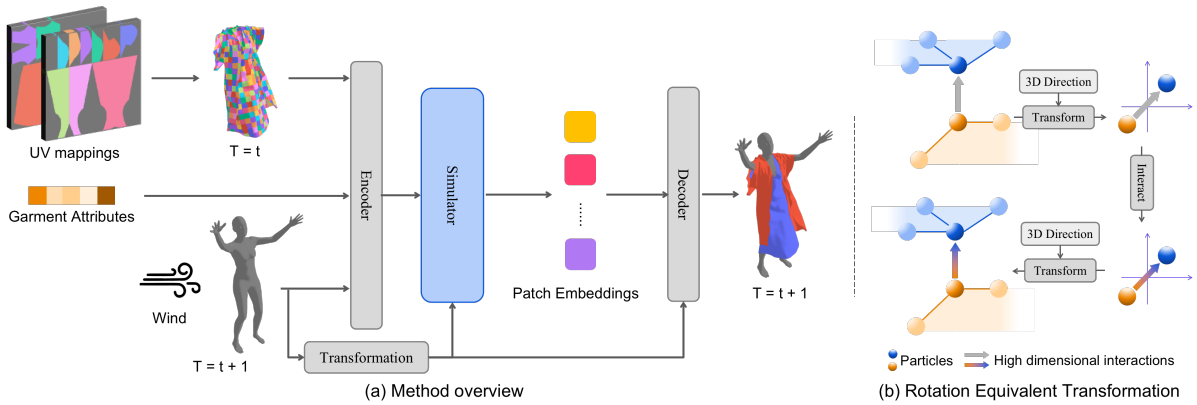


Figure 2: (a). Overview of LayersNet. Given driving factors at time $t+1$, i.e., the human body model and environmental wind in our study, LayersNet animates target garments at time t and predicts the new states of garments at time $t+1$. While all objects are represented by particles. we establish a two-level structural hierarchy for garments, as shown on the top left of the figure, where garments are made of patches given the UV mappings. Then we encode the particles and model the interactions among them by a simulator, which outputs the embeddings for each patch. We apply a decoder to decode the vertices' dynamics at time $t+1$ given neighbor patches features. The Rotation Equivalent Transformation (RET) is applied to both simulator and decoder. (b) Key ideas about RET. In high dimensional space, we transform the interactions between garments' and external forces' particles into canonical spaces, which are defined by 3D directions, such as vertex normals, of corresponding external forces', and extract the semantics of interactions, followed by the transformation back to the shared hidden space for aggregation. The high dimensional transformation is calculated by our rotation network, which converts rotation matrix in 3D space to hidden space.

states at time $t+1$. We adopt an encoder to embed the inputs into hidden space, and an decoder to decode the hidden features back to the 3D states.

Patch-based Garment Model. Since garments are composed of hundreds and thousands of particles, modeling interactions between densely connected particles inevitably leads to significant computational overhead. To reduce the number of interactions, we establish a two-level structural hierarchy for garments and represent each garment by patches, which consist of vertices of a fixed configuration. Patches are treated as special particles and interact with each other during simulations instead of densely connected vertices. Patch modeling holds several advantages. First, as basic units to represent garments, patches are topology independent. By modeling the dynamics of each patch, our model is more flexible and generalizable to unseen garments. Second, instead of simulating each vertex in a mesh, simulating patches significantly reduces the computational overhead, especially when the mesh is of high fidelity.

Formally, we find a mapping $\rho(\cdot)$ to map the vertex-based mesh to patch-based representation by:

$$P_g^t = \rho(M_g^t), \tag{2}$$

where $P_g^t = \{\boldsymbol{V}_p^t, \boldsymbol{E}_p^M, \boldsymbol{E}_p^W\}$. The patches' states $\boldsymbol{V}_p^t$ are the averaged vertices' states within the patches, and $\boldsymbol{E}_p^W$ are computed given $\boldsymbol{V}_p^t$. The mapping $\rho(\cdot)$ is based on the garments' UV maps as shown in Figure 2. In this way, our method can be updated as:

$$\hat{\boldsymbol{V}}_g^{t+1} = \Gamma(\boldsymbol{a}_g, \{P_g^{t-i}, M_b^{t+1-i}, \boldsymbol{w}^{t+1-i}\}_{i=0}^h). \tag{3}$$

Rotation Invariant Attention. Physics systems used for garment simulations possess two essential properties: rotation invariance and additivity. The rotation equivariance property states that the interactions' effects between objects remain the same regardless of the objects' rotations, while the additivity property implies that the total influence towards a particle equals the summation of each component's influence. By exploiting these two properties, we can segregate the impact of directed forces, such as forces brought by complex surface human bodies and directed wind, into individual interactions, solve them within their canonical space, and then aggregate the results. We assume the z-axis of the canonical space is the direction of human model vertex normal or the wind field, while the remaining two axes can be randomly selected, thanks to rotation invariance. To ensure that our Transformer-based model equivalently pays attention to features under different rotations, we apply decentralization and normalization for attention keys (Equation 5), and propose a rotation-invariant attention mechanism

$$\boldsymbol{q}_i = W_q \boldsymbol{v}_i, \quad \boldsymbol{r}_i = W_r \boldsymbol{v}_i, \quad \boldsymbol{s}_i = W_s \boldsymbol{v}_i, \tag{4}$$

$$\boldsymbol{f}_{i,j} = \frac{\boldsymbol{r}_i + \boldsymbol{s}_j - \boldsymbol{\mu}_{\boldsymbol{r}_i,\boldsymbol{s}_j}}{\sigma_{\boldsymbol{r}_i,\boldsymbol{s}_j}}, \tag{5}$$

$$\omega_{ij} = \mathrm{softmax}(\boldsymbol{q}_i^\top \boldsymbol{f}_{i,j}), \tag{6}$$

where $\boldsymbol{v}_i$ is state token, $\boldsymbol{q}_i$ is query token, $\boldsymbol{r}_i$ is receiver token and $\boldsymbol{s}_j$ is sender token, W_q, W_r, W_s are trainable parameters. $\boldsymbol{\mu}_{\boldsymbol{r}_i,\boldsymbol{s}_j} = (\boldsymbol{r}_i + \boldsymbol{s}_j)/2$ is the mean vector of $\boldsymbol{r}_i$ and $\boldsymbol{s}_j$, while $\sigma_{\boldsymbol{r}_i,\boldsymbol{s}_j}$ is the corresponding standard deviation. The choices of $\boldsymbol{\mu}_{\boldsymbol{r}_i,\boldsymbol{s}_j}$ and $\sigma_{\boldsymbol{r}_i,\boldsymbol{s}_j}$ ensure Equation 5 is

rotation equivariant, which decentralizes the feature vectors and normalizes them by the averaged L2 distance towards the center. The proof can be found in supplementary materials. Equation 5 can be further simplified as:

$$\boldsymbol{f}_{i,j} = \frac{\boldsymbol{r}_i + \boldsymbol{s}_j}{\|\boldsymbol{r}_i - \boldsymbol{s}_j\|}. \quad (7)$$

Rotation Equivalent Transformation. To directly extract rich semantics from high-dimensional spaces for interactions rather than 3D space, and rotate them into potential canonical space, we propose a rotation network to model high-dimensional rotations given the corresponding 3D rotations. Specifically, for each human body vertex $\boldsymbol{v}_{b_j}$, we calculate the rotation matrix $R_{b_j} \in \mathbb{R}^{3\times3}$ that transforms the 3D world space coordinates into local coordinates, where the z-axis is the normal $\boldsymbol{n}_{b_j}$ of $\boldsymbol{v}_{b_j}$. Since the physics system is rotation invariant, we can randomly sample a unit vector orthogonal to $\boldsymbol{n}_{b_j}$ as x-axis, and get the y-axis unit vector through cross product. To find the corresponding rotation matrix in the l-th layer with d dimension, we design a rotation network $\phi^l(\cdot) : \mathbb{R}^{3\times3} \to \mathbb{R}^{d\times d}$ as:

$$\phi^l(R) = W_R^l R (W_R^l)^\top, \quad (8)$$

$$\text{s.t.} \quad W_R^l (W_R^l)^\top = I, \quad (W_R^l)^\top W_R^l = I, \quad (9)$$

where $W_R^l \in \mathbb{R}^{d\times3}$ is the learnable parameter. Equation 9 ensures the rotation matrix in hidden space satisfying the property $\phi^l(R)(\phi^l(R))^\top = I$. The interactions between i-th garment patch and its neighbor human body vertex $b_j \in \mathcal{N}_i^b$ as well as the rest patches $k \in \mathcal{N}_i^p$ at l-th layer can be written as:

$$\boldsymbol{f}_{i,b_j}^R = \psi(\phi(R_{b_j})\boldsymbol{f}_{i,b_j}), \quad (10)$$

$$\boldsymbol{v}_i' = \sum_{b_j} \omega_{ib_j} (\phi(R_{b_j}))^\top \boldsymbol{f}_{i,b_j}^R + \sum_k \omega_{ik} \boldsymbol{f}_{i,k}, \quad (11)$$

where $\boldsymbol{v}_i'$ is the updated state token for i-th patch, and $\psi(\cdot)$ is multi-layer perception in practice. The first term in Equation 11 rotates the interaction features $\boldsymbol{f}_{i,b_j}^R$ from different canonical space back to the shared hidden space before aggregation. For gravity and wind, the directions of the forces are used to calculate the rotation matrices.

Finally, to recover the details of k-th vertex, we utilize its neighbor patches $p_i \in \mathcal{N}_k^p$ and the nearest point on human body indexed by b_j for decoding as follows:

$$\boldsymbol{v}_{p_i}^R = \phi(R_{b_j})\boldsymbol{v}_{p_i}, \quad \boldsymbol{v}_{b_j}^R = \phi(R_{b_j})\boldsymbol{v}_{b_j}, \quad (12)$$

$$\boldsymbol{\alpha}_k^{R,t+1} = \frac{1}{N_k} \sum_{p_i} g([R_{b_j}(\bar{\boldsymbol{x}}_k^t - \bar{\boldsymbol{x}}_{p_i}^t), \boldsymbol{v}_{p_i}^R, \boldsymbol{v}_{b_j}^R]), \quad (13)$$

$$\boldsymbol{\beta}_k^{t+1} = \Delta t \cdot (R_{b_j})^\top \boldsymbol{\alpha}_k^{R,t+1} + \bar{\boldsymbol{\beta}}_k^t, \quad (14)$$

$$\boldsymbol{x}_k^{t+1} = \Delta t \cdot \boldsymbol{\beta}_k^{t+1} + \bar{\boldsymbol{x}}_k^t, \quad (15)$$

where we first rotate the patches' features $\boldsymbol{v}_{p_i}$, human body vertex features $\boldsymbol{v}_{b_j}$, and the ground truth relative positions $\bar{\boldsymbol{x}}_k^t - \bar{\boldsymbol{x}}_{p_i}^t$ at time t before concatenation. We average the output of decoder $g(\cdot)$ as the predicted 3D acceleration $\boldsymbol{\alpha}_k^{R,t+1}$, which is further transformed back to global 3D coordinates to compute the velocity $\boldsymbol{\beta}_k^{t+1}$ and position $\boldsymbol{x}_k^{t+1}$ at time $t+1$. Δt is the time interval between each frame.

3.2. Training Details

To train our simulation-based model, we first apply a standard mean square error (MSE) loss as:

$$\mathcal{L}_{m,*}^{t+1} = \frac{1}{N} \sum_i \|\boldsymbol{x}_i^{t+1} - \bar{\boldsymbol{x}}_i^{t+1}\|_2^2, \quad (16)$$

where $\{\boldsymbol{x}_i^{t+1}\}_{i=1}^N, \{\bar{\boldsymbol{x}}_i^{t+1}\}_{i=1}^N$ are the predictions and ground truths at time $t+1$ respectively. We penalize the MSE loss on both garment vertices' positions $\mathcal{L}_{m,g}^{t+1}$ and the center of patches' positions $\mathcal{L}_{m,p}^{t+1}$ together as $\mathcal{L}_m^{t+1} = \mathcal{L}_{m,g}^{t+1} + \mathcal{L}_{m,p}^{t+1}$.

We adopt a loss term for the garment vertex normal to maintain the smoothness and consistency as:

$$\mathcal{L}_n^{t+1} = \frac{1}{N_v} \sum_i \|\boldsymbol{n}_i^{t+1} - \bar{\boldsymbol{n}}_i^{t+1}\|_2^2, \quad (17)$$

where $\boldsymbol{n}_i^{t+1}$ and $\bar{\boldsymbol{n}}_i^{t+1}$ are the vertex normals for prediction and ground truth, respectively.

To further reduce the collision rates between garments and human bodies, as well as between different layers of garments, we adopt collision loss:

$$\mathcal{L}_c^{t+1} = \frac{1}{N_c} \sum_i \max\left(d_\epsilon - (\boldsymbol{x}_i^{t+1} - \boldsymbol{x}_a^{t+1})\boldsymbol{n}_a^{t+1}, 0\right)^2, \quad (18)$$

where $\boldsymbol{x}_a^{t+1}$ is the nearest anchor point to $\boldsymbol{x}_i^{t+1}$, N_c is the number of collided vertices, and d_ϵ is the minimum distance of penetration. The collision loss between garments and ground truth human bodies, as well as between predictions of layers of garments, can be denoted as $\mathcal{L}_{c,\bar{b}}^{t+1}$ and $\mathcal{L}_{c,g}^{t+1}$ respectively. Thus, for predictions at time $t+1$, our training loss is written as:

$$\mathcal{L}^{t+1} = \lambda_m \mathcal{L}_m^{t+1} + \lambda_n \mathcal{L}_n^{t+1} + \lambda_b \mathcal{L}_{c,\bar{b}}^{t+1} + \lambda_g \mathcal{L}_{c,g}^{t+1}. \quad (19)$$

During training, we randomly rollout T_n steps without gradient given inputs at time $t-T_n$, which aims to add noise from the model itself. We only back-propagate gradients from one-step predictions on time $t+1$.

4. D-LAYERS Dataset

Most existing datasets are limited to single-layered garments driven solely by human bodies. Different garments,

Table 1: We display the influence of multi-layered garments and wind with different combinations for garment animations. We list the components of different splits and models' corresponding Euclidean errors (mm) below. Notice that in our D-LAYERS, all objects are scaled up 10 times than real-world size. We sample four splits from our dataset: inner garments are tight clothes without wind (**T**); inner garments are tight clothes with strong wind (**T+W**); inner garments are loose clothes without wind (**L**); inner garments are loose clothes with strong wind (**L+W**). The models marked by $*$ are trained and tested on the inner garment only. Notice that MGNet has worse generalization abilities due to garment-specific design. LayersNet achieves superior and robust performance in most cases especially those with multi-layered garments. More comparisons can be found in Appendix.

Splits	**Components**			**Methods on Inner Garment**		**Methods on Layered Garments**		
	InnerCloth	OuterCloth	WindStrength	MGNet* [43]	LayersNet*(Ours)	DeePSD [4]	GarSim [36]	LayersNet(Ours)
T	Jumpsuit	Jacket	≤ 50	5219.2±1565.8	**220.0±195.3**	1072.6±694.7	920.2±608.9	**489.9±447.7**
T+W	Jumpsuit	Jacket	> 250	5186.8±1754.8	**258.0±312.8**	1121.0±731.7	999.5±686.0	**508.5±562.7**
L	Dress	Jacket	≤ 50	4432.7±1438.0	**344.7±244.5**	887.8±460.5	929.6±509.7	**378.0±293.0**
L+W	Dress	Jacket	> 250	4595.0±1215.2	**347.4±288.9**	1083.1±492.2	1050.0±523.4	**467.8±403.7**

such as the upper T-shirt and lower pants, rarely interact with each other. Consequently, the problem can be easily solved by modeling garments as functions of human bodies with single-layered outfits predictions [24, 4]. Collecting a real-world dataset with dynamic multi-layered garments and outer forces is expensive and usually contains noisy artifacts, such as interpenetration [20] between scanned clothes and estimated SMPL-based human bodies, while synthetic data are easier to obtain and can provide more accurate dynamics in most cases, particularly for multi-layered clothes with narrow gaps. With this motivation, we generated D-LAYERS using a simulation engine and Blender[1], making it the first dynamic multi-layered garments dataset that considers the wind factor in addition to human bodies.

We construct our dataset by first collecting garment templates from SewPattern [13], which includes various types of garments such as jackets with hoods and dresses with waist belts. We then generate multi-layered combinations of outer and inner-layer clothes. Each multi-layered garment combination is draped onto an SMPL human body model [18], followed by a warm-up simulation in Blender to resolve interpenetrations. Finally, we simulate the dynamics of the garments given human motion sequences from CMU MoCap in AMASS [21] and sampled winds. To preserve high-frequency details in Blender, we scale up the human and garment meshes ten times their real-world size before simulation. Given the availability of the 3D meshes and attributes of garments, as well as the detailed scene settings for each sequence, D-LAYERS offers the potential to extend to other formats of data and support explorations of alternative topics such as optical flow estimations, 3D reconstructions from images, and physics parameter estimations. Supplementary materials provide additional details on the key settings in D-LAYERS. Here, we highlight the two main settings:

[1] https://www.blender.org/

Multi-layered Garments. Each multi-layered outfit in D-LAYERS consists of an inner and outer outfit with different garment attributes, such as mass, stiffness, and friction, leading to more diverse and flexible dynamics. For example, the outer garments can be softer or more rigid than the inner outfit. The outer outfit in our dataset is either a jacket or a jacket with a hood, providing a clear view of interactions from the inside and outside. Inner outfits refer to whole-body outfits, such as dresses, jumpsuits, and t-shirts with pants or skirts. We generate 4,900 combinations of multi-layered garments, which includes 9,872 different garments in total. The garment templates are of high fidelity, with vertices ranging from 5,000 to more than 15,000 for each garment, enabling us to capture more details in simulation.

Wind. Most existing datasets simplify real-life scenarios by driving garment animation solely through human bodies. To enrich the simulation settings and enable researchers to explore garment animation driven by multiple factors, we introduce randomly sampled wind in D-LAYERS. Wind is a common and prominent force field that influences garment dynamics in the real world. To simulate wind in our dataset, we randomly select several intervals of frames in each sequence and apply winds with varying directions and strengths as force fields. The directions and strengths are uniformly sampled from 0 to 400 in Blender. Within each interval, we assume that the wind affects the entire 3D space, with the direction and strength remaining constant.

5. Experiments

5.1. Baselines

We implement DeePSD [4], MGNet [43] and GarSim [36] as our baselines. MGNet is a standard garment-specific model, while DeePSD and GarSim model garments as functions of human bodies and achieve state-of-the-art performance in terms of 3D garment animations. We make the following extensions to baselines: 1. We add wind as extra

Table 2: Euclidean error (mm) on sampled D-LAYERS with maximum sequence length of 35 frames. The collision rates between different layers of garments are shown under **L-Collision**, while the collision rates between garments and human bodies are shown under **H-Collision**. Models trained with collision loss $\mathcal{L}_{c,b}$, $\mathcal{L}_{c,g}$ are marked by +. Our LayersNet achieves superior results in all cases.

Methods	**Jacket**	**Jacket + Hood**	**Dress**	**Jumpsuit**	**Skirt**
DeePSD+ [4]	1830.1±803.3	1566.0±527.1	1333.0±349.2	1219.0±186.8	1194.7±311.2
GarSim+ [36]	1412.1±886.8	1139.1±653.5	674.4±451.8	317.8±157.4	689.9±386.7
LayersNet(Ours)	571.9±451.9	493.9±354.2	397.2±342.2	264.0±200.2	301.3±79.3
LayersNet+(Ours)	**567.3±425.5**	**491.4±361.3**	**379.1±299.7**	**260.1±222.2**	**299.5±92.3**

Methods	**Pants**	**T-shirt**	**Overall**	**L-Collision(%)**	**H-Collision(%)**
DeePSD+ [4]	1185.7±213.3	1202.9±233.6	1563.4±486.8	8.78±5.12	19.47±6.38
GarSim+ [36]	317.8±150.1	447.6±303.8	1028.3±581.0	6.03±4.23	15.11±7.11
LayersNet(Ours)	234.4±206.3	273.3±169.0	472.8±343.5	**3.13±2.22**	10.68±4.53
LayersNet+(Ours)	**200.9±140.1**	**267.8±189.6**	**467.2±330.7**	3.77±2.60	**2.16±1.46**

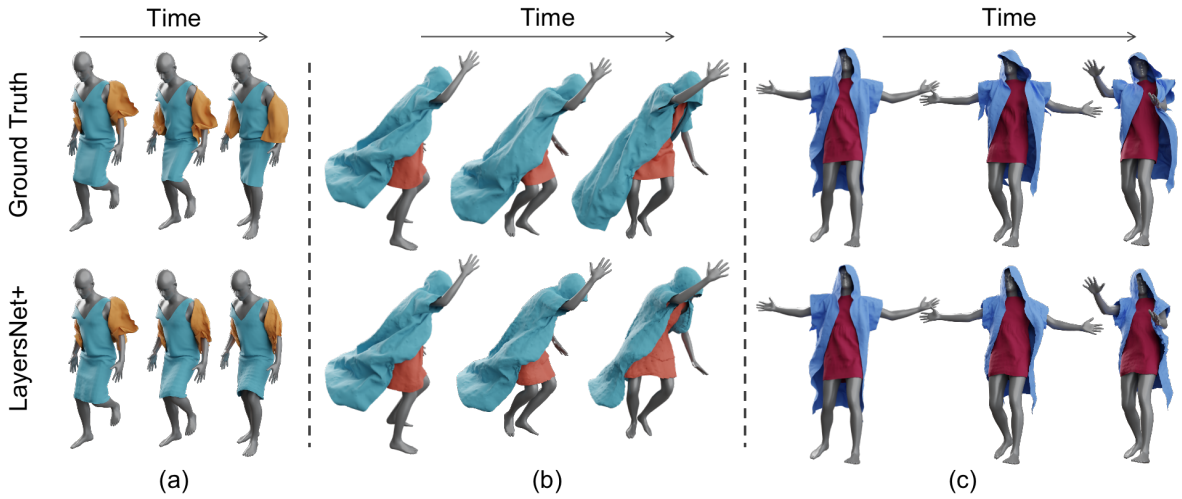


Figure 3: Qualitative results by our LayersNet+. Sequences in test set are mutually exclusive from training set samples. LayersNet+ is capable of generalizing to unseen scenarios with faithful and realistic rollouts in terms of vivid dynamics constrained by physics laws, low rates of garment-to-body and garment-to-garment interpenetration.

inputs; 2. We add the collision loss between different layers of garments. The second extension only applies to multi-layered clothes setting of DeePSD and GarSim, and does not apply to MGNet due to its specific design for single-layered clothes. All models are trained with ten epochs. We do not apply any post-processing for both the training and prediction stages. During the evaluation, we calculate the mean of Euclidean errors for each frame, then average the errors across all the frames within each sequence. The final results are the mean of errors from all sequences.

5.2. Garment Animations on D-LAYERS

Influence of Multi-Layered Garments and Wind. As shown in Table 1, we test models' abilities to animate garments on both simplified settings and general scenarios. The former assumes single-layered garments driven by human bodies, while the latter tests the models with multi-layered garments under the influences of both human bodies and wind. Specifically, we sample and divide our D-LAYERS into four splits according to the types of garments and the strength of wind as indicated in Table 1. We train and test models with either only inner garments, which is marked by *, or multi-layered garments on these four splits. Notice that we group winds with a strength less than 50 as not windy, where the wind has little influence on the garments. Each split contains 36K frames for training, 2K frames for validation, and 2K frames for test. The training set, validation set, and test set are mutually exclusive, thus they differ in human motions, garment topology and attributes.

As shown in Table 1, MGNet fails in our dataset due to the garment-specific design and low generalization abilities. Our LayersNet has lower errors, especially on splits with loose inner garments (L and L+W), suggesting the effectiveness and higher generalization abilities to animate loose clothes.

When trained on multi-layered garments, DeePSD and GarSim obtain higher Euclidean errors, indicating that they

Figure 4: Samples of qualitative results by DeePSD and GarSim with collision loss. They have difficulties in capturing the rich dynamics of highly flexible garments in our D-LAYERS, leading to difficulties in convergence. The low rates of garment-to-garment interpenetration result from the interpenetration-free initialized garment templates.

struggle with modeling complex, layered clothing. In contrast, LayersNet consistently demonstrates superior performance on all splits, handling both inner and outer garments effectively. The Euclidean errors remain similar across different splits, suggesting that our model exhibits greater robustness to varying garment topologies and external factors beyond human bodies.

General Garment Animations. As demonstrated in Table 2 and Figure 3, we further animate garments under more general conditions, featuring various combinations of multi-layered garments driven by human bodies and wind. Since DeePSD and GarSim outperform MGNet, we primarily compare our LayersNet with both DeePSD and GarSim. For training, we uniformly sample 50K frames from D-LAYERS, along with 6K frames for validation and 6K frames for testing. There is no overlap among these sample sets. All samples include both inner and outer garments, as well as random wind as an external factor.

As shown in Table 2, though GarSim achieves better performance than DeePSD, they exhibits high Euclidean errors across all garment types. Since GarSim relies heavily on the garments' priors, which are obtained from the linear blend skinning of garments' templates and tend to stay very close to the surface of the human body, it has lower Euclidean errors for skinny clothes, such as the jumpsuit, but struggles with loose garments, especially the highly flexible outer jackets. The intricate dynamics introduced by multi-layered garments and wind disrupt DeePSD and GarSim, causing convergence difficulties as depicted in Figure 4. As a result, DeePSD and GarSim fails to accurately predict the garments' lively movements and leads to extensive garment-to-body collisions and garment-to-garment interpenetration. The relatively low collision rates between garment layers stem from the interpenetration-free initialization of garment templates. This feature allows DeePSD and GarSim to automatically avoid some collisions when using linear blend skinning to deform the templates.

In contrast, LayersNet delivers superior performance in terms of Euclidean errors and collision rates, demonstrating the effectiveness of our simulation-based formulation powered by rotation equivalent transformation. Our method also shows outstanding generalization for various garment types. Notably, LayersNet achieves low collision rates and small Euclidean errors without penalizing collisions explicitly, resulting in more accurate outcomes. Since the core concept of simulation involves modeling object interactions, such as energy transitions and collisions, LayersNet can resolve collisions implicitly. By incorporating collision loss, LayersNet further minimizes interpenetration and strikes a balance between Euclidean errors and collision rates. We display the qualitative results of our LayersNet in Figure 3. Additional qualitative comparisons can be found in the supplementary materials.

Table 3: Ablation studies in terms of euclidean error (mm) and collision rates. We analyze the effectiveness of our Rotation Invariant Attention (RIA), Rotation Equivalent Transformation (RET), and the impacts of different collision loss. The vanilla LayersNet adopts the RIA in Equation 5, which improves LayersNet's performance over TIE [30]. LayersNet with default loss term $\mathcal{L}_{c,b}$,$\mathcal{L}_{c,g}$ adopts weights $\lambda_b = 1.0, \lambda_g = 0.1$, while LayersNet with optimal combinations of $\mathcal{L}^*_{c,b}$,$\mathcal{L}^*_{c,g}$ adopts weight $\lambda_b = 1.3$. RET improves LayersNet's performance with fewer interpenetrations.

LayersNet	Overall	L-Collision(%)	H-Collision(%)
TIE [30]	501.5±326.2	3.63±2.54	14.69±6.41
Vanilla	472.8±330.7	3.13±2.22	10.68±4.53
+ $\mathcal{L}_{c,b}$	**446.3±304.9**	3.10±2.16	9.51±4.55
+ RET, $\mathcal{L}_{c,b}$	449.2±315.1	5.21±3.34	2.72±1.60
+ RET, $\mathcal{L}_{c,b}$,$\mathcal{L}_{c,g}$	466.3±333.3	**2.62±2.28**	4.28±2.09
+ RET, $\mathcal{L}^*_{c,b}$,$\mathcal{L}^*_{c,g}$	467.2±330.7	3.77±2.60	**2.16±1.46**

Ablation Study. We investigate the effectiveness of our Rotation Invariant Attention (RIA), Rotation Equivalent Transformation (RET), and the impact of various collision losses in Table 3. Specifically, in our vanilla LayersNet, we only apply the RIA in Equation 5, which enables LayersNet to obtain higher accuracy than simply applying the attention mechanism in TIE [30]. LayersNet with garment-to-human collision loss attains lower Euclidean errors and reduces collision rates between garments and human bodies. When trained with RET, LayersNet dramatically decreases garment-to-human penetration by over 71%, from 9.51% to 2.72%, while maintaining low Euclidean errors and garment-to-garment collision rates. This suggests that RET effectively eliminates redundant information from different rotations and enhances LayersNet's ability to capture the semantics of complex interactions. Due to its modeling of particle-wise interactions, which implicitly accounts for collisions, LayersNet still achieves relatively low garment-to-garment penetration rates without $\mathcal{L}_{c,g}$. Although $\mathcal{L}_{c,g}$ slightly increases garment-to-body penetrations, an optimal combination of $\mathcal{L}_{c,g}$ and $\mathcal{L}_{c,b}$ jointly benefits LayersNet , producing accurate predictions with low errors.

6. Conclusion

In this paper, we introduce a Transformer-based simulation method, named LayersNet, designed to animate diverse garments represented by patch-wise particles within a two-level hierarchical structure. The newly proposed Rotation Equivalent Transformation leverages the rotation equivariance and additivity of physics systems, enabling LayersNet to effectively generalize across various garment animation scenarios. Moreover, we propose a large-scale, novel garment animation dataset called D-LAYERS, aiming to bridge the gap between experimental environments and realistic situations. D-LAYERS is a dynamic animation dataset governed by physical laws, encompassing 4,900 distinct combinations of multi-layered garments and a total of 700K frames, with sequences extending up to 600 frames in length. As demonstrated by our experiments, LayersNet delivers superior, robust performance, showcasing compelling generalization capabilities.

Acknowledgement. This study is supported under the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). It is also supported by Singapore MOE AcRF Tier 2 (MOE-T2EP20221-0011) and Shanghai AI Laboratory (P23KS00020, 2022ZD0160201).

References

[1] Alakh Aggarwal, Jikai Wang, Steven Hogue, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. Layered-garment net: Generating multiple implicit garment layers from a single image. In *ACCV*, 2022. 3

[2] Brandon M. Anderson, Truong-Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks. In *NeurIPS*, 2019. 3

[3] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. CLOTH3D: Clothed 3D humans. In *ECCV*, 2020. 2, 3

[4] Hugo Bertiche, Meysam Madadi, Emilio Tylson, and Sergio Escalera. DeePSD: Automatic deep skinning and pose space deformation for 3D garment animation. In *ICCV*, 2021. 2, 6, 7

[5] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3D people from images. In *ICCV*, 2019. 3

[6] Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik J. Bekkers, and Max Welling. Geometric and physical quantities improve E(3) equivariant message passing. In *ICLR*, 2022. 3

[7] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-modal 4D human dataset for versatile sensing and modeling. *ECCV*, 2022. 3

[8] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. SMPLicit: Topology-aware generative model for clothed people. In *CVPR*, 2021. 2

[9] Fabian Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. SE(3)-Transformers: 3D roto-translation equivariant attention networks. In *NeurIPS*, 2020. 3

[10] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. GemNet: Universal directional graph neural networks for molecules. In *NeurIPS*, 2021. 3

[11] Johannes Gasteiger, Chandan Yeshwanth, and Stephan Günnemann. Directional message passing on molecular graphs via synthetic coordinates. In *NeurIPS*, 2021. 3

[12] Artur Grigorev, Bernhard Thomaszewski, Michael J. Black, and Otmar Hilliges. HOOD: Hierarchical graphs for generalized modelling of clothing dynamics. *CVPR*, 2023. 2

[13] Maria Korosteleva and Sung-Hee Lee. Generating datasets of 3D garments with sewing patterns. In *NeurIPS Datasets and Benchmarks*, 2021. 6

[14] Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B. Tenenbaum, and Antonio Torralba. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. In *ICLR*, 2019. 3

[15] Y. D. Li, M. Tang, Y. Yang, Z. Huang, R. F. Tong, S. C. Yang, Y. Li, and Dinesh Manocha. N-Cloth: Predicting 3D cloth deformation with mesh-based networks. *Comput. Graph. Forum*, 2022. 2

[16] Junbang Liang, Ming C. Lin, and Vladlen Koltun. Differentiable cloth simulation for inverse problems. In *NeurIPS*, 2019. 3

[17] Yi Liu, Limei Wang, Meng Liu, Yuchao Lin, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3D molecular graphs. In *ICLR*, 2022. 3

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 2015. 6

[19] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J. Black. SCALE: Modeling clothed humans with a surface codec of articulated local elements. In *CVPR*, 2021. 2

[20] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3D people in generative clothing. In *CVPR*, 2020. 1, 2, 3, 6

[21] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019. 6

[22] Charlie Nash, Yaroslav Ganin, S. M. Ali Eslami, and Peter W. Battaglia. PolyGen: An autoregressive generative model of 3D meshes. In *ICML*, 2020. 3

[23] 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 *SIGGRAPH*, 2022. 2

[24] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style. In *CVPR*, 2020. 1, 2, 3, 6

[25] Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W. Battaglia. Learning mesh-based simulation with graph networks. In *ICLR*, 2021. 3

[26] Albert Pumarola, Jordi Sanchez, Gary P. T. Choi, Alberto Sanfeliu, and Francesc Moreno. 3DPeople: Modeling the geometry of dressed humans. In *ICCV*, 2019. 3

[27] Yi-Ling Qiao, Junbang Liang, Vladlen Koltun, and Ming C. Lin. Scalable differentiable physics for learning and control. In *ICML*, 2020. 3

[28] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. Learning to simulate complex physics with graph networks. In *ICML*, 2020. 3

[29] Igor Santesteban, Nils Thuerey, Miguel A. Otaduy, and Dan Casas. Self-supervised collision handling via generative 3D garment models for virtual try-on. In *CVPR*, 2021. 2, 3

[30] Yidi Shao, Chen Change Loy, and Bo Dai. Transformer with implicit edges for particle-based physics simulation. In *ECCV*, 2022. 3, 8

[31] Yu Shen, Junbang Liang, and Ming C. Lin. GAN-based garment generation using sewing pattern images. In *ECCV*, 2020. 2

[32] Nathaniel Thomas, Tess E. Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. *CoRR*, 2018. 3

[33] Nils Thuerey, Konstantin Weißenow, Lukas Prantl, and Xiangyu Hu. Deep learning methods for reynolds-averaged navier–stokes simulations of airfoil flows. *AIAA Journal*, 2020. 3

[34] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. SIZER: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing. In *ECCV*, 2020. 3

[35] Lokender Tiwari and Brojeshwar Bhowmick. DeepDraper: Fast and accurate 3D garment draping over a 3D human body. In *ICCVW*, 2021. 2

[36] Lokender Tiwari and Brojeshwar Bhowmick. Garsim: Particle based neural garment simulator. In *WACV*, 2023. 6, 7

[37] Benjamin Ummenhofer, Lukas Prantl, Nils Thuerey, and Vladlen Koltun. Lagrangian fluid simulation with continuous convolutions. In *ICLR*, 2020. 3

[38] Raquel Vidaurre, Igor Santesteban, Elena Garces, and Dan Casas. Fully convolutional graph neural networks for parametric virtual try-on. *Comput. Graph. Forum*, 2020. 1, 2

[39] Rui Wang, Karthik Kashinath, Mustafa Mustafa, Adrian Albert, and Rose Yu. Towards physics-informed deep learning for turbulent flow prediction. In *ACM SIGKDD*, 2020. 3

[40] Tuanfeng Y. Wang, Tianjia Shao, Kai Fu, and Niloy J. Mitra. Learning an intrinsic garment space for interactive authoring of garment animation. *ACM Trans. Graph.*, 2019. 1, 2

[41] Zehang Weng, Fabian Paus, Anastasiia Varava, Hang Yin, Tamim Asfour, and Danica Kragic. Graph-based task-specific prediction models for interactions between deformable and rigid objects. *IROS*, 2021. 3

[42] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *CVPR*, 2017. 3

[43] Meng Zhang, Duygu Ceylan, and Niloy J. Mitra. Motion guided deep dynamic 3D garments. *ToG*, 2022. 2, 6

[44] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. DeepHuman: 3D human reconstruction from a single image. In *ICCV*, 2019. 3