Towards Multi-Layered 3D Garments Animation

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

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
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 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 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 interreact 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 $M^t = \{V^t, E^M, E^W\}$, where $V^t = \{x^t_i, \dot{x}^t_i, \ddot{x}^t_i\}_{i=1}^N$ are the vertices’ positions, velocities, and accelerations, and $E^M$ denote the mesh edges. $E^W$ are the world space edges [25], where we dynamically connect node $i$ and node $j$ if $|x^t_i - x^t_j| < R_e$, 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 $e_{ij} \in E^M \cup E^W$. The interactions guided by $E^M$ enable learning internal dynamics of mesh, while interactions indicated by $E^W$ serve to compute external dynamics such as collisions.

We adopt abstract particles to represent the garments’ attributes and the wind. Specifically, we use $a_g$ to denote each garment’s attribute, such as friction and stiffness, and $w^t$ to denote the wind. Since the wind has constant strength in the whole 3D space, we use the quaternion rotation $\eta^t$ and the strength $s^t$ to represent the wind as $w^t = \{\eta^t, s^t\}$. 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:

$$V_g^{t+1} = \Gamma(a_g, \{M_g^{t-i}, M_b^{t+1-i}, w^{t+1-i}\}_{i=0}^h),$$

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 $V_g^{t+1}$ 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 a 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:

$$ P^t_g = \rho(M^t_g), $$

where $P^t_g = \{V^t_p, E^M_p, E^W_p\}$. The patches’ states $V^t_p$ are the averaged vertices’ states within the patches, and $E^W_p$ are computed given $V^t_p$. 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:

$$ \hat{V}^t_{g+1} = \Gamma(a_g, \{P^{t-i}_g, M^{t+1-i}_b, w^{t+1-i}_{j=0}\}). $$

**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

$$ q_i = W_q v_i, \quad r_i = W_r v_i, \quad s_i = W_s v_i, \quad (4) $$

$$ f_{i,j} = \frac{r_i + s_j - \mu_{r_i, s_j}}{\sigma_{r_i, s_j}}, \quad (5) $$

$$ \omega_{i,j} = \text{softmax}(q_i^T f_{i,j}), \quad (6) $$

where $v_i$ is state token, $q_i$ is query token, $r_i$ is receiver token and $s_j$ is sender token, $W_q, W_r, W_s$ are trainable parameters. $\mu_{r_i, s_j} = (r_i + s_j)/2$ is the mean vector of $r_i$ and $s_j$, while $\sigma_{r_i, s_j}$ is the corresponding standard deviation. The choices of $\mu_{r_i, s_j}$ and $\sigma_{r_i, s_j}$ 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:

$$ f_{i,j} = \frac{r_i + s_j}{\|r_i - s_j\|}. \quad (7) $$

**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 $v_{b_j}$, we calculate the rotation matrix $R_{b_j} \in \mathbb{R}^{3 \times 3}$ that transforms the 3D world space coordinates into local coordinates, where the z-axis is the normal $n_{b_j}$ of $v_{b_j}$. Since the physics system is rotation invariant, we can randomly sample a unit vector orthogonal to $n_{b_j}$ 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 $\phi^l(\cdot): \mathbb{R}^{3 \times 3} \rightarrow \mathbb{R}^{d \times 3}$ as:

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

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

where $W^l_R \in \mathbb{R}^{d \times 3}$ 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 $b_j \in N_i^p$ as well as the rest patches $k \in N_i^p$ at l-th layer can be written as:

$$ f_{i,b_j}^l = \psi(\phi(R_{b_j}) f_{i,b_j}), \quad (10) $$

$$ v_i^t = \sum_{b_j} \omega_{ib_j} (\phi(R_{b_j}))^\top f_{i,b_j}^l + \sum_k \omega_{ik} f_{i,k}, \quad (11) $$

where $v_i^t$ is the updated state token for i-th patch, and $\psi(\cdot)$ is multi-layer perceptron in practice. The first term in Equation 11 rotates the interaction features $f_{i,b_j}^l$ 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 $p_i \in N_i^p$ and the nearest point on human body indexed by $b_j$ for decoding as follows:

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

$$ \alpha_{k,t+1}^R \leftarrow \frac{1}{N_k} \sum_{p_i} g([R_{b_j}(\bar{x}_k^t - \bar{x}_p^t), v_{p_i}^R, v_{b_j}^R]) \quad (13) $$

$$ \beta_{k,t+1}^+ = \Delta t \cdot (R_{b_j})^\top \alpha_{k,t+1}^R + \beta_k^t, \quad (14) $$

$$ x_{k,t+1}^+ = \Delta t \cdot \beta_{k,t+1}^+ + \bar{x}_k^t, \quad (15) $$

where we first rotate the patches’ features $v_{p_i}$, human body vertex features $v_{b_j}$, and the ground truth relative positions $\bar{x}_k^t - \bar{x}_p^t$ at time $t$ before concatenation. We average the output of decoder $g(\cdot)$ as the predicted 3D acceleration $\alpha_{k,t+1}^R$, which is further transformed back to global 3D coordinates to compute the velocity $\beta_{k,t+1}^+$ and position $x_{k,t+1}^+$ at time $t+1$. $\Delta t$ 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:

$$ L_{m,s}^{t+1} = \frac{1}{N} \sum_i \|x_i^{t+1} - \bar{x}_i^{t+1}\|^2_2, \quad (16) $$

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

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

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

where $n_i^{t+1}$ and $\bar{n}_i^{t+1}$ 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:

$$ L_c^{t+1} = \frac{1}{N_c} \sum_i \max(d_c - (x_i^{t+1} - x_n^{t+1}) n_n^{t+1}, 0)^2, \quad (18) $$

where $x_n^{t+1}$ is the nearest anchor point to $x_i^{t+1}$, $N_c$ is the number of collided vertices, and $d_c$ 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 $L_{c,b}^{t+1}$ and $L_{c,g}^{t+1}$ respectively. Thus, for predictions at time $t+1$, our training loss is written as:

$$ L^{t+1} = \lambda_m L_{m}^{t+1} + \lambda_n L_n^{t+1} + \lambda_b L_{c,b}^{t+1} + \lambda_g L_{c,g}^{t+1}. \quad (19) $$

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,
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 Blender1, 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 $L_{c,b}$, $L_{c,g}$ are marked by +. Our LayersNet achieves superior results in all cases.

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

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

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
GarSim to automatically avoid some collisions when using linear blend skinning to deform the templates. This feature allows DeePSD and GarSim to obtain higher accuracy than simply applying the RIA in Equation 5, which enables LayersNet to realize the semantics of complex interactions. Due to its modeling of 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.

**Ablation Study.** We investigate the effectiveness of our Rotation Invariant Attention (RIA), Rotation Equivalent Transformation (RET), and the impact 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.

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

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.
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

[26] Albert Pumarola, Jordi Sanchez, Gary P. T. Choi, Alberto Sanfelici, and Francesc Moreno. 3DPeople: Modeling the geometry of dressed humans. In ICCV, 2019. 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, Anastasiiia Varava, Hang Yin, Tamim Asfour, and Danica Kragic. Graph-based task-specific prediction models for interactions between deformable and rigid objects. IROS, 2021. 3