SwinGar: Spectrum-Inspired Neural Dynamic Deformation for Free-Swinging Garments

Tianxing Li, Rui Shi, Qing Zhu, Takashi Kanai

***Abstract*—Our work presents a novel spectrum-inspired learning-based approach for generating clothing deformations with dynamic effects and personalized details. Existing methods in the field of clothing animation are limited to either static behavior or specific network models for individual garments, which hinders their applicability in real-world scenarios where diverse animated garments are required. Our proposed method overcomes these limitations by providing a unified framework that predicts dynamic behavior for different garments with arbitrary topology and looseness, resulting in versatile and realistic deformations. First, we observe that the problem of bias towards low frequency always hampers supervised learning and leads to overly smooth deformations. To address this issue, we introduce a frequency-control strategy from a spectral perspective that enhances the generation of high-frequency details of the deformation. In addition, to make the network highly generalizable and able to learn various clothing deformations effectively, we propose a spectral descriptor to achieve a generalized description of the global shape information. Building on the above strategies, we develop a dynamic clothing deformation estimator that integrates graph attention mechanisms with long short-term memory. The estimator takes as input expressive features from garments and human bodies, allowing it to automatically output continuous deformations for diverse clothing types, independent of mesh topology or vertex count. Finally, we present a neural collision handling method to further enhance the realism of garments. Our experimental results demonstrate the effectiveness of our approach on a variety of free-swinging garments and its superiority over state-of-the-art methods.**

***Index Terms*—Clothing deformation, dynamics, spectral analysis, graph learning.**

I. Introduction

VIRTUAL humans, which are digital characters designed to resemble real humans, have been used in various industries. Realistic clothing is crucial for the appearance of virtual humans, making clothing animation an important topic in computer graphics. Physics-based approaches [1], [2] apply basic physics laws to animate cloth, but they require extensive computation and are not practical for real-time applications. Alternately, learning-based methods [3]–[5] have been proposed to predict garment deformations close to simulation results, making them promising solutions for interactive cloth animation due to their efficiency.

So far, several learning-based methods [6]–[10] have demonstrated the ability to generate plausible garment deformations under static poses. However, these models only consider isolated states in the current time, which results in a lack of dynamics in the animation, particularly for loose-fitting garments like dresses. More recently, researchers have also explored solutions for dynamic deformations using GRU (Gate Recurrent Unit) networks [11], [12], inertia loss terms [13], [14], or operating on image space [15]–[17]. However, these methods are restricted to specific garment objects and require individual training for each garment, which limits scalability. The most recent work [18] addresses this with hierarchical graphs, but its efficiency is curtailed due to the difficulty in parallel processing of sequences, preventing optimal utilization of GPU hardwares.

Undoubtedly, creating a unified learning-based model to approximate the behavior of animated garments is a difficult task due to the variety of clothing types, dynamic deformations, and nonlinear details involved. Neural networks with a carefully-designed architecture hold promise for this task, but they often display a bias towards low-frequency information, which has been criticized in other studies [4], [6], [13]. Employing a narrow-bandwidth kernel to preserve the fine-scale details like in [6] provides a spectral solution but cannot fully address the inherent bias, especially in complicated scenarios with a variety of garment shapes. On the other hand, the common practice of directly using spatial position to represent clothing also poses a challenge for the task. This representation is highly sensitive and not generalizable, leading to overfitting and poor generalization performance of the model. Therefore, exploring solutions to these challenges is crucial.

We present a learning-based method for predicting the deformation of free-swinging garments (Fig. 1). Our approach is centered on a general estimator based on graph attention mechanism and long short-term memory, capable of handling garments with arbitrary topologies. The key to accomplishing such a complex task lies in the introduction of spectral analysis techniques, which facilitate effective control over the learning of low- and high-frequency garment deformations and the generation of discriminative global shape representations for diverse garments. With our designed estimator, high-quality predictions of dynamic deformations for unseen garments can be achieved without additional training. Besides the estimator for dynamic deformation, we propose a novel collision handling method. This technique not only further removes the residual garment-body penetration in deformation

Manuscript received 28 Aug. 2023; revised 7 Nov. 2023; revised 17 Dec. 2023; accepted 19 Dec. 2023.

Tianxing Li, Rui Shi, and Qing Zhu are with the Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China (E-mail: litianxing@bjut.edu.cn; ruishi@bjut.edu.cn (corresponding author); ccgszq@bjut.edu.cn).

Takashi Kanai is with the Department of General Systems Studies, the University of Tokyo, Tokyo 153-8902, Japan (E-mail: kanait@acm.org).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes more implementation details and experimental results.

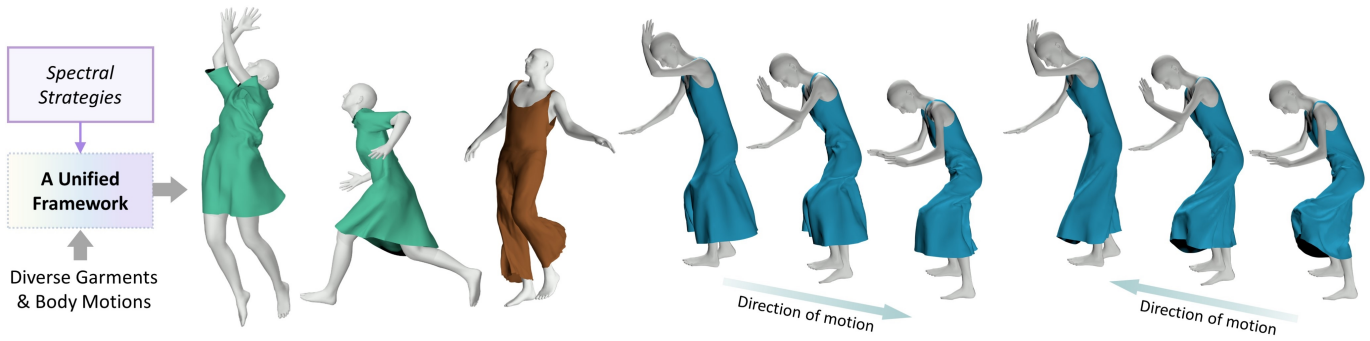


Fig. 1: We propose a learning-based method to automatically generate dynamic deformations across a variety of garments within a unified framework. By incorporating spectral strategies, our method achieves realistic generation of high-quality details while also exhibiting good generalization capabilities.

estimation using collision loss alone [4], [7], [8], but it also overcomes the barrier of requiring to learn multiple models for scenarios involving different garments [19]. Crucially, our method takes into account the fidelity of clothing details when eliminating collisions, thereby improving the overall realism of the clothing animation. In summary, our main contributions include:

- **A frequency control strategy for deformation learning.** To the best of our knowledge, we are the first to identify a relationship between the Fourier spectrum of the hidden layers and the learning of garment deformations. By integrating Lipschitz normalization into the graph attention-based network, we address the intrinsic spectral bias, resulting in an improvement in the quality of garment deformation.
- **A spectral global clothing descriptor.** Considering the diverse range of customized garments in practice, we propose a spectral descriptor that provides robust and discriminative representation for various types of garments, allowing the model to effectively learn the personalized deformations relevant to garments.
- **A general dynamic clothing estimator.** We present a novel estimator that leverages the combination of graph attention mechanism with long short-term memory, enabling the generation of dynamic deformations for unseen garments without the need for repeated training.
- **A neural collision-handling method.** We propose a collision handling method based on neural distance fields, designed to handle scenarios involving multiple garments with a unified model. By introducing corrective displacements that consider local mesh consistency, we can remove collisions while maintaining the natural details of the garments and preventing bulge artifacts.

Our proposed method has been validated through extensive experiments, demonstrating its advantages over state-of-the-art methods. The results highlight the potential of our approach to advance the field and facilitate practical applications, where the detailed insight into the ideas of learning-based deformation and spectral analysis can contribute to more future research.

II. Related Work

Physics-based simulations [20]–[24] take into account the physical properties of clothing, allowing for a high degree of realism in deformation effects. The algorithm starts by constructing a model based on the physical properties and dynamics of the garment, and then numerically solves the model equations to obtain the deformed mesh state at each time step. Researchers have developed methods [1], [25] to simulate hyperrealistic clothing, but the computational cost required for accurate results is enormous. Hence, studies have looked at ways to improve the efficiency and stability of garment simulation, including adding detailed wrinkles on low-frequency meshes [26], [27], utilizing parallel computation solutions with GPUs [28], [29], and trading some accuracy for performance in the approximation [30]. On the other hand, physics-based simulations also face the challenge of setting simulation parameters. Typically, each change of cloth properties or resetting of the mesh structure requires manual fine-tuning of parameters, which requires a certain level of expertise and significant time cost. Despite the emergence of research [31] on the automatic acquisition of simulation parameters, there is still a need for computationally inexpensive and easier-to-set-up deformation methods if diverse garments are to be applied to digital scenes.

Learning-based methods [19], [32]–[38] have gained popularity as an alternative to physics-based simulations. These methods use models for estimation and directly output the desired garment deformation.

For complicated garments of game characters, NeuroSkinning [39] explored how to apply graph neural networks to 3D mesh deformation and proposed a skinning weight approximation method that can be applied to arbitrary topological structures of meshes. Subsequently, graph learning-based approaches [40]–[43] for automatic generation of skinning weights and blend shapes have been proposed. Although the above methods provide deformation approximation models with good generalization ability and efficiency, they are still limited to producing folds and wrinkles of garments with vivid visual expression. To this end, researchers have focused on garment deformation that generates rich details for dressing the parametric human body SMPL [44]. TailorNet [6] decomposes

deformations under static pose into low- and high-frequency components, and then utilizes multiple multi-layer perceptrons (MLP) for deformation approximation, which can generate wrinkles for garment meshes with the fixed topology and number of vertices. To improve the generalization ability of deformation models, researchers use PointNet-based network [3], [4] and fully convolutional graph neural networks [5] to model deformations for various garments with different numbers of vertices. Similarly, using the graph mesh-based network, N-Cloth [45] can predict plausible garment deformation for arbitrary triangle meshes, while not being limited to SMPL bodies. One recent technique, DeePSD [7], also employs GCN. However, the weights and blend shapes it generates rely solely on the initial state of the garment, without accounting for the effects of animation.

Researchers have also made efforts to eliminate the dependency on huge volumes of ground truth data, by proposing a neural simulator trained by unsupervised learning [8]. In addition, there are studies that follow an unsupervised scheme and introduce the inertia loss term, inspired by physics-based simulation, to achieve dynamic clothing effects [13], [14]. For loose-fitting garments, recent work [11] generates virtual bones for dresses and infers the dynamic deformation of garments from motion sequences. However, they are difficult to generate different dynamic garments using a single deformation approximation model. More recently, HOOD [18] introduces a hierarchical-graph-based network that enables diverse garment deformation prediction. However, due to the complex design involving multiple sources of features such as vertices, edges, and garment-body relationship, achieving parallel processing of motion sequences remains a challenge with their proposal.

Another line of research focuses on clothing deformation in image space. To create realistic clothing deformation from scan data, DeepWrinkles [46] employs two complementary modules: one models the global shape using a linear subspace model, while the other enhances high-frequency details in normal maps via a generative adversarial network. Additional research [15] builds on this detail enhancement, framing it as a style transfer task. Though these methods yield photorealistic results, they may inaccurately capture some 3D shape details in certain areas and can be sensitive to lighting or viewpoint.

III. Methodology

At the core of our work is a learning-based approach that takes a body in motion and a garment in its initial state as inputs, and outputs a garment deformation with dynamic effects and individualized details. More specifically, given an arbitrary 3D garment mesh in the initial state M^0, a set of SMPL [44] bodies with shape parameters β and continuous poses $(\theta^{t-m}, ..., \theta^t)$ from the previous state $t-m$ to the current state t, our goal is to predict the deformed garment mesh with dynamic effects based on the state of the body and the properties of the clothing itself. Formally, we define our dynamic deformation estimator $\mathcal{W}$ as:

$$M^t = \mathcal{W}(M^0, \beta, (\theta^{t-m}, ..., \theta^t)). \tag{1}$$

Fig. 2 illustrates the complete process of clothing deformations. While the learning process of M^t is end-to-end, the deformation generation can be split into two stages. Firstly, we generate the global coarse deformation M_c^t by approximating dynamic weights and unposed blend shapes. Secondly, we generate M^t by further predicting the corrective blend shapes based on M_c^t. This two-stage strategy can effectively reduce the complexity of learning the nonlinear deformation task, as evidenced by prior studies [5], [6], [9], [12].

Given the aforementioned tasks, intuitively, it is crucial to accurately obtain and process various feature types. In the following sections, we start by introducing a frequency control strategy in feature processing of garment mesh graphs to enhance the quality of the high-frequency detail part (Sec. III-A). Next, to provide networks with valuable information to effectively distinguish between diverse garments, we propose a compact global descriptor based on the concept of spectrum, offering a comprehensive characterization of the garment shape (Sec. III-B). Then, we detail the dynamic deformation estimator, illustrating the information flow for the two-stage process (Sec. III-C). Lastly, we present a collision-handling method based on the neural signed distance field to further ensure the realism of the predicted garment (Sec. III-D).

A. Frequency Control Strategy

Due to the wide range of garments with varying topologies and vertex counts, it is essential for our network to be capable of handling these diverse inputs. To address this challenge, we adopt graph attention network (GAT) [47] and perform feature extraction using graph attention layers on arbitrary garment meshes. Let $V = [\boldsymbol{v}_1, ..., \boldsymbol{v}_N] \in \mathbb{R}^{d \times N}$ denote the graph features, where d is the dimension of features for each vertex, and N is the number of vertices. The process of handling features in graph attention layer f_{Att} can be formulated as:

$$f_{\mathrm{Att}}(V) = V \mathrm{softmax}(g(V))^\top, \tag{2}$$

where $g(V) : \mathbb{R}^{d \times N} \to \mathbb{R}^{N \times N}$ denotes a linear transformation used to calculate attention scores. Here, we execute the masked attention by calculating scores for nodes along with their neighborhood in the graph as in [47]. The $\mathrm{softmax}$ function is employed as a normalization operation.

Graph attention-based networks have emerged as state-of-the-art methods in a wide range of 3D data processing applications. However, as previously argued, we have observed that these networks, along with other general graph networks, face challenges in learning deformations due to spectral biases. To address this issue, we aim to enhance the network's ability to learn high-frequency components by adjusting its spectrum (Fig. 3). It has been demonstrated in image detail enhancement tasks that finding an upper bound on the Fourier coefficients of the learning layer can effectively control the Fourier spectrum of a convolutional network, yielding favorable results [48]. Inspired by this, the derived problem in this study becomes how to find the upper bound on the Fourier coefficients for the graph attention layer. To this end, according to the theory of harmonic analysis [49], we can enforce f_{Att} to be Lipschitz continuous such that its Fourier coefficients are constrained.

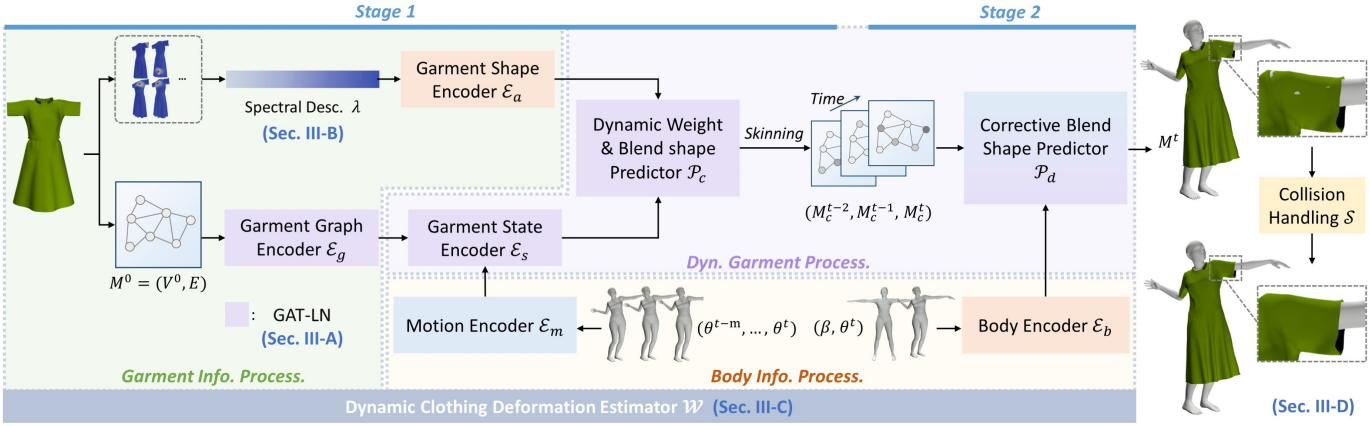


Fig. 2: Overview of our neural dynamic deformation method. The deformation estimation can be divided into two stages involving three processing units: garment information, body information, and dynamic garment processing. In the first stage, garment graph encoder and garment shape encoder extract the initial graph and global shape features of the garment mesh, respectively. Simultaneously, motion encoder and body encoder process body motion, shape, and pose features. The high-level graph features and motion features are utilized by the garment state encoder to generate garment state features, which combined with the garment global features, are used to predict dynamic weights and unposed blend shapes through the predictor, resulting in coarse-level garment deformation. In the second stage, subsequent graphs $(M_c^{t-2}, M_c^{t-1}, M_c^t)$ are processed by a corrective blend shape predictor and fused with body attributes to yield dynamic garment deformation. Finally, a neural collision handling method refines the predicted garments to produce the final garment deformation results.

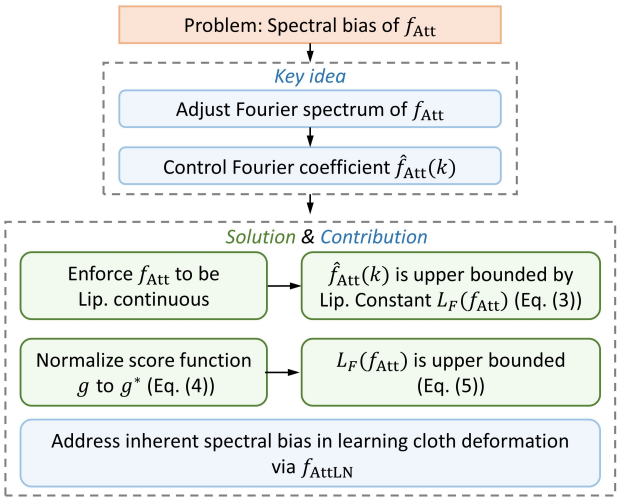


Fig. 3: Step-by-step procedure for addressing the spectral bias issue in graph attention layers from a Fourier spectrum perspective.

Specifically, if f_{Att} is Lipschitz continuous, there exists a constant L satisfies $\|f_{\text{Att}}(V_u) - f_{\text{Att}}(V_w)\|_{\text{F}} \leq L\|V_u - V_w\|_{\text{F}}$ for any input of $V_u, V_w \in \mathbb{R}^{d\times N}$, where $\|\cdot\|_{\text{F}}$ represents the Frobenius norm of the matrix. The minimum of such constant values is called the Lipschitz constant, denoted as $L_F(f_{\text{Att}})$. Consequently, the k-th Fourier coefficient $\hat{f}_{\text{Att}}(k)$ of the attention layer is bounded by:

$$|\hat{f}_{\text{Att}}(k)| \leq \frac{L_F(f_{\text{Att}})}{k^2}. \tag{3}$$

Here, without any control of the attention layer, its Lipschitz constant $L_F(f_{\text{Att}})$ is uncertain, which leads to no upper bound on the Fourier coefficient in Eq. (3). At this point, a further derived problem arises: determining the appropriate upper bound for the Lipschitz constant $L_F(f_{\text{Att}})$.

In the work by [50], the authors innovatively propose bounding $L_F(f_{\text{Att}})$ through the introduction of a normalized attention score function, and they provide a comprehensive derivation of this process. This pioneering work has been a direct influence on our approach. Our contribution lies in the adaptation of this method to a new domain - clothing deformation. Specifically, we normalize the score function $g(\cdot)$, defining it as g^*:

$$g^*(V) = \frac{\alpha g(V)}{\max\left\{\|g(V)\|_{(2,\infty)}, \|V^\top\|_{(\infty,2)} L_{F,(2,\infty)}(g)\right\}}, \tag{4}$$

where $\alpha \geq 0$ is the scale controller of scores; $\|\cdot\|_{(2,\infty)}$ and $\|\cdot\|_{(\infty,2)}$ respectively denote $(2,\infty)$-norm and $(\infty,2)$-norm for the matrix; $L_{F,(2,\infty)}(g)$ stands for the spectral norm of the parameters of g. The proof of Eq. (4) is detailed in [50]. With the designed normalization, $L_F(f_{\text{Att}})$ is Lipschitz continuous, and Eq. (3) can be expressed as:

$$|\hat{f}_{\text{Att}}(k)| \leq \pi \frac{L_F(f_{\text{Att}})}{k} \leq \frac{e^\alpha \sqrt{b/N} + \alpha\sqrt{8}}{k^2}. \tag{5}$$

As a result, the attention layer f_{Att} with g^* is Lipschitz continuous, and its Fourier coefficient $\hat{f}_{\text{Att}}(k)$ can be upper-bounded. This allows us to have control over the network's ability to learn high-frequency information through the parameter α. We also refer to the graph attention layer that has the Lipschitz-normalized score function g^* described above as the GAT-LN layer f_{AttLN}.

B. Clothing Spectral Descriptor

Existing dynamic garment deformation models [11], [13] are mostly limited to the specific garment, while static garment

Fig. 4: Color plots of the first six approximated eigenvectors of the long-sleeved dress mesh, where vertices in red represent high values distributed in areas such as the sleeves, shoulders, waist, and neckline. This suggests that the spectral description can reveal meaningful global information of the mesh shape.

deformation models that use graph neural networks [7], [45] attempt to learn across multiple garment types but often rely only on vertex positions, which overlooks important global information and leads to inaccurate predictions. Research has demonstrated that spectral shape analysis is effective in capturing mesh information [51]. In this work, we investigate the use of spectral descriptors to reveal meaningful global information about garments with varying shapes.

Given a garment mesh with N vertices, we construct an affinity matrix $A \in \mathbb{R}^{N\times N}$, where the i,j-th entry of A is the affinity between the i-th and the j-th vertices. Specifically, we use the geodesic distance calculated by Dijkstra's algorithm to define the affinity between two vertices. The defined affinity matrix has the advantage of incorporating intrinsic shape properties and being invariant to bending and rigid transformations, allowing for the effective learning of diverse garment deformations. Next, we perform spectral decomposition on the affinity matrix A. The eigenvectors of the affinity matrix form the normalized representation of the garment shape, while the eigenvalues specify how the shape varies along the axes. This allows us to consider the eigenvalues as the spectral descriptor.

In practice, garments often have thousands or tens of thousands of vertices, which can make the process of constructing and decomposing affinity matrices computationally expensive. To expedite this process, we apply Nyström approximation method [52] to efficiently approximate the eigenvalues of the original affinity matrix A. In particular, we perform furthest point sampling, a technique where we start by randomly selecting a vertex and then iteratively sample from the remaining vertices that are farthest from the set of already-sampled vertices until z vertices are selected ($z \ll N$). For these sampled vertices, we then calculate the affinity matrix $B \in \mathbb{R}^{z\times z}$. This smaller affinity matrix can be easily spectral-decomposed $B = U\Lambda U^\top$, allowing us to obtain an approximation of the eigenvectors $\hat{Q}$ of the original affinity matrix by using the Nyström method. Then, the corresponding affinity matrix can be approximated as $\hat{A} = \hat{Q}\Lambda\hat{Q}^\top$. Finally, we combine the eigenvalues into a vector $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_z]$, which we refer to as our clothing spectral descriptor. Note that the length z of the designed spectral descriptor is determined by the number of samples selected from the original garment shape and remains fixed. In contrast to the direct representation of clothing shapes by vertex coordinates, the proposed spectral descriptor offers a more compact and global representation of the garment shape. This enables the model to efficiently learn multiple types of garment deformations. To demonstrate a clearer visualization of the global information description, we show a color plot of the approximated eigenvectors $\hat{Q}$ (instead of the eigenvalues, which are difficult to represent visually) in Fig. 4.

C. Dynamic Clothing Deformation Estimator

Due to the multi-source nature of the deformation-related information, the estimator is divided into several units responsible for processing garment information, body information, and dynamic garment.

Garment information processing. We first construct a mesh graph for initial garment $M^0 = (V^0, E)$, where V^0 represents the vertex features and E denotes the mesh edges. For each vertex, we define its features as: $\boldsymbol{v}_i = [n_i, x_i, a_i]^\top$, which includes vertex normal $n_i \in \mathbb{R}^3$, position $x_i \in \mathbb{R}^3$, and the garment-body fit attribute $a_i \in \mathbb{R}^1$ [9]. These features capture important characteristics of the garment and describe the fit to the target body. To extract abstract garment information, we then pass this mesh graph through a garment graph encoder $\mathcal{E}_g$ to extract high-dimensional local graph features. On the other hand, we use spectral global description to analyze the garment as a whole, using the descriptor $\boldsymbol{\lambda}$. This descriptor is then fed into a garment shape encoder $\mathcal{E}_a$ to extract the comprehensive shape information. Notably, the information for each garment is processed just once at the rest pose.

Body information processing. The garment deformation is also closely influenced by body state. Therefore, we adopt a motion encoder $\mathcal{E}_m$ composed of LSTM to sequentially process the motion poses from time $t-m$ to t $(\theta^{t-m}, ..., \theta^t)$, where θ denotes the concatenation of axis-angle of each joint and the translation of the body relative to the preceding frame. In practice, we set m to 64 to include the relative overall motion information. Additionally, we apply a body encoder $\mathcal{E}_b$ to project the body shape β and current pose θ^t to the latent space, resulting in the latent body representation.

Dynamic garment processing. In the context of garment dynamic deformation, the state of the garment is influenced by both its inherent properties and the characteristics of the motion with which it interacts. To encode the dynamic garment state, we perform an element-wise multiplication between the processed graph features from the garment graph encoder $\mathcal{E}_g$ and the processed motion features from the motion encoder $\mathcal{E}_m$, and feed the product into the garment state encoder $\mathcal{E}_s$ to extract garment state features. Next, loose garments such as dresses require skinning weights that are strongly correlated with both the clothing shape and the motion. To achieve this, we use a dynamic weight predictor $\mathcal{P}_c$. This predictor performs multiplication between the garment shape features and state information, after which the results are normalized via softmax, thus generating garment and motion-related skinning weights W^t. The predictor also contains an unposed blend shape prediction part, which further processes the garment state features through GAT-LN layers and combines them with the motion feature to generate the unposed blend shape B^t. Subsequently, using the resulting skinning weights and unposed blend shapes, we apply linear blend

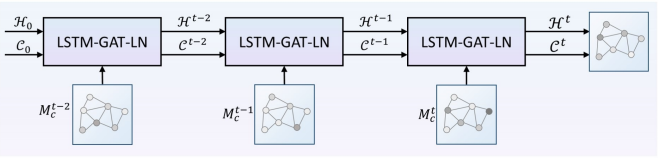


Fig. 5: Illustration of the inside of LSTM-GAT-LN. Sequences of graphs are transformed using the LSTM, while the state-to-state and input-to-state transformations follow the attention mechanism.

skinning to generate the intermediate deformation M_c^t. Note that while the body translation information is included in the body information processing, it has not been applied to the skinning step yet. The obtained M_c^t is order-dependent that incorporates underlying dynamics resulting from dynamic weights and unposed blend shapes.

To enhance the detailed folds of garments and ensure continuity, it is essential to process the time-series garment mesh data further. To do so, we construct a mesh graph $M_c^t = (V^t, E)$, where each vertex feature includes the current position of the vertex and its distance from all body skeletal joints. To process time-series graph data in a detail-aware manner, we design LSTM-GAT-LN as basis to construct corrective blend shape predictor $\mathcal{P}_d$. Specifically, LSTM-GAT-LN is employed to map a series of graphs into corrective blend shape. For efficiency, we use three consecutive graphs $(M_c^{t-2}, M_c^{t-1}, M_c^t)$. The LSTM-GAT-LN is defined as:

$$
\begin{aligned}
i^t &= \sigma\left(f_{\mathrm{AttLN}}(M_c^t) + f_{\mathrm{AttLN}}(\mathcal{H}^{t-1}) + w_i\mathcal{C}^{t-1}\right),\\
f^t &= \sigma\left(f_{\mathrm{AttLN}}(M_c^t) + f_{\mathrm{AttLN}}(\mathcal{H}^{t-1}) + w_f\mathcal{C}^{t-1}\right),\\
o^t &= \sigma\left(f_{\mathrm{AttLN}}(M_c^t) + f_{\mathrm{AttLN}}(\mathcal{H}^{t-1}) + w_o\mathcal{C}^{t-1}\right),\\
\mathcal{C}^t &= i_t \circ \tanh\left((f_{\mathrm{AttLN}}(M_c^t) + f_{\mathrm{AttLN}}(\mathcal{H}^{t-1})\right) + f^t \circ \mathcal{C}^{t-1},\\
\mathcal{H}^t &= o^t \circ \tanh\left(\mathcal{C}^t\right),
\end{aligned}
\tag{6}
$$

where i^t, f^t, o^t denote the input, forget, and output gate respectively; $\mathcal{C}^t, \mathcal{H}^t$ indicate the cell state and hidden state; w_i, w_f, w_o are the weights of the cell state; σ is the sigmoid function, and $\circ$ means the Hadamard product. The structure of LSTM-GAT-LN is shown in Fig. 5. It considers both the spatial correlation between structural graph nodes and the temporal correlation between sequential graphs, enabling efficient learning of high-level representations from complex clothing graph data. Next, to make the dynamic details fully reflect the attributes of the garment and the body, the processed graph after the LSTM-GAT-LN is multiplied with the latent body features from the body encoder $\mathcal{E}_b$ and forward into a two-layer GAT-LNs to generate the final detail correction. This correction is at the vertex level and will be added to the intermediate garment M_c^t. Finally, we also apply body translation to the result, thereby obtaining the dynamic garment deformation M^t.

To train the dynamic clothing deformation estimator, besides the position loss, we also employ consistency, gravity and collision loss terms [8] to help learn garment deformations close to the ground truth, while ensuring continuity and collision-free.

D. Collision Handling

With the dynamic garment deformation estimator, we are able to achieve realistic garment deformation effects that include continuous, detailed folds. However, we observed that collisions still occur in the predicted results. Despite the collision loss term provides some soft constraints on penetration between the garment-body pair during model optimization, it may not be effective in handling collisions for unseen data during inference. To address this issue, we propose a neural collision-handling method that can accurately detect and appropriately respond to garment collisions, correcting the garments to be collision-free while preserving realistic details.

Neural fields are capable of representing the physical properties of the object across space, and have been successfully applied in several 3D tasks [36], [53]. For the potential collision between our dynamic garments and bodies, we use the signed distance field (SDF) of the body to represent the penetration between the body and the garment:

$$
s(x_i) = \mathrm{sgn}(x_i, M_b) d(x_i, M_b), \tag{7}
$$

where $d(x_i, M_b) \in \mathbb{R}$ is the unsigned nearest distance from garment vertex x_i to body mesh M_b, and $\mathrm{sgn}(\cdot)$ is a sign function indicating whether the x_i is inside (positive) or outside (negative) of the body.

The function of Eq. (7) is non-analytic, and the computation for each vertex can be time-consuming. Therefore, we propose employing a neural model $\mathcal{S}$ to approximate the SDF for fast collision detection. Neural models have previously been employed with success in the reconstruction of 3D rigid [54] and articulated objects [55]. Building on this foundation, our approach repurposes a similar architecture to cater specifically to SMPL human body models, enabling the detection of collisions and facilitating the process of collision removal. Our model $\mathcal{S}$ comprises multiple fully-connected layers to process clothing vertices, along with an additional branch of fully-connected layers for handling body shape and pose features. To facilitate effective parameter optimization, we design a loss function that guides the learning process:

$$
\mathcal{L}_{\mathrm{SDF}} = \frac{1}{N}\sum_{i=1}^{N}\left(\|\mathcal{S}(x_i) - s(x_i)\| + \|\nabla_x \mathcal{S}(x_i) - n_i\|\right) + \mu_e \mathcal{L}_{\mathrm{E}}, \tag{8}
$$

$$
\mathcal{L}_{\mathrm{E}} = \frac{1}{N}\sum_{i=1}^{N}(\|\nabla_x \mathcal{S}(x_i)\| - 1)^2, \tag{9}
$$

where $\|\cdot\|$ is the L_2 norm. The loss function encourages the prediction $\mathcal{S}(x_i)$ to be similar to the ground truth value $s(x_i)$ and its gradients $\|\nabla_x \mathcal{S}(x_i)\|$ to be similar to the normal n_i. Also, we use an Eikonal term [56] $\mathcal{L}_{\mathrm{E}}$ that constrains the predicted gradient value $\|\nabla_x \mathcal{S}(x_i)\|$ to be close to 1. μ_e is the balancing weight and set as 0.15.

Once a garment vertex collision has been quickly detected by the model, we need to make reasonable adjustments to the position of the collision vertex. For the collision vertex

x_i with a positive value of $\mathcal{S}(x_i)$, the corrective displacement $\Delta(x_i) \in \mathbb{R}^3$ is defined as:

$$\Delta(x_i) = \frac{\nabla_{x_i}\mathcal{S}(x_i)}{\|\nabla_{x_i}\mathcal{S}(x_i)\|}\left(|\mathcal{S}(x_i)| + \delta_i\right), \tag{10}$$

where $\nabla_{x_i}\mathcal{S}(x_i)$ is normalized as the direction of displacement, and the magnitude is calculated as the sum of two terms. The first term of the magnitude is the SDF value $|\mathcal{S}(x_i)|$ (*i.e.*the collision vertex x_i is moved just to the body boundary) and the second term δ_i is a further detail correction based on the state of the non-collision neighboring vertices $\mathcal{N}_i$. Specifically, $\delta_i = \sum_{k\in\mathcal{N}_i} w_k|\mathcal{S}(x_k)|/|\mathcal{N}_i|$ indicates the weighted average of the absolute SDF value $|\mathcal{S}(x_k)|$ of the adjacent vertex $x_k \in \mathcal{N}_i$, where the weight w_k is determined by the Gaussian function applied to the distance between x_i and x_k, which is then normalized. Here, δ_i serves two purposes: first, it allows for partial avoidance of edge-to-edge collisions that may occur if only the vertices were moved to the surface of the body using the magnitude of $|\mathcal{S}(x_i)|$; second, it helps maintain local consistency near the corrected vertex, preventing excessive bulges.

IV. Implementation Details

Architecture. As described in Fig. 2, our dynamic deformation estimator $\mathcal{W}$ comprises three processing units: garment information processing, body information processing, and dynamic garment processing.

The garment information processing unit consists of a garment graph encoder $\mathcal{E}_g$ and a shape encoder $\mathcal{E}_a$. Specifically, $\mathcal{E}_g$ has two GAT-LN layers, with hidden feature sizes of [64, 128] and multi-head numbers of [24, 24]. The frequency control parameter α is set to 2, which is consistent across all subsequent GAT-LN layers. The garment graph features are generated by $\mathcal{E}_g$. The shape encoder $\mathcal{E}_a$, used for handling the spectral descriptor $\lambda = [\lambda_1, ..., \lambda_{256}]$, and comprises three fully connected layers of sizes [256, 128, 24]. The garment shape features are generated by $\mathcal{E}_a$. In our complete estimator $\mathcal{W}$, all f_{AttLN} layers utilize tanh, and all fully-connected layers employ ReLU as the activation function.

The body information processing unit is composed of a motion encoder $\mathcal{E}_m$ and a body encoder $\mathcal{E}_b$. Specifically, $\mathcal{E}_m$ incorporates a standard LSTM and outputs the motion feature with a size of 128. The body encoder $\mathcal{E}_b$, designed to handle body shape and current pose features, is structured with three fully-connected layers, possessing hidden sizes of [256, 128, 64].

The dynamic garment processing unit comprises a garment state encoder $\mathcal{E}_s$, a dynamic weight and blend shape predictor $\mathcal{P}_c$, and a corrective blend shape predictor $\mathcal{P}_d$. Firstly, the garment state encoder $\mathcal{E}_s$ has a f_{AttLN} layer with a hidden feature size of 128 and 24 multi-heads. The garment graph features and motion features are element-wisely multiplied before being fed into $\mathcal{E}_s$. Secondly, the dynamic weight and blend shape predictor $\mathcal{P}_c$ has two parts: 1) a dynamic weight predictor and 2) an unposed blend shape predictor. The dynamic weight predictor multiplies the garment state features with garment shape features, compresses the feature channel, and normalizes the result using the softmax function. The unposed blend shape predictor processes the garment state features through two f_{AttLN} layers with hidden feature sizes of [256, 128] and three heads, and then multiplies the result with the motion features to generate the unposed blend shape. After that, we utilize the linear blend skinning, which deform the garment according to the skeleton rotation, yielding the intermediate garment M_c^t. Lastly, a corrective blend shape predictor $\mathcal{P}_d$ is used to handle the temporal graphs. $\mathcal{P}_d$ consists of an LSTM-GAT-LN with hidden feature size of 64 and 8 heads, along with two single-head f_{AttLN} layers with hidden feature sizes of [64, 3]. The body features generated by $\mathcal{E}_b$ are multiplied with the outputs of the LSTM-GAT-LN. Following this, the multiplication result is forwarded into the two f_{AttLN} layers to generate the corrective displacements. These displacements are added to M_c^t, and translations are applied, resulting in the generation of dynamic detail garment M^t.

Additionally, we introduce a collision handling model $\mathcal{S}$ for post-processing. This model is designed with eight fully-connected layers (512 hidden feature sizes), split into two branches: three layers for garment vertex position, three for body shape and pose features, and two for feature fusion. Each branch incorporates a residual connection that links the output of the first layer to the layer preceding the fusion. For the activation function, we utilize the Softplus activation function with an internal parameter set to 100. During the training, the body translations are temporarily removed prior to collision handling and subsequently reintroduced to generate the final deformation.

Dataset. We obtained loose garments, including long t-shirts, dresses of varying lengths and sleeve styles, and jumpsuits, from the public CLOTH3D dataset [57], and draped them over SMPL bodies. To animate these characters wearing various garments, we collected motion sequences (*e.g.*, dancing, running, throwing, strutting, etc.) from the CMU Mocap dataset [58] at a frame rate of 30. For ease in training, the initial position of these characters is set to the origin. Then we utilized the Blender's cloth physics with the silk fabric to create the ground truth data. The training set consists of about 50 garments, nine bodies, and a total of 50,000 poses (roughly 250 motions). The test set includes 15 different garments, each draped over a body with a randomly selected shape, and collectively yielding a total of 7,500 poses (about 30 motions) across all garments. Importantly, there is no overlap between the training and test sets.

Training losses. To train the dynamic clothing deformation estimator, in addition to the mean squared error loss function of the vertex position, we also draw inspiration from other learning-based models like [8], and utilize the following loss functions for optimization:

$$\mathcal{L}_{\mathrm{vert}} = k_v\frac{1}{N}\sum_{i}^{N}\|X_i - X_i^{\mathrm{GT}}\|^2, \tag{11}$$

$$\mathcal{L}_{\mathrm{consistency}} = k_e\|E - E^{\mathrm{GT}}\|^2 + k_b\Delta(\mathrm{N})^2, \tag{12}$$

$$\mathcal{L}_{\mathrm{gravity}} = -k_g MXg, \tag{13}$$

$$\mathcal{L}_{\mathrm{collision}} = k_c\|\min\left(h(X) - \epsilon, 0\right)\|^2, \tag{14}$$

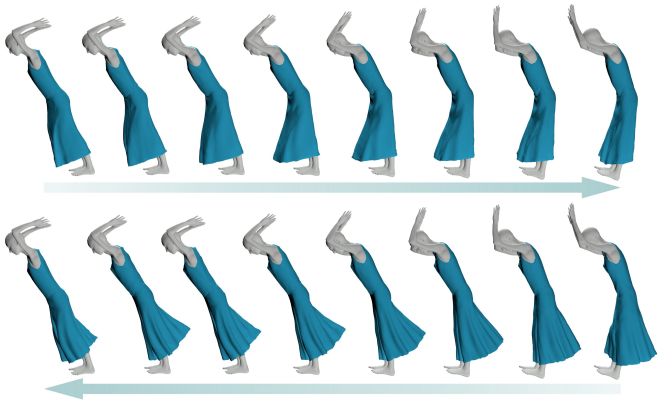


Fig. 6: Example of a forward and backward motion sequence of dressing swinging. Our method can successfully generate dynamic effects.

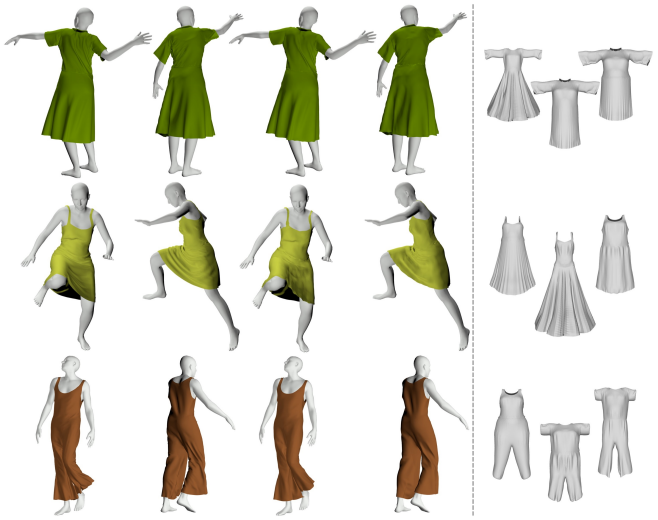


Fig. 7: Qualitative results of our method for various garments and motions.

where $\mathcal{L}_{\text{vert}}$ is the vert position loss and $X^{\text{GT}} \in \mathbb{R}^{N\times 3}$ stands for the ground truth vertex positions; $\mathcal{L}_{\text{consistency}}$ is the consistency loss and consists of two terms: an edge term a bending term, where $E \in \mathbb{R}^{N_E}$ (N_E: the number of edges) is the predicted edge lengths, E^{GT} is the edge lengths of the ground truth, and Δ represents the Laplace-Beltrami operator applied to the face normal $\mathrm{N} \in \mathbb{R}^{N_F\times 3}$ (N_F: the number of faces). $\mathcal{L}_{\text{gravity}}$ is the gravity loss, M is the vertices mass, and $g = (0, 0, -9.8)$ is the gravitational acceleration. $\mathcal{L}_{\text{collision}}$ is the collision loss, where $h(X)$ is the signed distance from the garment vertex to the nearest body vertex, and ϵ refers to the small positive value to prevent interpenetration. The hyperparameters of balancing weights are set as $k_v = 50$, $k_e = 10$, $k_b = 3$, $k_g = 0.2$, $k_c = 1.5$ for the optimization process.

Training implementation. We trained our dynamic deformation estimator $\mathcal{W}$ using two nVIDIA RTX A6000 GPUs with a batch size of 64. To optimize the training process, we used the Adam optimizer with an initial learning rate of 1e-3 and applied cosine annealing to decay the learning rate. We initialized the GAT-LN layers using the Glorot initialization and LSTM and fully connected layers using the Kaiming initialization [59]. Both of initializations are performed with their default settings. We made these choices based on their proven effectiveness in improving training convergence and performance. We first use only supervised loss for training, and switch to include all losses when the reduction in vertex distance error is less than 2% within 50 epochs. Our model was trained for roughly 5000 epochs and took around 5 days. To train our collision handling model, we randomly selected vertices from different bodies and clothes, and sampled the interior points of the body. The collision model weights are initialized with a geometric initialization from [60]. The setting of training of this model is [56].

V. Experiments

A. Results and Evaluation

Dynamic effect. In Fig. 6, given the dressing and swinging motions from the test set, our approach is capable of predicting the unseen garment's dynamic effect. The trend of body movement can be clearly observed by examining the positions of the head and arms. In the top row, the body undergoes a swing from a forward-leaning state to a slightly curved state, resulting in the front hemline of the garment transitioning from a hanging state to a position close to the calf. In contrast, in the bottom row, the body motion is from back to front, causing the back side of the dress to gradually lift up with the movement. Despite the same pose, our dynamic deformation estimator successfully predicts the distinct garment dynamics based on the previous states. More intuitive dynamic results are available in the supplemental video.

Realism. Our approach is characterized by a unified framework that enables dynamic deformation of diverse unseen garments. In Fig. 7, we evaluate the proposed method using three test garments. To demonstrate the method's generalization ability, we list the top three training data with the highest similarity to the corresponding test data on the right side of the figure. This similarity is calculated by computing the Frobenius norm of the affinity matrices between the test garment and all training garments in pairs. Notably, the training data and test data obviously differ in aspects like sleeve shapes, hemline lengths, trouser leg styles, and levels of looseness. Even when encountering new garments, the overall deformation predicted by our method appear realistic and maintain a high level of fidelity to the ground truth. The proposed method exhibits powerful generalization capabilities, negating the requirement for repetitive training across diverse clothing types. This enhancement in efficiency and applicability proves highly valuable in numerous scenarios requiring a large number of outfit variations.

B. Ablation Study

Frequency control strategy. To demonstrate the effectiveness of our proposed frequency control strategy, we investigate

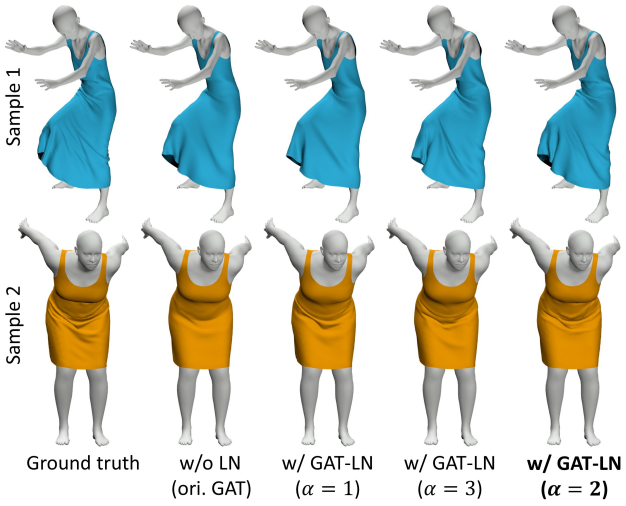


Fig. 8: Qualitative results of ablation on GAT-LN layers. Our experiments are conducted on two distinctly different garments, and we compare the ground truth deformation, approximated deformation without Lipschitz normalization, and with Lipschitz normalization controlled by different values of parameters.

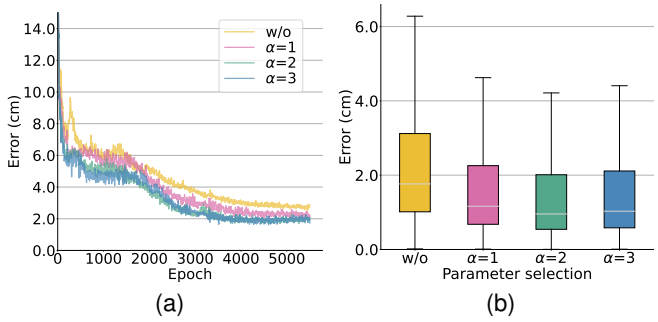


Fig. 9: Quantitative evaluation of ablation on the GAT-LN layers: (a) RMSE of the predicted deformation, (b) error distribution of predicted deformation.

the impact of GAT-LN layers f_{AttLN} on the deformation results of both loose and tight-fitting dresses. We approximate the garments' deformations using both the original graph attention network (GAT) without Lipschitz normalization and with Lipschitz normalization using different parameter values ($\alpha = 1$, 2, and 3). The qualitative results in Fig. 8 demonstrate that using GAT without Lipschitz normalization leads to a loss of high-frequency details in both garments. In contrast, applying f_{AttLN} with $\alpha = 2$ and 3 effectively generates more natural effects. Additionally, we found that the garment deformation around the belly position of sample 1 is better represented in the result for $\alpha = 2$. Furthermore, we conduct a quantitative evaluation of the Root Mean Square Error (RMSE) on test data with different settings (Fig. 9a). The absence of Lipschitz normalization leads to unstable performance and largest errors, whereas using f_{AttLN} with the parameter values of $\alpha = 2$ and 3 exhibits similar performance. According to Fig. 9b, $\alpha = 2$ yields slightly better with lower prediction error for

TABLE I: Quantitative results of ablation on the spectral descriptor. We conduct four experiments to assess the impact of the proposed spectral descriptor: without any global descriptor, applying graph pooling to acquire global information, substituting Euclidean distance for geodesic distance in the affinity matrix, and utilizing our original method.

	$\mathcal{E}_{dist}$ (cm)	$\mathcal{E}_{norm}$ (°)
w/o global desc.	2.37	9.02
w/ graph pooling	2.15	8.67
w/ Euclidean distance	2.24	8.95
w/ spectr. desc. (ours)	**1.90**	**7.53**

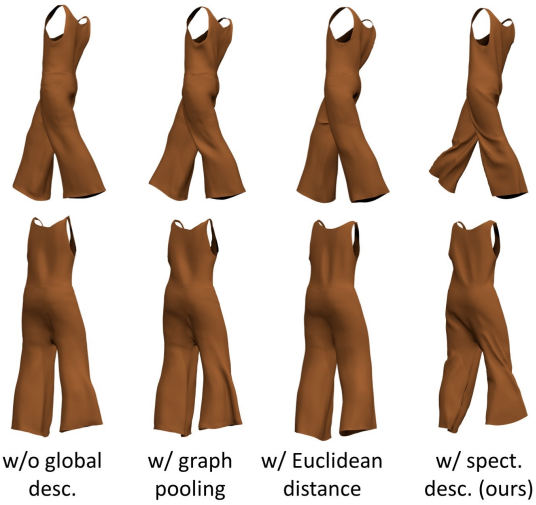


Fig. 10: Qualitative results of ablation on the spectral descriptor. Note the dynamic effects observed in the deformation trend of the trouser legs, along with the intricate details of the shoulder straps and back. Our method demonstrates superior performance compared to the other alternatives.

our data. Overall, the results provide evidence of the benefits of using the proposed GAT-LN layers and selecting suitable parameters to enhance the model's performance in generating realistic garment details.

Spectral descriptor. We report the effectiveness of the proposed spectral descriptor by using average vertex distance $\mathcal{E}_{dist}$ and average facet angular deviation $\mathcal{E}_{norm}$ between the predictions and the ground truth. In Tab. I, we first remove the spectral descriptor and do not use any global representations in the network. This absence of global information hinders the model's ability to generalize to new garments, resulting in the largest error. Next, following a similar manner as in [7], we use a fully-connected layer and a max-pooling layer after graph encoder $\mathcal{E}_g$ to obtain the global features, and then concatenate them with local features from $\mathcal{E}_g$. This implementation leads to improved accuracy compared to the case without global representation. This improvement is also expected, given that the inclusion of global information enables the model to perform effectively with new garments. Next, we maintain our proposed network structure while replacing the geodesic distance of the affinity matrix A with the Euclidean distance. However, this modification doesn't lead to the global description being invariant or robust to bending, resulting

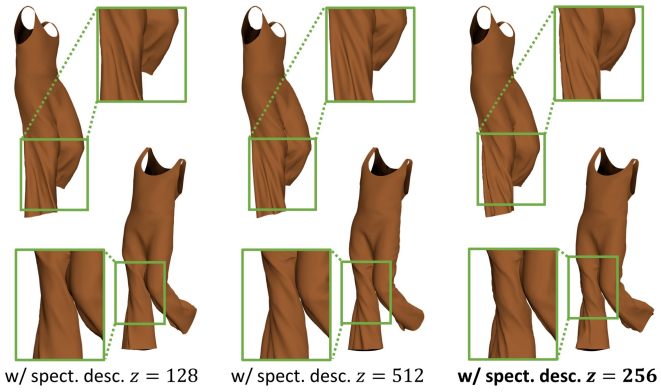


Fig. 11: Qualitative evaluation of ablation on the spectral descriptor with different dimensionality. The trouser leg area is zoomed in on. With a dimension of 128, some details are lost. In contrast, dimensions of 256 and 512 better encapsulate global information, leading to richer fold patterns. In our study, we opt for a dimension of 256 for the spectral descriptor.

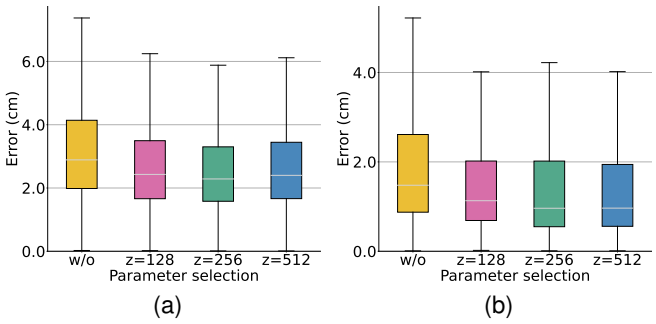


Fig. 12: Quantitative evaluation of ablation on the spectral descriptor: (a) prediction error of a challenging sample (*i.e.*, jumpsuit), (b) mean error of all test samples.

in minimal improvement in prediction accuracy. Finally, we show the results using our proposed spectral descriptor $\boldsymbol{\lambda}$, which yields the highest prediction accuracy. This achievement can be attributed to the fact that our descriptor effectively captures essential garment shape features, providing valuable global prior knowledge to the model. In Fig. 10, we present the qualitative results on a challenging garment, a wide-legged jumpsuit. This jumpsuit has a large shape difference from garments in the training set. Notably, the approach without a global description and utilizing a replacement global representation exhibits shortcomings in regions such as the back, trouser leg, and shoulder strap (second row). In contrast, the utilization of our spectral descriptor effectively captures the dynamic swinging effect of the loose trouser leg during forward steps, along with intricate wrinkles and folds, resulting in a more successful inference.

We further validate the effect of the dimensionality z of the spectral descriptor $\boldsymbol{\lambda} = [\lambda_1, ..., \lambda_z]$ on the results . We explore the impact of different descriptor vector lengths where we set z to 128, 256, and 512, and present qualitative results in Fig. 11. We found that the results using $z = 256, 512$ demonstrate stronger shape representation capabilities, leading to visually

TABLE II: Average ratio (%) of garment-body collision. We measure three types of garments with new motion sequences in the test set.

	collision loss only	SSCH [19]	w/ collision handling
T-shirt	0.97	0.10	0.12
Dress	0.61	0.07	0.07
Jumpsuit	1.18	0.14	0.15

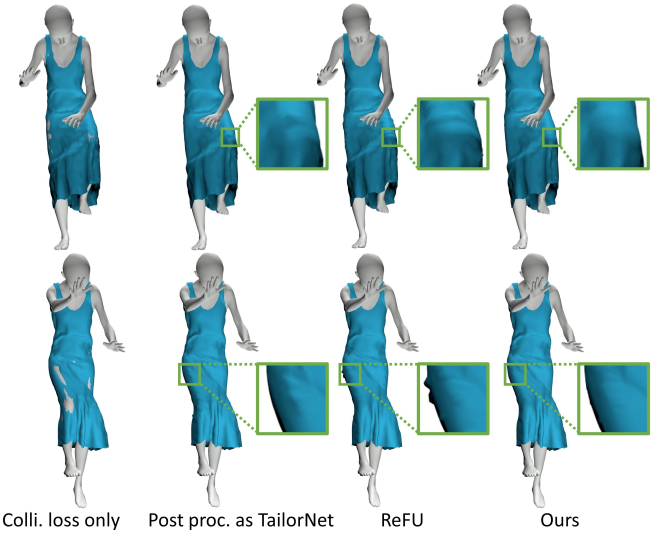


Fig. 13: Qualitative evaluation of collision handling. We compare our approach with the most common-used collision loss strategies, as well as the post-processing in TailorNet [6] and ReFU [61]. Our method effectively removes collisions while the garment surface is more natural and free of bulge artifacts.

pleasing outputs with finer details. Fig. 12a displays the quantitative results, confirming that the proposed spectral descriptor enhances the performance of the model. Specifically, the use of the spectral descriptor significantly reduces the deformation error, with $z = 256$ resulting in the lowest error. Fig. 12b presents the mean error of all test data, which is approximately 1cm lower than the error of the challenging jumpsuit overall. The results provide support for the effectiveness of the spectral descriptor in accurately deforming dynamic garments.

Collision handling. To quantify the improvement in collision handling, we provide an evaluation conducted on three types of garments from the test set. These garments are subjected to new motion sequences including walking, jumping, and climbing. In Tab. II, we report the average percentage of garment vertices inside the body. As observed, when dealing with previously unseen garments and motions, a certain percentage of collided vertices are still present in the results without post-processing, *i.e.*, using collision loss only. Our collision handling neural model detects and corrects these collided vertices using the approximate SDF, effectively reducing the collision rate by around 88%. This improvement is achieved with high efficiency, requiring only around 0.2ms per frame. As a result, clothing animation approaches a nearly collision-free state that is visually imperceptible to the naked

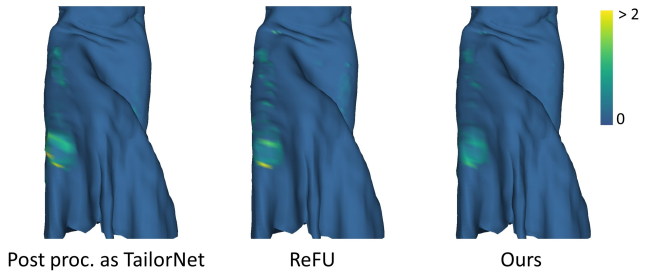


Fig. 14: Comparison of local mesh consistency after collision removal using various methods. A Laplacian-based consistency measure is employed, with the original mesh before the collision removal serving as the reference baseline. Correspondingly, a value of 0 signifies the non-collision (and thus uncorrected) state. Regions with degraded consistency (such as bulges) in the corrected outcomes are highlighted in yellow, indicating higher values.

TABLE III: Quantitative evaluation of mesh local consistency following three post-processing methods. The average Laplacian-based consistency deviation is reported for collision removal vertices across three different garment types.

	TailorNet [6]	ReFu [61]	Ours
T-shirt	0.56	0.47	0.34
Dress	0.30	0.28	0.21
Jumpsuit	0.63	0.55	0.42

eye. Moreover, it becomes evident that garments with more relaxed fits, such as dress, tend to exhibit lower collision rates in dynamic poses. Additionally, we compare ours with the state-of-the-art learning-based collision handling method SSCH [19]. SSCH offers an end-to-end solution for garment deformation that manages garments in both pose space and canonical space to be collision-free. The results show that this method is highly effective in avoiding a substantial portion of collisions. In comparison, our method has two distinct advantages. Firstly, as we directly predict the SDF, this format can be more readily adapt to scenarios involving multiple garments with a single model. Secondly, based on our empirical observations, we have found that many datasets may contain problematic data with some degree of intra-body penetration, *e.g.*, a hand partially penetrating the body during a hand-crossing movement. In such cases, using SDF for post-processing leads to more stable performance during model optimization than the end-to-end method of learning garment deformations directly. Nonetheless, for most collision issues, both ours and SSCH can effectively address them.

Next, we report the effect of our proposed collision handling solution in Fig 13. As depicted, while penalizing the penetration of clothing and body imposes a certain degree of constraint on model learning, it cannot ensure a collision-free state in some extreme test poses. Based on the collision scenario, we implement a post-processing step in TailorNet [6] to efficiently resolve collisions. However, this step results in undesirable bulge artifacts, as indicated by the enlarged area. The reason behind these artifacts is that the method crudely employs a fixed distance larger than the penetration depth to remove the vertices of garments that penetrate the body's interior without considering the positions of surrounding vertices. Then, we experiment with the ReFU approach [61] to displace the penetration vertices away from the body, guided by an estimated scale. However, since the scale estimation does not yet incorporate the states of neighboring vertices, the outcome still exhibits issues like "pump out". At the rightmost, we present the result produced by our proposed method. By incorporating adaptive adjustments for the collided vertices, the corrected mesh surface appears more natural and achieves higher visual quality. Next, for a quantitative evaluation of collision processing quality, we introduce a Laplacian-based metric to quantify the degree of local consistency in the region around the collision resolved vertices. Specifically, for the mesh before and after the collision removal, we compute their Laplacian matrices $L^C, L \in \mathbb{R}^{N\times N}$ with the half-cotangent weighting function. Then, we define the local consistency deviations as $d_{cons} = |\mathrm{diag}(L) - \mathrm{diag}(L^C)|$, with lower values indicating higher consistency. As shown in Fig. 14, we compare the local consistency between ours and two other post-processing methods. For those collision-resolved vertices, the average consistency deviation are shown in Tab. III. This result further validates that our collision handling method generates high quality processing details. However, it is also important to recognize that the post-processing techniques employed by TailorNet generally have the advantage of higher reliability.

C. Comparisons

Method capacity. We conduct a comparison of our method with recent learning-based approaches for garment deformation, including SNUG [14], Neural Cloth Simulation (NCS) [13], and HOOD [18]. A comprehensive analysis of deformation characteristics, model generalization, batch support, and collision situation is presented in Table IV. Our method stands out by achieving dynamic clothing deformations for arbitrary garments, irrespective of their topologies and vertex counts, as well as different body shapes. This feature enhances the applicability of our approach in scenarios demanding diverse garment types, like virtual try-on. Furthermore, our method is flexible in terms of training and execution batch sizes, leading to enhanced efficiency in both training and running processes. Regarding the collision situation, we define an imperceptible as one where the collision rate between the garment and the body remains under 0.15%. In contrast to other methods where approximately 0.28% of clothing vertices intersect with the body, our approach excels in producing collision-free garment animations that are virtually imperceptible to the naked eye.

Qualitative evaluation. We also present qualitative results from different methods in Fig. 15. For implementation, we base our comparisons on publicly available code. It is worth mentioning our approach and HOOD achieve their results using a single trained model, whereas other methods necessitate separate training for each individual garment. While SUNG only released the model inference codes, we reconstructed the entire model by leveraging their codes and the details provided in the paper. The technique described in SUNG, borrowing

Fig. 15: Comparison with state-of-the-art neural dynamic clothing deformation methods. Rows from top to bottom : ground truth, SNUG [14], Neural Cloth Simulation (NCS) [13], HOOD [18], and our prediction. Columns represent deformation with various garments and motions: (a)-(b) moving forward and hanging leg raise, (c)-(e) palm striking and kicking, (f)-(h) layup.

SMPL skinning weights from the closest body vertex in rest pose, is ineffective for loose garments. We thus implemented a modified version by sampling and averaging the body weights of surrounding vertices for the garment weights. While there are minor discrepancies in the architecture and training details, they do not influence the qualitative results and the key idea in the original paper. For NCS, following their description and empirical evidence, we reimplemented their method by gradually increasing the impact of collision and inertial force loss from about 1/10 to 1 unit, in line with stretch and shear conditions. Similarly, we replace their original skinning weights by adopting the same method as for SNUG, which involves randomly sampling and averaging body skinning weights. These improve the dynamics of complex garments and enhances convergence stability. We also reimplemented HOOD in accordance with their description and original code.

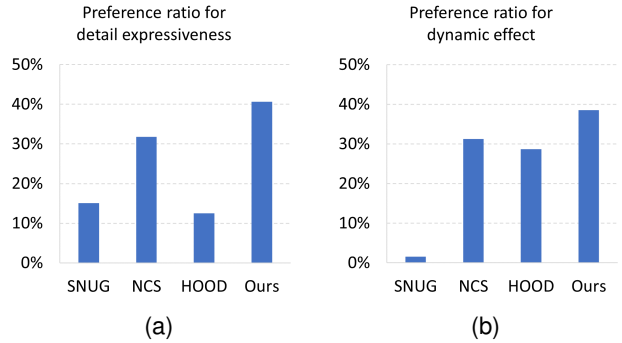


Fig. 16: Perceptual study. The preference ratio for (a) detail expressiveness and (b) dynamic effect.

As observed in the results, for motions with low dynamics: forward movement (a) and hanging leg lifts (b), prior methods can yield roughly reasonable deformations, but some finer details are lacking in HOOD's prediction. In contrast, our method demonstrates its capability to predict a substantial portion of folds and dynamic trends, primarily due to the effectiveness of our proposed spectral strategies. In scenarios where movements involve actions such as the character's palm strikes and body twisting in cases (c) and (d), we can observe a continuous swinging motion of the dress hem from left to right. However, SNUG does not adequately capture this dynamic behavior, particularly in the fold direction of the dress. This deficiency in capturing dynamics can be attributed to the inherent non-dynamic nature of SNUG, as thoroughly analyzed in [13]. This static behavior also persists in the test samples (e)-(h), where the dress fails to exhibit natural movement in sync with the character's motions. Conversely, both NCS and HOOD demonstrate the ability to predict a majority of clothing dynamics, with the resulting deformations aligning well with the character's movements. Nevertheless, HOOD does exhibit a limitation in accurately predicting the hemline deformation during the leg lift movement (e). Additionally, all methods, except for ours, struggle with generating detailed deformations, particularly noticeable in the shoulder region of motion (c). Regarding the layup motion (f)-(h), our method stands out among other learning-based approaches by producing highly plausible and dynamic garment deformations. Unlike methods that utilize unsupervised schemes, our approach employs supervised learning, allowing us to create a model capable of accurately predicting specific deformation patterns based on input conditions. While supervised learning involves initial effort in preparing ground truth data, it offers targeted predictions that are particularly valuable for precise control over the training process. Additionally, our approach allows for more direct comparisons with ground truth, facilitating a clearer evaluation of prediction accuracy compared to unsupervised methods, which can be more challenging to assess. Furthermore, our supervised model benefits from transfer learning, enabling easy adaptation to different garment types through direct application or fine-tuning. This efficiency stands in contrast to the repetitive training required by methods like SNUG and NCS.

TABLE IV: Comparison with state-of-the-art methods in terms of deformation dynamics, model generalization, batch support, and collision situation.

	Dyn-amics	Genera-lization	Batch support	Imperceptible collision	Learning scheme
SNUG	✓	✗	✓	✗	unsupervised
NCS	✓	✗	✓	✗	unsupervised
HOOD	✓	✓	✗	✗	unsupervised
Ours	✓	✓	✓	✓	Supervised

TABLE V: Timing performance comparison with state-of-the-art graph learning-based methods.

	Dynamic	Speed	Speed (batch)
GarNet++	✗	1.2 fps	16.5 fps
FitGar	✗	47.6 fps	380.9 fps
HOOD	✓	13.4 fps	13.4 fps
Ours	✓	38.1 fps	307.8 fps

User study. Ultimately, users' perceptions of garment deformation rely heavily on visual satisfaction, rendering the need for ground truth as a reference for fairness unnecessary. Thus, we compare these methods with a focus on user preferences, aiming to uncover potential differences in perceptual clothing animation. In this perceptual experiment, our goal is to assess which prediction displays more detailed expressiveness and dynamic effect. We recruited 32 participants to watch videos featuring garment animations from four methods. Among the participants, 18 had experience in computer animation research, while 14 were newcomers to the field. To ensure fairness, the order in which videos from different methods were shown was randomized. Participants were then asked to rate the four methods, allocating scores of 3, 2, 1, or 0 based on their preferences for deformation detail and dynamics after viewing the videos. As shown in Fig. 16, our method achieves the highest preference ratios in both visual quality metrics: deformation detail and dynamics. This demonstrates the advantages of our approach over state-of-the-art methods in terms of visual perception.

D. Model Performance

Our model can infer one frame with a garment consisting of about 10K vertices and 20K faces in approximately 26 ms, achieving a frame rate of 38 fps on a computer equipped with an Intel Core i9-13900K CPU and an nVIDIA GeForce RTX 4090 GPU. By running batched frames with a batch size of 10, the model can achieve a frame rate of about 300 fps. Moreover, as the video memory capacity expands, our model can process larger batches of data in parallel, thereby effortlessly achieving frame rates well beyond real-time requirements. In general, the inference speed of neural network models increases linearly with increased video memory, requiring only minimal modifications to existing code.

In Tab. V, we also provide performance comparisons with other learning-based clothing deformation methods, which include GarNet++ [4], FitGar [9], and HOOD [18]. All

these methods utilize graph neural networks for inferring deformations. While GarNet++ and FitNet are designed for static deformations, the advanced HOOD method is capable of generating dynamic clothing effects. However, HOOD's inference speed is restricted due to its limitation of batch size 1. In contrast, our approach showcases exceptional efficiency, achieving an inference speed approximately 18 times faster than HOOD. On the other hand, when compared to the garment-specific design of SNUG and NCS, our method (alongside the other methods detailed in Tab. V) necessitates a more complex graph network structure to satisfy wider-ranging generalization needs. It's worth noting that while SNUG and NCS may have superior efficiency, our approach still stands out by being approximately 25 times faster than physics-based simulation techniques such as those employed in cloth simulation tools [62].

VI. Conclusion and Future Work

We have presented a novel approach to efficiently estimate dynamic garment deformation using a unified model. Our work can learn the dynamic behavior of arbitrary garments under consistent body motion, without special constraints on garment topology, vertex count, and degree of fitness. We achieve this by introducing the frequency-control strategies for the deformation network and a spectral global descriptor for diverse garment representation. These techniques enable our deformation estimator to generate personalized and vivid details. We believe that this spectral technique also has the potential to address issues that have been encountered in other animation areas when utilizing graph neural networks. In addition, we propose a neural collision handling method that automatically detects and corrects penetrations between garment-body pairs, resulting in more realistic and natural-looking results.

There are still a few weaknesses that need to be addressed for further improvement. First, while we were able to mitigate the over-smoothing effect by applying a frequency control strategy to graph attention layers, we did not extend this strategy to other layers in the network, which may have resulted in the spectral bias problem not being fully addressed. In the future, exploring spectral control for the entire network is a promising direction. Second, our current collision handling method relies on the estimated SDF of the human body for detecting and correcting garment penetrations. While we have validated its effectiveness in handling SMPL bodies that were seen during training, the network may not be able to accurately estimate subtle collisions of unseen bodies. In the future, the problem of garment-body collisions can be further explored based on implicit surface studies such as NASA [63], DeepSDF [54], and A-SDF [55]. Notably, A-SDF shows promise due to the ability to handle articulation variations and the generalization capacity for unseen joint angles. This capability suggests the potential for A-SDF to serve as an innovative loss function in the future, for addressing self-collision challenges in garment simulations.

Acknowledgments

This work has been partially supported by Beijing Natural Science Foundation 4232017 and JSPS KAKENHI 22K12331.

References

[1] J. Li, G. Daviet, R. Narain, F. Bertails-Descoubes, M. Overby, G. E. Brown, and L. Boissieux, "An implicit frictional contact solver for adaptive cloth simulation," *ACM Trans. Graph.*, vol. 37, no. 4, jul 2018. [Online]. Available: https://doi.org/10.1145/3197517.3201308

[2] M. Li, D. M. Kaufman, and C. Jiang, "Codimensional incremental potential contact," *ACM Trans. Graph.*, vol. 40, no. 4, jul 2021. [Online]. Available: https://doi.org/10.1145/3450626.3459767

[3] E. Gundogdu, V. Constantin, A. Seifoddini, M. Dang, M. Salzmann, and P. Fua, "GarNet: A two-stream network for fast and accurate 3D cloth draping," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8738–8747.

[4] E. Gundogdu, V. Constantin, S. Parashar, A. Seifoddini, M. Dang, M. Salzmann, and P. Fua, "GarNet++: Improving fast and accurate static 3D cloth draping by curvature loss," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 181–195, 2022.

[5] R. Vidaurre, I. Santesteban, E. Garces, and D. Casas, "Fully convolutional graph neural networks for parametric virtual try-on," *Comput. Graph. Forum*, vol. 39, no. 8, pp. 145–156, 2020.

[6] C. Patel, Z. Liao, and G. Pons-Moll, "TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7363–7373.

[7] H. Bertiche, M. Madadi, E. Tylson, and S. Escalera, "DeePSD: Automatic deep skinning and pose space deformation for 3D garment animation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 5451–5460.

[8] H. Bertiche, M. Madadi, and S. Escalera, "PBNS: Physically based neural simulation for unsupervised garment pose space deformation," *ACM Trans. Graph.*, vol. 40, no. 6, dec 2021. [Online]. Available: https://doi.org/10.1145/3478513.3480479

[9] T. Li, R. Shi, and T. Kanai, "Detail-aware deep clothing animations infused with multi-source attributes," *Comput. Graph. Forum*, vol. 42, no. 1, pp. 231–244, 2023.

[10] L. De Luigi, R. Li, B. Guillard, M. Salzmann, and P. Fua, "DrapeNet: Garment generation and self-supervised draping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023.

[11] X. Pan, J. Mai, X. Jiang, D. Tang, J. Li, T. Shao, K. Zhou, X. Jin, and D. Manocha, "Predicting loose-fitting garment deformations using bone-driven motion networks," in *ACM SIGGRAPH*, 2022. [Online]. Available: https://doi.org/10.1145/3528233.3530709

[12] I. Santesteban, M. A. Otaduy, and D. Casas, "Learning-based animation of clothing for virtual try-on," *Comput. Graph. Forum*, vol. 38, no. 2, pp. 355–366, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13643

[13] H. Bertiche, M. Madadi, and S. Escalera, "Neural cloth simulation," *ACM Trans. Graph.*, vol. 41, no. 6, 2022. [Online]. Available: https://doi.org/10.1145/3550454.3555491

[14] I. Santesteban, M. A. Otaduy, and D. Casas, "SNUG: Self-supervised neural dynamic garments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8130–8140.

[15] M. Zhang, T. Wang, D. Ceylan, and N. J. Mitra, "Deep detail enhancement for any garment," *Comput. Graph. Forum*, vol. 40, no. 2, pp. 399–411, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.142642

[16] M. Zhang, T. Y. Wang, D. Ceylan, and N. J. Mitra, "Dynamic neural garments," *ACM Trans. Graph.*, vol. 40, no. 6, 2021. [Online]. Available: https://doi.org/10.1145/3478513.3480497

[17] M. Zhang, D. Ceylan, and N. J. Mitra, "Motion guided deep dynamic 3D garments," *ACM Trans. Graph.*, vol. 41, no. 6, 2022. [Online]. Available: https://doi.org/10.1145/3550454.3555485

[18] A. Grigorev, B. Thomaszewski, M. J. Black, and O. Hilliges, "HOOD: Hierarchical graphs for generalized modelling of clothing dynamics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 16 965–16 974. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01627

[19] I. Santesteban, N. Thuerey, M. A. Otaduy, and D. Casas, "Self-supervised collision handling via generative 3D garment models for virtual try-on," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11 758–11 768.

[20] J. Weil, "The synthesis of cloth objects," in *ACM SIGGRAPH*, 1986, p. 49–54. [Online]. Available: https://doi.org/10.1145/15922.15891

[21] D. Baraff and A. Witkin, "Large steps in cloth simulation," in *ACM SIGGRAPH*, 1998, p. 43–54. [Online]. Available: https://doi.org/10.1145/280814.280821

[22] T. Liu, A. W. Bargteil, J. F. O'Brien, and L. Kavan, "Fast simulation of mass-spring systems," *ACM Trans. Graph.*, vol. 32, no. 6, 2013. [Online]. Available: https://doi.org/10.1145/2508363.2508406

[23] M. Müller, B. Heidelberger, M. Hennix, and J. Ratcliff, "Position based dynamics," *J. Vis. Comun. Image Represent.*, vol. 18, no. 2, p. 109–118, apr 2007. [Online]. Available: https://doi.org/10.1016/j.jvcir.2007.01.005

[24] M. Macklin, M. Müller, and N. Chentanez, "XPBD: Position-based simulation of compliant constrained dynamics," in *Proc. Int. Conf. Motion Games*, 2016, pp. 49–54. [Online]. Available: https://doi.org/10.1145/2994258.2994272

[25] C. Jiang, T. Gast, and J. Teran, "Anisotropic elastoplasticity for cloth, knit and hair frictional contact," *ACM Trans. Graph.*, vol. 36, no. 4, jul 2017. [Online]. Available: https://doi.org/10.1145/3072959.3073623

[26] H. Wang, "GPU-based simulation of cloth wrinkles at submillimeter levels," *ACM Trans. Graph.*, vol. 40, no. 4, jul 2021. [Online]. Available: https://doi.org/10.1145/3450626.3459787

[27] Z. Chen, H.-Y. Chen, D. M. Kaufman, M. Skouras, and E. Vouga, "Fine wrinkling on coarsely meshed thin shells," *ACM Trans. Graph.*, vol. 40, no. 5, 2021. [Online]. Available: https://doi.org/10.1145/3462758

[28] M. Tang, T. Wang, Z. Liu, R. Tong, and D. Manocha, "I-Cloth: Incremental collision handling for gpu-based interactive cloth simulation," *ACM Trans. Graph.*, vol. 37, no. 6, 2018. [Online]. Available: https://doi.org/10.1145/3272127.3275005

[29] C. Li, M. Tang, R. Tong, M. Cai, J. Zhao, and D. Manocha, "P-Cloth: Interactive complex cloth simulation on multi-gpu systems using dynamic matrix assembly and pipelined implicit integrators," *ACM Trans. Graph.*, vol. 39, no. 6, nov 2020. [Online]. Available: https://doi.org/10.1145/3414685.3417763

[30] M. Ly, J. Jouve, L. Boissieux, and F. Bertails-Descoubes, "Projective dynamics with dry frictional contact," *ACM Trans. Graph.*, vol. 39, no. 4, 2020. [Online]. Available: https://doi.org/10.1145/3386569.3392396

[31] J. Liang, M. Lin, and V. Koltun, "Differentiable cloth simulation for inverse problems," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[32] T. Y. Wang, T. Shao, K. Fu, and N. J. Mitra, "Learning an intrinsic garment space for interactive authoring of garment animation," *ACM Trans. Graph.*, vol. 38, no. 6, 2019. [Online]. Available: https://doi.org/10.1145/3355089.3356512

[33] Q. Ma, J. Yang, A. Ranjan, S. Pujades, G. Pons-Moll, S. Tang, and M. J. Black, "Learning to dress 3D people in generative clothing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6468–6477. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00650

[34] M. Mihajlovic, Y. Zhang, M. J. Black, and S. Tang, "LEAP: Learning articulated occupancy of people," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10 456–10 466.

[35] G. Tiwari, B. L. Bhatnagar, T. Tung, and G. Pons-Moll, "SIZER: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing," in *Proc. Eur. Conf. Comput. Vis.*, vol. 12348, 2020, pp. 1–18. [Online]. Available: https://doi.org/10.1007/978-3-030-58580-8_1

[36] I. Santesteban, M. Otaduy, N. Thuerey, and D. Casas, "ULNeF: Untangled layered neural fields for mix-and-match virtual try-on," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 12 110–12 125.

[37] C. Rodriguez-Pardo, M. Prieto-Martin, D. Casas, and E. Garces, "How will it drape like? capturing fabric mechanics from depth images," *Comput. Graph. Forum*, vol. 42, no. 2, pp. 149–160, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14750

[38] A. Casado-Elvira, M. C. Trinidad, and D. Casas, "PERGAMO: Personalized 3D garments from monocular video," *Comput. Graph. Forum*, vol. 41, no. 8, pp. 293–304, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14644

[39] L. Liu, Y. Zheng, D. Tang, Y. Yuan, C. Fan, and K. Zhou, "NeuroSkinning: Automatic skin binding for production characters with deep graph networks," *ACM Trans. Graph.*, vol. 38, no. 4, 2019. [Online]. Available: https://doi.org/10.1145/3306346.3322969

[40] Z. Xu, Y. Zhou, E. Kalogerakis, C. Landreth, and K. Singh, "RigNet: Neural rigging for articulated characters," *ACM Trans. Graph.*, vol. 39, no. 4, 2020. [Online]. Available: https://doi.org/10.1145/3386569.3392379

[41] T. Li, R. Shi, and T. Kanai, "DenseGATs: A graph-attention-based network for nonlinear character deformation," in *Proc. Symp. Interactive 3D Graph. Games*, 2020, pp. 5:1–5:9.

[42] P. Li, K. Aberman, R. Hanocka, L. Liu, O. Sorkine-Hornung, and B. Chen, "Learning skeletal articulations with neural blend shapes," *ACM Trans. Graph.*, vol. 40, no. 4, jul 2021. [Online]. Available: https://doi.org/10.1145/3450626.3459852

[43] T. Li, R. Shi, and T. Kanai, "MultiResGNet: Approximating nonlinear deformation via multi-resolution graphs," *Comput. Graph. Forum*, vol. 40, no. 2, pp. 537–548, 2021.

[44] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, 2015. [Online]. Available: https://doi.org/10.1145/2816795.2818013

[45] Y. Li, M. Tang, Y. bo Yang, Z. Huang, R. Tong, S. Yang, Y. Li, and D. Manocha, "N-Cloth: Predicting 3D cloth deformation with mesh-based networks," *Comput. Graph. Forum*, vol. 41, no. 2, pp. 547–558, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14493

[46] Z. Lähner, D. Cremers, and T. Tung, "Deepwrinkles: Accurate and realistic clothing modeling," in *Proc. Eur. Conf. Comput. Vis.*, 2018. [Online]. Available: https://doi.org/10.1007/978-3-030-01225-0_41

[47] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ

[48] Z. Shi, P. Mettes, S. Maji, and C. G. M. Snoek, "On measuring and controlling the spectral bias of the deep image prior," *Int. J. Comput. Vis.*, vol. 130, no. 4, pp. 885–908, 2022. [Online]. Available: https://doi.org/10.1007/s11263-021-01572-7

[49] Y. Katznelson, *An introduction to harmonic analysis*, 3rd ed. Cambridge University Press, 2004.

[50] G. Dasoulas, K. Scaman, and A. Virmaux, "Lipschitz normalization for self-attention layers with application to graph neural networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 2456–2466. [Online]. Available: http://proceedings.mlr.press/v139/dasoulas21a.html

[51] D. Smirnov and J. Solomon, "Hodgenet: Learning spectral geometry on triangle meshes," *ACM Trans. Graph.*, vol. 40, no. 4, 2021. [Online]. Available: https://doi.org/10.1145/3450626.3459797

[52] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the nystrom method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 214–225, 2004.

[53] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. sitzmann, and S. Sridhar, "Neural fields in visual computing and beyond," *Comput. Graph. Forum*, vol. 41, no. 2, pp. 641–676, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14505

[54] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "DeepSDF: Learning continuous signed distance functions for shape representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 165–174.

[55] J. Mu, W. Qiu, A. Kortylewski, A. L. Yuille, N. Vasconcelos, and X. Wang, "A-SDF: learning disentangled signed distance functions for articulated shape representation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 12 981–12 991.

[56] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, "Implicit geometric regularization for learning shapes," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3569–3579.

[57] H. Bertiche, M. Madadi, and S. Escalera, "Cloth3D: clothed 3D humans," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 344–359.

[58] Carnegie-Mellon, "CMU graphics lab motion capture database," http://mocap.cs.cmu.edu/, 2010, accessed: 2023.

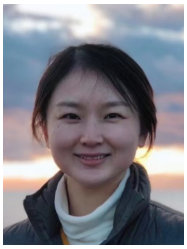
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034. [Online]. Available: https://doi.org/10.1109/ICCV.2015.123

[60] M. Atzmon, N. Haim, L. Yariv, O. Israelov, H. Maron, and Y. Lipman, *Controlling Neural Level Sets*. Red Hook, NY, USA: Curran Associates Inc., 2019.

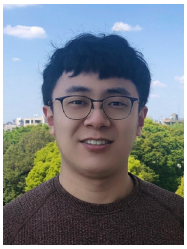
[61] Q. Tan, Y. Zhou, T. Wang, D. Ceylan, X. Sun, and D. Manocha, "A repulsive force unit for garment collision handling in neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 451–467.

[62] "Marvelous designer," https://www.marvelousdesigner.com/, accessed: 2023.

[63] B. Deng, J. P. Lewis, T. Jeruzalski, G. Pons-Moll, G. Hinton, M. Norouzi, and A. Tagliasacchi, "NASA neural articulated shape approximation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, p. 612–628. [Online]. Available: https://doi.org/10.1007/978-3-030-58571-6_36



Tianxing Li is a lecturer in the Faculty of Information Technology, Beijing University of Technology, Beijing, China. Her current research interests include computer animation, visualization, and pattern recognition. She received her Ph.D. degree in computer science from the University of Tokyo, Tokyo, Japan, in 2021.



Rui Shi is a lecturer in the Faculty of Information Technology, Beijing University of Technology, Beijing, China. His current research interests include explainable artificial intelligence, computer animation, and visualization. He received his Ph.D. degree in computer science from the University of Tokyo, Tokyo, Japan, in 2022.

Qing Zhu is a professor in the Faculty of Information Technology, Beijing University of Technology, Beijing, China. Her research interests include multimedia information processing technology, virtual reality technology, and information integration technology. She received her Ph.D. degree in electronic information and communication from Waseda University, Tokyo, Japan, in 2000.

Takashi Kanai is an associate professor in the Graduate School of Arts and Sciences, the University of Tokyo. His research interests include geometry processing and physics-based animation in computer graphics. He received his Ph.D. degree in engineering from the University of Tokyo in 1998. He is a member of ACM, ACM SIGGRAPH, IIEEJ (the Institute of Image Electronics Engineers of Japan), and IPSJ (Information Processing Society of Japan).