CHA: Physics-Based Dynamics Modeling for Clothing-Editable Human Avatars

Yong Deng, Baoxing Li, Xu Zhao* *Member, IEEE*

***Abstract*—This paper proposes a novel framework for generating human avatars from monocular videos. Existing methods struggle to balance dynamic effects and modeling efficiency due to the need for per-subject retraining. To address this, we introduce CHA, a physics-based approach for Clothing-editable Human Avatars, which decouples reconstruction and driving, thereby eliminating the need for per-subject fitting, and utilizes physics-based cloth dynamics modeling to generate realistic clothing dynamics. Our framework comprises three core components: First, an avatar reconstruction module employs an attention mechanism to fuse multi-frame normal information, ensuring complete human surface recovery. Second, a cloth recovery module exploits the static equilibrium state to restore standard clothing templates from observed garments, facilitating downstream dynamics modeling. Third, we propose a new physics-based cloth dynamics module to generate realistic cloth-aware avatar animations. The module generalizes well across different clothing sizes and styles by employing dual-layer graph networks with local predictions and a dynamic KNN-based graph construction that is not limited by mesh topology. Furthermore, we employ a multi-frame joint semantic segmentation algorithm to separate clothing templates, making our method applicable to clothing editing applications, such as virtual try-on. Extensive quantitative and qualitative experiments demonstrate that our avatars exhibit enhanced physical accuracy in clothing dynamics, well-preserved individual body characteristics, and distinct clothing layers.**

***Index Terms*—Human avatar, cloth dynamics modeling, monocular video, graph networks, attention mechanism.**

I. Introduction

HUMAN avatar modeling [1] aims to reconstruct animatable 3D digital humans from visual observations, enabling a wide range of applications, including multimedia content generation and editing, immersive virtual and augmented reality scenarios, and virtual try-on. Recent approaches [69]–[72] commonly leverage parametric human body models to represent pose and shape, and adopt implicit neural surface representations to capture high-fidelity geometry and appearance. These methods employ non-rigid deformation prediction modules or unified implicit representations to model pose-dependent dynamic deformations of human avatars.

Despite notable progress, current human avatar methods from monocular RGB videos still face two major challenges: (1) *Lack of clothing dynamics*. Methods such as [26]–[29] follow a subject-agnostic reconstruction paradigm, in which avatar reconstruction and driving are decoupled into two independent stages. They first employ implicit neural networks to reconstruct detailed 3D human surfaces and then animate the avatars by extracting deformation fields from parametric models such as SMPL [7]. This design generalizes across subjects and avoids per-subject retraining. However, the extracted deformation fields do not explicitly model pose-dependent garment deformations, leading to limited expressiveness in capturing fine-grained clothing variations across different poses. (2) *Low modeling efficiency*. In contrast, methods such as [12]–[15] adopt a subject-specific fitting strategy, in which non-rigid surface deformations are directly embedded into implicit neural representations. After optimization, the learned implicit models can directly generate the human surface for a given pose, thereby implicitly capturing clothing dynamics. However, since the learned deformation is tightly coupled with the subject's geometry, these methods require per-subject optimization or retraining, leading to low modeling efficiency.

Given the aforementioned challenges, existing monocular methods struggle to simultaneously achieve realistic clothing dynamics and efficient avatar modeling. Furthermore, treating the body and clothing as a unified entity limits the flexibility for independent modifications. Therefore, we propose CHA, a novel physics-based approach for clothing-editable human avatars. As shown in Fig. 1, our method utilizes multi-frame images extracted from monocular videos as input to generate human avatar animations. *For modeling efficiency*, following the subject-agnostic reconstruction paradigm, we separate modeling and driving into two independent stages, thereby avoiding repetitive training for new subjects. Additionally, we map the feature space to a canonical space to directly reconstruct the avatar in the standard pose, thereby avoiding additional transformations. *For clothing dynamics*, we integrate cloth dynamics modeling [67] into our framework to address the lack of dynamics. By leveraging the physical properties of cloth, we guide the modeling network to capture dynamic deformation patterns. Moreover, to enable independent modeling of clothing dynamics, we propose a semantic segmentation-based algorithm to separate clothing templates from the inner body model. The independent clothing templates facilitate highly flexible and customizable modifications.

As shown in Fig. 2, our method consists of four modules: avatar reconstruction, cloth separation, cloth recovery, and cloth dynamics modeling. In the avatar reconstruction step, to mitigate surface information loss due to the absence of subject-specific fitting, we introduce an attention-based multi-frame normal fusion strategy for comprehensive avatar reconstruction. Next, we extract clothing templates using a multi-frame joint semantic segmentation algorithm, and then construct the inner body model by extracting head geometry from the reconstructed avatars and combining it with body geometry derived from the parametric model. To remove dynamic details

Yong Deng, Baoxing Li and Xu Zhao are with the School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai, China. Email: {dengyong2016, lbx11sjtu, zhaoxu}@sjtu.edu.cn. * Corresponding author.

Fig. 1. **Animations of our avatars.** CHA generates human avatars from multi-frame image sequences, featuring advanced clothing dynamics, sharp individual characteristics, and distinct clothing layers. The independent clothing templates enable highly customizable modifications and are readily applicable to clothing editing applications, such as virtual try-on.

and create standardized clothing templates, we introduce a cloth recovery module that minimizes energy consumption to reach a static equilibrium state. Finally, the cloth dynamics modeling module is used to generate the avatar animations through physics-based dynamics modeling.

To realize realistic avatar animations, the cloth dynamics modeling module is responsible for predicting physically plausible clothing deformations. To accommodate diverse reconstructed garments with varying styles and sizes, we propose a novel physics-based modeling approach that combines local prediction with a dual-layer graph architecture. Local prediction reduces sensitivity to garment variability, while the hierarchical graph ensures effective deformation propagation. Specifically, the coarse-level graph captures and transmits global deformations across the entire garment surface, ensuring responsiveness to overall pose changes. Meanwhile, the fine-level graph focuses on local vertex neighborhoods, enabling the modeling of fine-grained deformations. Furthermore, we introduce a global optimization algorithm during the dynamics modeling stage that enforces spatiotemporal consistency in clothing dynamics while avoiding collisions between garments. During the animation process, the parametric model drives the inner body motion, while the cloth dynamics module animates the reconstructed clothing templates, generating temporally coherent, physically grounded animation sequences.

In summary, the main contributions of this paper are as follows:

1) A novel human avatar modeling framework balancing avatar modeling efficiency and clothing dynamic effects, with a hierarchical clothing representation that seamlessly extends to clothing editing applications.
2) A human avatar reconstruction method that fuses multi-frame normal information via an attention mechanism and leverages human body priors from parametric models to achieve complete and detailed human avatar reconstruction in the canonical space.
3) A semantic segmentation-based cloth separation algorithm and a cloth static recovery algorithm that extract clothing templates and restore them into standardized static templates for independent dynamics modeling.
4) A physics-based clothing dynamics modeling method based on dual-layer graph networks that captures both global and local clothing dynamics, achieving state-of-the-art performance on the CLOTH3D [3] dataset with superior generalization across garment types.

II. Related Work

A. Human Avatar Modeling

Surface Field-Based Methods. Current methods primarily use implicit neural networks to represent the human body surface. Commonly used implicit functions include the signed distance field (SDF) [24] and the occupancy field (OF) [25], which represent the human surface using level sets. Some methods [27], [28], [30]–[33] first reconstruct the human surface and then warp it to the posed space. ECON [28] reconstructs the human surface and maps it to a standard pose, potentially causing geometric distortion due to unseparated limbs. CAR [33] directly predicts the human surface in a canonical space by transforming features accordingly. These methods rely on linear blend skinning for motion control but fail to capture clothing dynamics. Therefore, some methods [12], [16]–[18] propose using implicit networks to jointly fit the human surface and its driving patterns. Vid2avatar [12] represents the human geometry as a pose-conditioned SDF and then uses neural volumetric rendering for learning. After implicit fitting, they take pose inputs to generate the corresponding models. This unified implicit representation demands extended fitting times for each subject, leading to low modeling efficiency.

Radiance Field-Based Methods. Some methods [15], [64]–[66] employ Neural Radiance Fields (NeRF) [22] for human modeling via volumetric rendering. Animatable NeRF [34] represents human avatars using NeRF, where the neural network takes position and view direction as inputs, outputting density and color. The learned skinning weight field enables animation of the avatar by deforming the radiance field according to skeletal motion. Approaches like [19], [20] represent avatars as a set of 3D Gaussians, defined by position, covariance, opacity, and color, achieving faster rendering speeds through efficient and compact representation. These Gaussian-based methods reduce memory consumption and enable real-time rendering, making them more practical for interactive applications. While radiance field-based approaches yield high-quality and photorealistic rendering results with rich view-dependent effects, they often suffer from lower geometric accuracy compared to explicit mesh-based models. Additionally, they require time-consuming per-subject training, which limits their immediate applicability to new subjects.

Existing methods face challenges in avatar modeling efficiency while capturing detailed clothing dynamics. To address these limitations, we propose a novel physics-based avatar

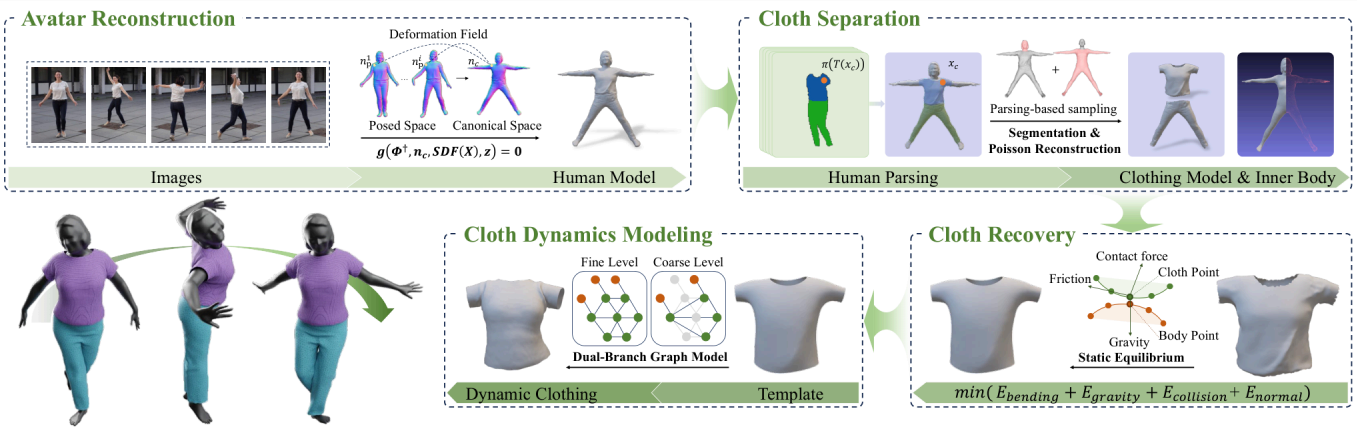

Fig. 2. **CHA overview.** Our framework comprises four primary modules: (1) Avatar Reconstruction Module, which reconstructs the human avatar in the canonical space using temporal information. (2) Cloth Separation Module, which employs semantic segmentation to obtain garment regions for segmenting the clothing template from human models. (3) Cloth Recovery Module, which optimizes the garment template based on its physical properties in static equilibrium, producing a standardized template for subsequent driving. (4) Cloth Dynamics Modeling Module, which predicts garment dynamic deformations using a dual-branch graph network grounded in physics-based deformation models.

generation framework that separates body and clothing representations. Specifically, we leverage implicit reconstruction to facilitate rapid avatar creation without per-subject training, and decouple the clothing from the body surface to allow for independent clothing dynamics modeling. The framework supports downstream tasks such as virtual try-on like [6], but features a more realistic inner body model.

B. Cloth Dynamics Modeling

Physics-Based Simulation. Physics-based simulation [35] models the dynamic behavior of clothing using physical laws and mathematical formulations. Common approaches include differentiable cloth simulation [36], [37], the finite element method (FEM) [38], mass-spring systems [39], and fluid dynamics-based methods [40]. Building upon these simulation models, PhysAvatar [76] employs a mass-spring system for cloth simulation with learned garment material parameters to model clothing dynamics. In a similar spirit, DiffAvatar [74] incorporates a differentiable cloth simulation based on continuum mechanics to enable inverse optimization of simulation-ready clothing dynamics. Despite their high accuracy and physical realism, these simulations often suffer from high computational complexity and significant time costs.

Learning-Based Methods. To overcome simulation limitations, some methods [41]–[45], [63] use ground truth supervision to train networks for dynamic prediction. However, collecting ground truth data is complex. Consequently, some self-supervised learning methods [49]–[51] leverage clothing's physical properties to supervise the deformation model. SNUG [46] uses physics-based losses, such as stress and bending, to train a neural network that predicts per-vertex offsets. These offsets are applied to the template to represent clothing dynamics. CaPhy [75] introduces a physics-aware neural deformation model based on a 4-layer GRU network that jointly learns clothing dynamics and latent garment physical parameters. HOOD [47] employs a hierarchical graph representation and graph neural networks to enhance the model's generalization ability. However, the HOOD node graph is extracted from regular garment meshes, whereas the segmented templates often suffer from topological inconsistencies, sparsity, holes, and irregular boundaries. As a result, directly using the original HOOD graph structure makes generalization challenging. GAPS [48] performs a geometrically aware mechanism to compute body participation in clothing dynamics, improving the handling of loose-fitting garments.

The aforementioned methods are not readily applicable to segmented clothing templates due to irregular mesh topology and incomplete mesh connectivity. To address this limitation, we adopt a dual-layer graph network that captures garment deformations from coarse to fine scales. In contrast to HOOD, which relies on implicitly hierarchical message passing, our design enforces an explicit functional decomposition to prevent the smoothing or dilution of fine-grained details during multi-level propagation. Specifically, the coarse-level graph focuses on modeling global deformations, while the fine-level graph specializes in learning high-frequency details. This separation is further reinforced by a two-stage training strategy that explicitly guides the learning of each level. Furthermore, we dynamically construct the graph structure using KNN [73], which provides robust adjacency under unstable segmented topologies. These design distinctions allow our graph network to robustly capture the dynamics of clothing templates.

III. Method

Our goal is to create a drivable, clothing-editable, and dynamically realistic 3D human avatar. An overview of our framework is shown in Fig. 2. To enhance human avatar reconstruction, we propose a human reconstruction method based on temporal information (Sec. III-A). Next, human semantic segmentation is leveraged to extract clothing templates for independent dynamics modeling (Sec. III-B). To standardize the clothing templates in the canonical space, we optimize them using energy properties in a static equilibrium state, thereby removing excessive clothing details arising from reconstruction (Sec. III-C). Finally, we adopt a physics-based

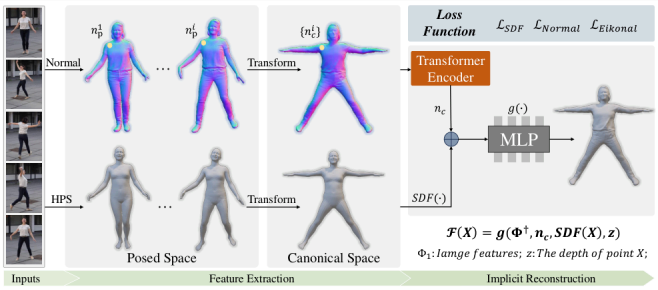

Fig. 3. **Architecture for avatar reconstruction.** The model takes sampled frames from monocular videos as input, predicting normal maps and the SMPL-X [78] parameters using the normal prediction (Normal) network and human pose and shape estimation (HPS) network.

cloth deformation model for clothing dynamics prediction, resulting in a fully animated 3D human avatar (Sec. III-D).

A. Avatar Reconstruction

Following [33], we construct avatars in the canonical space. Instead of repetitive optimization, we integrate temporal information to generate human models. To extract temporal features, we select N complementary frames based on the estimated global orientation of the parametric model.

Basic Structure. As shown in Fig. 3, given a set of sampled RGB images $\{\mathcal{I} \in \mathbb{R}^{H\times W\times 3}\}$, we employ a normal prediction network to estimate the corresponding normal maps $\{\mathcal{N} \in \mathbb{R}^{H\times W\times 3}\}$. For a spatial point x_c in the canonical space, we transform it to the posed space via linear blend skinning [7] to obtain the corresponding point x_p:

$$x_p = T(x_c) = \sum_{i=1}^{N_J} w_i G_i(\theta, J) x_c, \tag{1}$$

where w is the skinning weights of joints J on vertices and $G(\theta, J)$ represents the bone transformations of N_J joints under the estimated pose θ [2]. Subsequently, x_p is projected into the image space using a weak perspective camera projection $\pi(\cdot)$, and the corresponding normal features are extracted via bilinear sampling $\mathcal{B}(\cdot)$:

$$n_p = \mathcal{B}(\mathcal{N}, \pi(T(x_c))), \tag{2}$$

where n_p is the normal vector corresponding to the point x_c in image space. We use an attention mechanism $\mathcal{A}(\cdot)$ [53] to fuse normal information n_p^i from different frames to form a unified normal feature. To better extract useful normal features, the visibility $\mathcal{V}^i$ of x_c in the posed space of i frame is concatenated with the normal features, forming a new feature vector n_c:

$$n_c = \mathcal{A}(\{n_p^i, \mathcal{V}^i\}). \tag{3}$$

Finally, we combine image features $\Phi^\dagger$, the temporally fused normal features n_c, and the SDF features of the parametric model $SDF(\cdot)$ and input them into an implicit network to infer the human avatar. The zero level set of the human body surface can be represented as:

$$\mathcal{F}(x_c) = g(\Phi^\dagger, n_c, SDF(x_c), z) = 0, \tag{4}$$

where z is depth of point x_c.

Training Loss. Our loss function includes geometric loss $\mathcal{L}_{SDF}$, normal loss $\mathcal{L}_{Normal}$, and regularization loss $\mathcal{L}_{Eikonal}$:

$$\mathcal{L} = \mathcal{L}_{SDF} + \lambda_n \mathcal{L}_{Normal} + \lambda_e \mathcal{L}_{Eikonal}, \tag{5}$$

where λ_n and λ_e are hyper-parameters. The geometric loss is the L1 error between the predicted SDF values $\mathcal{F}(\cdot)$ and the ground truth $\mathcal{F}^*(\cdot)$ of the human mesh. The human model is transformed into the posed space for consistency supervision:

$$\mathcal{L}_{SDF} = \frac{1}{N_v} \sum_{i=1}^{N_v} |\mathcal{F}(x_c^i) - \mathcal{F}^*(T(x_c^i))|, \tag{6}$$

where N_v is the number of query points. The normal loss is the L1 error between the predicted normals $\mathcal{N}_{pred}$ and the normals of the reconstructed model $\mathcal{N}_{recon}$:

$$\mathcal{L}_{Normal} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} |\mathcal{N}_{pred}(i,j) - \mathcal{N}_{recon}(i,j)|, \tag{7}$$

where H and W denote the height and width of the normal maps. The regularization term uses the Eikonal loss [68] to constrain the SDF by enforcing unit gradients of normals.

B. Cloth Separation

To independently model clothing dynamics, we decouple the clothing template and the inner body model from the reconstructed avatar (Sec. III-A). First, we employ an off-the-shelf, pre-trained Sapiens-2B model [54] to segment the clothing regions from the input images. This model is used as-is, without any additional training or fine-tuning. Then, the reconstructed 3D vertices are projected onto the 2D image plane using camera intrinsic and extrinsic parameters. Classification of human components is performed in the image space based on the projected pixel coordinates and the segmentation masks. Given the presence of limb occlusions, misclassifications may occur. To mitigate this, we employ multi-frame joint classification to exclude vertices not belonging to the clothing. This classification is based on the principle that, except for occluded regions, vertices belonging to the clothing should fall within the corresponding 2D clothing regions when projected onto each image frame. Specifically, for a vertex x_c in the canonical space, we obtain its corresponding position x_p^t in the posed space at frame t via linear skinning, as described in Sec. III-A. The semantic label s of this vertex is then determined through multi-frame joint classification:

$$s = \arg\max_{n\in C} \sum_{t\in\mathcal{T}} v_t I_t(\pi(x_p^t)), \tag{8}$$

where C denotes the set of semantic classes, $\mathcal{T}$ denotes the set of frames, and I_t denotes the 2D semantic segmentation result of frame t. $I_t(\pi(x_p^t))$ gives the class probability at the projected location $\pi(\cdot)$ of the posed point x_p in frame t. The visibility v_t of each point x_p^t is estimated based on the angle between its associated surface normal and the viewing direction. In practice, joint classification is conducted using five frames that are also used for avatar reconstruction. After assigning semantic labels to all vertices, the clothing template is naturally extracted. Finally, the inner body model is derived

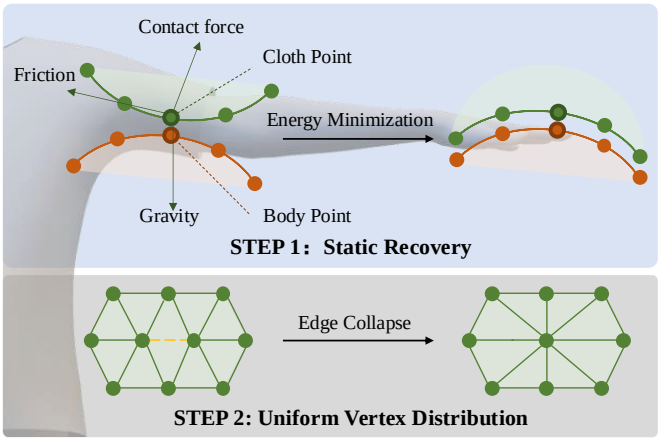

Fig. 4. **Steps for cloth recovery.** Firstly, optimize the cloth template by utilizing the principle of minimizing cloth energy under static equilibrium. Secondly, apply the edge collapse algorithm to balance the cloth vertices.

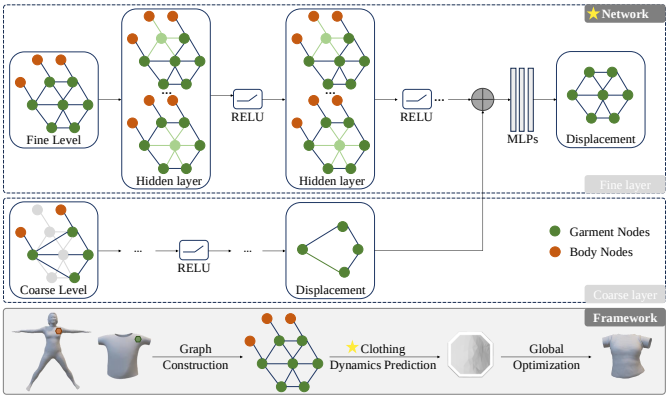

Fig. 5. **Schematic diagram of cloth dynamics modeling.** A dual-layer graph network infers local clothing dynamics, followed by a global optimization algorithm for refining global dynamics. The steps corresponding to the network are marked with ★.

by sampling head data from the reconstruction model and limb data from the parametric model, followed by Poisson reconstruction [55]. The inner body model is then deformed using the parametric model's deformation domain, determined via vertex nearest-neighbor computation. Finally, the clothing template is driven by the inner body model via the cloth dynamics module (Sec. III-D).

C. Cloth Recovery

The separated clothing template contains the original dynamic details from the reconstruction. As shown in Fig. 4, we standardize the clothing template through two steps: static recovery and uniform vertex distribution.

Static Recovery. When stationary, clothing's bending and draping cause energy accumulation. According to the minimum potential energy principle, the system naturally reaches a minimum energy state in equilibrium without external forces. To reduce excessive details, we optimize the vertex positions in wrinkled regions. The energy terms considered in the optimization include bending, gravity, collision, and normal:

$$E_R = E_{bending} + E_{gravity} + E_{collision} + E_{normal}. \quad (9)$$

As shown in Fig. 4, when the garment is in a state of static equilibrium, based on the principle of minimum potential energy, the system tends to minimize its energy in the absence of external forces. Accordingly, we penalize both the bending energy and gravitational potential energy of the garment. The bending term is obtained by modeling the energy formed by the dihedral angle θ of each edge:

$$E_{bending} = \sum_{edges} k_b \theta^2, \quad (10)$$

where k_b is bending stiffness. The gravity term is formulated as follows:

$$E_{gravity} = \sum_{vertices} -mg^{\mathsf{T}} x, \quad (11)$$

where m is the vertex mass estimated by the area of the connected faces, g is the gravitational acceleration, and x is the vertex position. The collision term is designed to prevent the garment from penetrating the inner body:

$$E_{collision} = \sum_{vertices} k_c \max(\varepsilon_b - d(x), 0), \quad (12)$$

where k_c is collision stiffness, ε_b is the safe boundary and $d(x)$ is the distance to inner body. To obtain a smoother garment template, we add a normal term to penalize sudden changes in the normals:

$$E_{normal} = \sum_{edges} k_n (1 - n_1 \cdot n_2), \quad (13)$$

where n_1 and n_2 are the normals of the adjacent face patches of the edge, and k_n is the coefficient of this term.

Uniform Vertex Distribution. The recovered garment template exhibits a high density of vertices in regions with rich original details, which in turn results in reduced computational efficiency due to redundant geometric resolution. Therefore, we employ the edge collapse algorithm [56] to reduce the number of vertices in these dense regions. In this process, an edge is selected to merge two vertices, with each merge reducing two triangles, three edges, and one vertex. To preserve local geometric features while improving computational performance, the merging cost $C = L \times K$, where L is the edge length and K is the average curvature of the faces connected by the edges, is computed and the edge with the lowest cost is selected for each merge. This operation is repeated iteratively until a desired vertex density is achieved across the mesh.

D. Cloth Dynamics Modeling

Our clothing templates are derived from reconstructed human models. To adapt to different reconstructed clothing templates, we designed a local-based dual-layer graph network, as illustrated in Fig. 5, to predict dynamic details. The local-based prediction enables the model to handle diverse clothing types. Furthermore, we propose a global optimization algorithm to enhance global dynamic details. During training, the network leverages a parametric model to guide garment deformation, and during inference, it predicts vertex displacements based on input pose parameters.

Fig. 6. Qualitative results of our avatar generation method under various poses and subjects. CHA is capable of generating realistic human avatars from multi-frame images and producing corresponding clothing dynamics under different poses.

Graph Construction. We build the garment graph using a k-nearest neighbor (KNN) strategy. For each vertex, we connect it to its k closest neighbors in 3D space to form an undirected graph. This dynamic construction does not rely on the original mesh topology, making it more robust to irregular shapes and non-uniform vertex distributions commonly found in segmented garment templates. It provides a flexible and stable structure for graph-based learning.

Cloth Dynamics Prediction. To improve the network's generalization across different clothing types, we divide the clothing templates into local patches for detailed prediction and employ a graph [62] to unify the garment structure. Given that garment dynamics are globally correlated, we propose a dual-layer graph network to capture and propagate dynamic changes during local predictions. Within each graph layer, the message passing mechanism is identical to that used in HOOD [47]. The dual-layer graph network explicitly separates responsibilities: the coarse level is dedicated to learning global garment dynamics, while the fine level focuses on modeling local geometric details. This explicit functional decomposition helps avoid the smoothing or dilution of fine-grained details that can arise in HOOD's hierarchical graph network during long-range information propagation. The graph model is downsampled and used as input to the coarse level. Additionally, we adopt a two-stage training approach: first, we train the coarse level, then encode its results as features that are input into the fine level to assist with local detail prediction. The dual-layer graph structure ensures that local detail predictions dynamically follow the region's motion.

The model is trained using a self-supervised method with a physics-based loss function, as detailed in [46], to guide the model's learning process. The loss is defined as follows:

$$\mathcal{L}_D = \mathcal{L}_{inertia} + \mathcal{L}_{strain} + \mathcal{L}_{bending} + \mathcal{L}_{gravity} + \mathcal{L}_{collision}. \tag{14}$$

The inertial term ensures that the velocity changes of the clothing are consistent with the body motion:

$$\mathcal{L}_{inertia} = \frac{1}{2\Delta t^2}(x^t - x^{t-1})^{\mathsf{T}} M (x^t - x^{t-1}), \tag{15}$$

where M is the mass matrix and Δt is the time step. The strain term models the energy during the compression and stretching of the clothing. Its definition is as follows:

$$\mathcal{L}_{strain} = \alpha tr(G)^2 + \beta tr(G^2), \tag{16}$$

where G is the Green strain tensor and α, β are the Lamé constants [57]. The other terms are the same as those introduced in Sec. III-C for static recovery. During training, garment material–related parameters, including bending and stretching coefficients, Lamé constants, and mass density, are randomized by sampling from empirically validated ranges, which enables the model to learn a wide spectrum of fabric behaviors without requiring explicit supervision.

Global Optimization. Three optimization terms are introduced to improve the continuity in both time and space:

$$E_G = E_{spat} + E_{temp} + E_{sep}. \tag{17}$$

The spatial term penalizes abrupt changes of curvature η in the mesh, eliminating unnatural transitions:

$$E_{spat} = \sum_{(i,j)\in edges} k_s(\eta_i - \eta_j)^2, \tag{18}$$

where k_s is the spatial term coefficient. The temporal term enforces smoothness constraints on velocity v and acceleration a, penalizing abrupt changes in the structure over time, thus achieving complementary optimization temporally:

$$E_{temp} = \sum_{vertices} k_t(\|v_t - v_{t-1}\|^2 + \|a_t - a_{t-1}\|^2), \tag{19}$$

where k_t is the temporal term coefficient. The separation term is used to prevent the intersection between garments:

$$E_{sep} = \sum_{i\in P_t} \sum_{j\in P_b} \max(\delta - D(x_i, x_j), 0), \tag{20}$$

TABLE I
QUANTITATIVE EVALUATION OF DYNAMICS MODELING. THE GEOMETRIC CONSISTENCY ε_g, ENERGY CONSUMPTION ε_e, AND COLLISION RATE ε_c ON CLOTH3D [3] ARE PRESENTED. WE EVALUATED FOUR CLOTHING CATEGORIES, INCLUDING T-SHIRTS, PANTS, DRESSES, AND TANK TOPS. LOWER VALUES INDICATE BETTER DYNAMICS. THE BEST RESULTS ARE HIGHLIGHTED IN GREEN, AND THE SECOND-BEST RESULTS IN ORANGE.

Method	T-shirt			Pants			Dress			Tank		
	ε_g	ε_e	ε_c	ε_g	ε_e	ε_c	ε_g	ε_e	ε_c	ε_g	ε_e	ε_c
SNUG [46]	0.075	0.132	4.821	0.068	0.116	2.962	0.142	0.154	5.316	0.067	0.843	2.725
HOOD [47]	0.092	0.096	2.814	0.067	0.675	2.124	0.097	0.106	2.065	0.076	0.475	1.942
GAPS [48]	0.081	0.071	0.761	0.055	0.059	0.564	**0.073**	**0.080**	0.768	0.071	0.046	0.477
Ours	**0.065**	**0.051**	**0.156**	**0.053**	**0.035**	**0.126**	0.087	0.094	**0.463**	**0.060**	**0.029**	**0.118**

where $D(\cdot)$ is the distance between vertices on the top and bottom garments. P_t and P_b are the sets of vertices of the top and bottom garments. δ is the safety boundary. If the distance is smaller than the safety boundary, an intersection risk between the garments is considered.

IV. EXPERIMENTS

A. Datasets and Metrics

Training Data. The *MVP-Human* dataset [58] is employed to train the avatar reconstruction network. It comprises 400 subjects, each with 15 scans in various poses and 8 images per pose from different viewpoints. We utilize *CLOTH3D* [3] and *AMASS_CMU* [59] as the clothing templates and motion sequences, respectively, to train a graph network for modeling clothing dynamics. The *CLOTH3D* contains seven categories of clothing templates along with SMPL-based simulation sequences. Like DrapeNet [49], we select 600 top garments and 300 bottom garments from the dataset for training. *AMASS_CMU* includes 52 motion sequences.

Testing Data. The *MonoPerfCap* [60] and *Adobe* [77] datasets are used to evaluate the performance of avatar generation. The MonoPerfCap contains 13 sequences, of which 2 sequences include 3D ground truth. The Adobe dataset is an indoor dataset containing three subjects and ten action sequences, where each frame is associated with a corresponding 3D mesh. We select 200 simulation sequences from *CLOTH3D* for the quantitative evaluation of the cloth dynamics modeling module, which were not used during training.

Evaluation Metrics. ε_g is used to evaluate the geometric similarity between the cloth dynamics modeling results and the simulation ground truth, calculated using the Chamfer distance between vertices. ε_e represents the energy consumption in forming clothing dynamics on the test data, including inertia, stress, bending, and gravity, as defined in Sec. III-D. ε_c is the percentage of collisions between clothing vertices and the inner body model. Meanwhile, we evaluate the reconstruction accuracy using the Point-to-Surface distance (*P2S*), Chamfer distance (*Chamfer*), and Normal error (*Normal*). In addition, we report the mean Intersection over Union (*mIoU*) and mean Accuracy (*mAcc*) to assess the segmentation accuracy by projecting the 3D clothing template in the posed space back to the 2D image space and comparing it with the corresponding ground-truth clothing masks.

TABLE II
QUANTITATIVE EVALUATION OF AVATAR RECONSTRUCTION. COMPARISON WITH OTHER SINGLE-FRAME (**SF**) AND MULTI-FRAME (**MF**) METHODS ON THE MONOPERFCAP AND ADOBE DATASETS.

Method	MonoPerfCap [60]			Adobe [77]		
	P2S↓	Chamfer↓	Normal↓	P2S↓	Chamfer↓	Normal↓
PIFu (SF) [27]	4.568	3.785	0.217	4.684	3.923	0.243
CAR (SF) [33]	2.683	2.106	0.141	2.824	2.321	1.156
Ours (SF)	2.253	1.915	0.137	2.416	2.072	0.144
PIFu+AP (MF)	2.715	2.068	0.224	3.113	2.251	0.231
Ours+AP (MF)	2.025	1.861	0.147	2.182	1.941	0.143
Ours+AM (MF)	**1.718**	**1.636**	**0.115**	**1.825**	**1.722**	**0.126**

TABLE III
COMPARISON OF EXISTING HUMAN AVATAR MODELING METHODS. THE COMPARISON IS CONDUCTED WITH RESPECT TO HIERARCHICAL REPRESENTATION (**HIERARCHY**), EXPLICIT MODELING OF CLOTHING DYNAMICS (**DYNAMICS**), GENERALIZATION TO UNSEEN SUBJECTS WITHOUT RETRAINING (**GENERAL**), AND APPROXIMATE PER-SUBJECT MODELING TIME (**TIME**).

Method	Hierarchy	Dynamics	General	Time
Subject-agnostic reconstruction methods				
ClothWild [61]	✓	✗	✓	∼ 10.0 s
ICON [29]	✗	✗	✓	∼ 1.4 min
ECON [28]	✗	✗	✓	∼ 2.0 min
CAR [33]	✗	✓	✓	∼ 5.0 min
Subject-specific fitting methods				
Vid2Avatar [12]	✗	✓	✗	∼ 25.0 h
HumanNeRF [15]	✗	✓	✗	∼ 10.0 h
InstantAvatar [14]	✗	✓	✗	∼ 3.0 min
Ours	✓	✓	✓	∼ 2.5 min

B. Comparison and Validation

Quantitative Analysis. We present a quantitative comparison of our model with previous methods, including SNUG [46], HOOD [47], and GAPS [48]. As shown in Tab. I, our clothing dynamics prediction results are closer to the simulation ground truth, as indicated by the ε_g metric. The local-based predictions enhance the generalization capability of our method, while they slightly reduce the dynamic prediction performance for dresses. Additionally, the results of the ε_e and ε_c metrics demonstrate that our method achieves lower energy consumption and a reduced collision rate between clothing and the inner body.

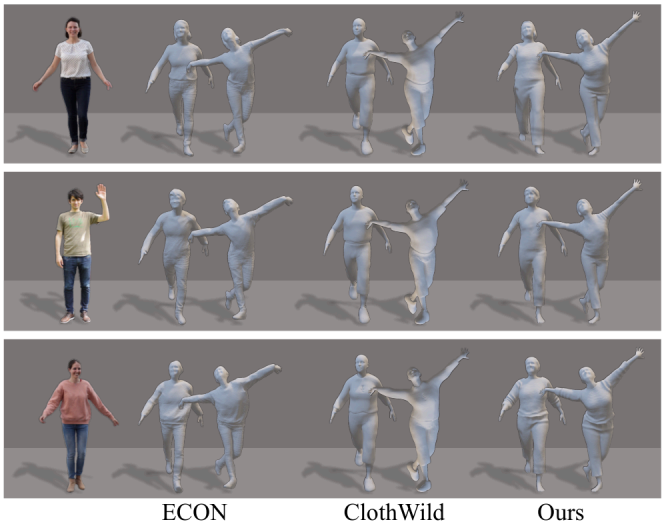

Fig. 7. Qualitative comparison of avatar generation methods with other frameworks, ECON [28] and ClothWild [61].

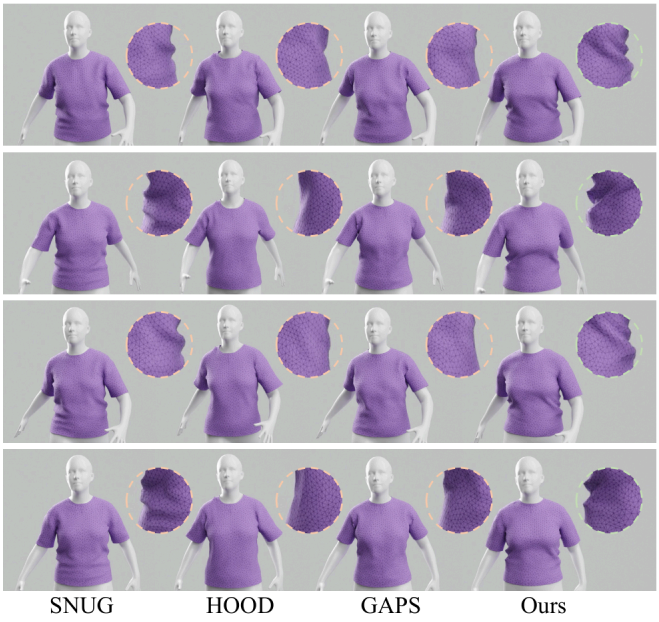

Fig. 8. Qualitative comparison of the dynamics modeling module with other state-of-the-art methods, including SNUG [46], HOOD [47], and GAPS [48].

The evaluation of avatar reconstruction is presented in Tab. II, comparing the baseline CAR [33] and the classical implicit method PIFu [27]. We conduct evaluations on the outdoor *MonoPerfCap* dataset [60] and the indoor *Adobe* dataset [77]. For multi-frame inputs, we first reconstruct the human model in a canonical pose and then transform it to the corresponding pose of each frame for evaluation. With a single-frame input, our method outperforms CAR due to the incorporation of geometric priors. With multi-frame inputs (5 frames), our attention mechanism-based (**AM**) feature aggregation method outperforms PIFu's average pooling (**AP**) aggregation. According to the normal error, certain geometric details are smoothed out by the average pooling aggregation.

Qualitative Analysis. Fig. 6 presents the qualitative evaluation results of our proposed avatar generation method. As shown in the figure, CHA is capable of generating realistic human avatars for various subjects. Moreover, for different poses, our method generates detailed and realistic clothing dynamics. To highlight the advantages of our framework, we perform a qualitative comparison with other avatar generation frameworks, ECON [28] and ClothWild [61], as shown in Fig. 7. We select frames from the video with poses closest to the standard pose as input to the model to demonstrate the best performance of the single-frame method for comparison. Compared to ECON, CHA separates the clothing template for dynamics modeling, achieving more realistic clothing dynamics and clearer clothing hierarchy. In addition, compared to the generative model-based approach, ClothWild, our method produces avatars with better individual features, such as hair.

Fig. 8 presents a qualitative comparison of our cloth dynamics modeling module with advanced dynamics modeling methods, including SNUG [46], HOOD [47], and GAPS [48]. As shown in the zoomed-in region, aided by physical modeling and deformation propagation via a dual-layer graph network, our approach produces dynamic representations more consistent with physical laws. Compared to other methods, our method produces more natural clothing dynamics while achieving superior generalization capabilities. Moreover, Fig. 9 presents the dynamics modeling results of the same clothing template applied to different body types. As illustrated, our method effectively accommodates human body variations and generates dynamic effects for challenging poses.

Efficiency Analysis. We improve avatar modeling efficiency from three aspects: (1) Our modeling framework separates modeling and driving into two independent stages, allowing our method to handle new subjects without retraining. (2) Our avatar reconstruction module is capable of directly integrating multi-frame information to reconstruct a T-pose human avatar, without requiring secondary deformation to obtain a canonical-pose representation. (3) Compared to other clothing dynamics modeling methods, our method can adapt to different clothing types and sizes, eliminating the need to train a separate network for each type of clothing. Tab. III presents a comparison of the single-subject avatar modeling time among representative subject-agnostic reconstruction methods, subject-specific fitting methods, and our approach. Compared to existing methods, our method strikes a balanced trade-off between generalization, dynamics modeling, and avatar modeling efficiency. In practice, the single-avatar modeling process of our method takes approximately 2.5 minutes, achieving a competitive level of efficiency.

C. Ablation study

Avatar Reconstruction. Our method integrates information from multiple frames to jointly reconstruct the human avatar. Tab. IV illustrates the impact of the number of image inputs on the reconstruction results. The testing was conducted on two ground-truth sequences from MonoPerfCap [60]. Although using more input frames provides richer human body information, we ultimately choose 5 frames as input. This is because increasing the number of frames significantly raises

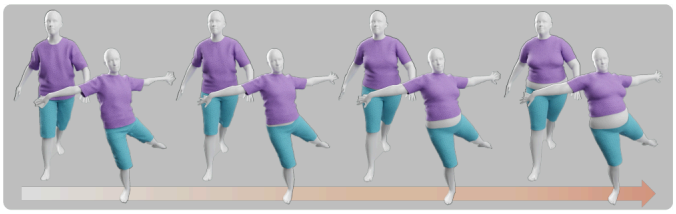

Fig. 9. Generalization capability of cloth dynamics modeling across different body types. The clothing template can be adapted to different body types through dynamics modeling.

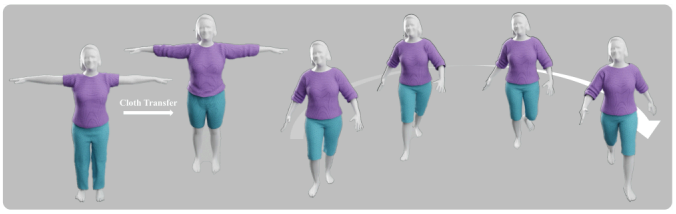

Fig. 10. Illustrative diagram of clothing transfer application. The avatar generation framework for easy integration into virtual try on applications with corresponding animations.

TABLE IV
ABLATION STUDY ON THE INFLUENCE OF INPUT NUMBER ON AVATAR RECONSTRUCTION ACCURACY.

N-Images	P2S↓	Chamfer↓	Normal↓
1	2.253	1.915	0.137
3	1.832	1.783	0.121
5	1.718	1.636	**0.115**
7	**1.672**	**1.601**	0.118

the complexity of feature fusion. When the network depth and capacity remain fixed, it becomes difficult to effectively extract useful information from a larger, potentially redundant or inconsistent input space.

Cloth Separation. To extract the clothing template, we jointly classify clothing vertices by leveraging multi-frame semantic segmentation results. Tab. V reports the effect of the number of input frames on clothing segmentation accuracy. Specifically, we evaluate three settings: sampling one or three frames from the images used for avatar reconstruction, and using all five reconstruction frames. In general, achieving an mIoU above 80% and an mAcc above 85% indicates reliable segmentation performance. As shown in the results, using all five reconstruction frames achieves reliable clothing segmentation accuracy, enabling effective extraction of the clothing template.

Dynamics modeling. In the cloth dynamics modeling module, global optimization is used to address unnatural transitions in local predictions and refine clothing details. Tab. VI demonstrates the impact of different optimization terms on the dynamics of clothing. From the ε_g metric, it can be seen that the temporal and spatial terms bring the deformation of the clothing closer to the ground truth simulation. Although the separation term increases energy consumption ε_e, it achieves a lower collision rate between clothes ε_c, which better reflects realistic clothing behavior.

TABLE V
ABLATION STUDY ON THE EFFECT OF THE NUMBER OF INPUT FRAMES ON CLOTHING SEGMENTATION ACCURACY.

N-Frames	mIoU(%) ↑	mAcc(%) ↑
1	45.5	58.6
3	72.8	79.1
5	**80.7**	**86.3**

TABLE VI
ABLATION STUDY OF THE DYNAMICS MODELING ON T-SHIRTS, SHOWING THE IMPACT OF OPTIMIZATION TERMS ON OVERALL PERFORMANCE.

	$\varepsilon_g \downarrow$	$\varepsilon_e \downarrow$	$\varepsilon_c \downarrow$
w/o $E_{spatial}$	0.064	0.050	0.214
w/o $E_{temporal}$	0.067	**0.045**	0.178
w/o $E_{separation}$	0.056	0.047	0.342
All	**0.054**	0.051	**0.156**

D. Application

We extract clothing templates and inner body models from the reconstructed avatar for independent cloth dynamics modeling. Our cloth dynamics modeling module allows clothing templates of various sizes to adapt to the inner body model. Hence, our framework can be easily applied to clothing transfer applications. As shown in Fig. 10, we can easily transfer other clothing templates to human avatars and generate realistic human animations. Moreover, the independent inner body models and garment templates allow us to edit their dimensions and shapes separately.

V. CONCLUSION

We proposed a novel framework for generating clothing-editable human avatars, named CHA. Our approach divides modeling and driving into independent phases. We employ an avatar reconstruction module that extracts monocular temporal human information to reconstruct the complete human surface. We designed a physics-based clothing dynamics modeling module with clothing generalization capabilities, addressing the issue of poor dynamic performance in avatar animations. Our method showcases superior clothing dynamics, sharper individual features, and distinct clothing layer structures. Furthermore, this method does not require implicit fitting for each subject or multiview data inputs. The extraction of clothing templates and inner body models enables the human avatars to be easily applied to virtual try-on tasks.

Limitations. Distinguishing the original clothing details from dynamic details is challenging, leading the cloth recovery module to inadvertently erase some characteristic features when removing dynamic details. Moreover, due to their bias toward body-conforming reconstructions, cloth recovery algorithms often fail to recover the canonical loose-fitting template structure of dresses. Finally, while our method supports different garment parameter configurations, such parameters are manually specified and not adaptively inferred from the input data. Automatically estimating garment properties from visual observations is a promising avenue for future work.

REFERENCES

[1] M. Sun, D. Yang, D. Kou, Y. Jiang, W. Shan, Z. Yan, and L. Zhang, "Human 3d avatar modeling with implicit neural representation: A brief survey," in *2022 14th international conference on signal processing systems (ICSPS)*. IEEE, 2022, pp. 818–827.

[2] Q. Sun, Y. Xiao, J. Zhang, S. Zhou, C.-S. Leung, and X. Su, "A Local Correspondence-Aware Hybrid CNN-GCN Model for Single-Image Human Body Reconstruction," in *IEEE Transactions on Multimedia*, vol. 25, pp. 4679–4690, 2023.

[3] H. Bertiche, M. Madadi, and S. Escalera, "Cloth3d: clothed 3d humans," in *European Conference on Computer Vision*. Springer, 2020, pp. 344–359.

[4] X. Zuo et al., "SparseFusion: Dynamic Human Avatar Modeling From Sparse RGBD Images," in *IEEE Transactions on Multimedia*, vol. 23, pp. 1617–1629, 2021.

[5] X. Yang, X. Chen, D. Gao, S. Wang, X. Han, and B. Wang, "Have-fun: Human avatar reconstruction from few-shot unconstrained images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 742–752.

[6] Z. Xu, W. Chang, Y. Zhu, L. Dong, H. Zhou, and Q. Zhang, "Building High-Fidelity Human Body Models From User-Generated Data," in *IEEE Transactions on Multimedia*, vol. 23, pp. 1542–1556, 2021.

[7] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: a skinned multi-person linear model," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 1–16, 2015.

[8] C. H. Esteban and F. Schmitt, "Silhouette and stereo fusion for 3d object modeling," *Computer Vision and Image Understanding*, vol. 96, no. 3, pp. 367–392, 2004.

[9] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 8, pp. 1362–1376, 2009.

[10] D. Vlasic, I. Baran, W. Matusik, and J. Popović, "Articulated mesh animation from multi-view silhouettes," in *Acm Siggraph 2008 papers*, 2008, pp. 1–9.

[11] D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik, "Dynamic shape capture using multi-view photometric stereo," in *ACM SIGGRAPH Asia 2009 papers*, 2009, pp. 1–11.

[12] C. Guo, T. Jiang, X. Chen, J. Song, and O. Hilliges, "Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 858–12 868.

[13] A. Karthikeyan, R. Ren, Y. Kant, and I. Gilitschenski, "AvatarOne: Monocular 3D Human Animation," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2024, pp. 3635–3645.

[14] T. Jiang, X. Chen, J. Song, and O. Hilliges, "Instantavatar: Learning avatars from monocular video in 60 seconds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16922–16932.

[15] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman, "Humannerf: Free-viewpoint rendering of moving people from monocular video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, 2022, pp. 16 210–16 220.

[16] K. Shen, C. Guo, M. Kaufmann, J. J. Zarate, J. Valentin, J. Song, and O. Hilliges, "X-avatar: Expressive human avatars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 911–16 921.

[17] X. Chen, T. Jiang, J. Song, J. Yang, M. J. Black, A. Geiger, and O. Hilliges, "gdna: Towards generative detailed neural avatars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 427–20 437.

[18] S. Saito, J. Yang, Q. Ma, and M. J. Black, "Scanimate: Weakly supervised learning of skinned clothed avatar networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2886–2897.

[19] S. Lin, Z. Li, Z. Su, Z. Zheng, H. Zhang, and Y. Liu, "Layga: Layered gaussian avatars for animatable clothing transfer," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[20] Z. Li, Z. Zheng, L. Wang, and Y. Liu, "Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 711–19 722.

[21] Z. Li, Z. Zheng, Y. Liu, B. Zhou, and Y. Liu, "Posevocab: Learning joint-structured pose embeddings for human avatar modeling," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11.

[22] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[23] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023.

[24] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174.

[25] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4460–4470.

[26] T. He, Y. Xu, S. Saito, S. Soatto, and T. Tung, "Arch++: Animation-ready clothed human reconstruction revisited," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 046–11 056.

[27] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2304–2314.

[28] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black, "Econ: Explicit clothed humans optimized via normal integration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 512–523.

[29] Y. Xiu, J. Yang, D. Tzionas, and M. J. Black, "ICON: Implicit Clothed humans Obtained from Normals," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Mar. 2022, pp. 13286–13296.

[30] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll, "Multi-garment net: Learning to dress 3d people from images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5420–5430.

[31] Y.-H. Kwon, J. H. Yoon, and M.-G. Park, "Text2Avatar: Articulated 3D Avatar Creation With Text Instructions," in *IEEE Transactions on Multimedia*, vol. 27, pp. 3797–3806, 2025.

[32] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung, "Arch: Animatable reconstruction of clothed humans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3093–3102.

[33] T. Liao, X. Zhang, Y. Xiu, H. Yi, X. Liu, G.-J. Qi, Y. Zhang, X. Wang, X. Zhu, and Z. Lei, "High-fidelity clothed avatar reconstruction from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8662–8672.

[34] S. Peng, J. Dong, Q. Wang, S. Zhang, Q. Shuai, X. Zhou, and H. Bao, "Animatable neural radiance fields for modeling dynamic human bodies," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 314–14 323.

[35] D. Terzopoulos, J. Platt, A. Barr, and K. Fleischer, "Elastically deformable models," in *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, 1987, pp. 205–214.

[36] Y. Li, T. Du, K. Wu, J. Xu, and W. Matusik, "Diffcloth: Differentiable cloth simulation with dry frictional contact," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 1, pp. 1–20, 2022.

[37] J. Liang, M. Lin, and V. Koltun, "Differentiable cloth simulation for inverse problems," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[38] T. Kim, "A finite element formulation of baraff-witkin cloth," in *Computer Graphics Forum*, vol. 39, no. 8. Wiley Online Library, 2020, pp. 171–179.

[39] T. Liu, A. W. Bargteil, J. F. O'Brien, and L. Kavan, "Fast simulation of mass-spring systems," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, pp. 1–7, 2013.

[40] M. Macklin, M. Müller, and N. Chentanez, "Xpbd: position-based simulation of compliant constrained dynamics," in *Proceedings of the 9th International Conference on Motion in Games*, 2016, pp. 49–54.

[41] Z. Lahner, D. Cremers, and T. Tung, "Deepwrinkles: Accurate and realistic clothing modeling," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 667–684.

[42] C. Patel, Z. Liao, and G. Pons-Moll, "Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7365–7375.

[43] I. Santesteban, M. A. Otaduy, and D. Casas, "Learning-based animation of clothing for virtual try-on," in *Computer Graphics Forum*, vol. 38, no. 2. Wiley Online Library, 2019, pp. 355–366.

[44] T. Y. Wang, T. Shao, K. Fu, and N. J. Mitra, "Learning an intrinsic garment space for interactive authoring of garment animation," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, pp. 1–12, 2019.

[45] R. Li, B. Guillard, E. Remelli, and P. Fua, "Dig: Draping implicit garment over the human body," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 2780–2795.

[46] I. Santesteban, M. A. Otaduy, and D. Casas, "Snug: Self-supervised neural dynamic garments. ieee," in *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, no. 5, 2022, p. 9.

[47] A. Grigorev, M. J. Black, and O. Hilliges, "Hood: Hierarchical graphs for generalized modelling of clothing dynamics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 965–16 974.

[48] R. Chen, L. Chen, and S. Parashar, "Gaps: Geometry-aware, physics-based, self-supervised neural garment draping," in *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 116–125.

[49] L. De Luigi, R. Li, B. Guillard, M. Salzmann, and P. Fua, "Drapenet: Garment generation and self-supervised draping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1451–1460.

[50] I. Santesteban, N. Thuerey, M. A. Otaduy, and D. Casas, "Self-supervised collision handling via generative 3d garment models for virtual try-on," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 763–11 773.

[51] H. Bertiche, M. Madadi, and S. Escalera, "Pbns: Physically based neural simulation for unsupervised garment pose space deformation," *ACM Transactions on Graphics*, vol. 40, no. 6, p. 198, 2021.

[52] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 975–10 985.

[53] A. Waswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[54] R. Khirodkar, T. Bagautdinov, J. Martinez, S. Zhaoen, A. James, P. Selednik, S. Anderson, and S. Saito, "Sapiens: Foundation for human vision models," in *European Conference on Computer Vision*. Springer, 2025, pp. 206–228.

[55] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proceedings of the fourth Eurographics symposium on Geometry processing*, vol. 7, no. 4, 2006.

[56] S. Jia, X. Tang, and H. Pan, "Fast mesh simplification algorithm based on edge collapse," in *Intelligent Control and Automation: International Conference on Intelligent Computing, ICIC 2006 Kunming, China, August 16–19, 2006*. Springer, 2006, pp. 275–286.

[57] J. Montes, B. Thomaszewski, S. Mudur, and T. Popa, "Computational design of skintight clothing," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 105–1, 2020.

[58] X. Zhu, T. Liao, X. Zhang, J. Lyu, Z. Chen, Y. Wang, K. Guo, Q. Cao, S. Z. Li, and Z. Lei, "Mvp-human dataset for 3-d clothed human avatar reconstruction from multiple frames," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 5, no. 4, pp. 464–475, 2023.

[59] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5442–5451.

[60] D. Xiang, F. Prada, C. Wu, and J. Hodgins, "Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 322–332.

[61] G. Moon, H. Nam, T. Shiratori, and K. M. Lee, "3d clothed human reconstruction in the wild," in *European conference on computer vision*. Springer, 2022, pp. 184–200.

[62] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.

[63] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8739–8748.

[64] Z. Zheng, H. Huang, T. Yu, H. Zhang, Y. Guo, and Y. Liu, "Structured local radiance fields for human avatar modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 893–15 903.

[65] B. Liu, J. Lei, B. Peng, Z. Zhang, J. Zhu, and Q. Huang, "Advancing Generalizable Occlusion Modeling for Neural Human Radiance Field," in *IEEE Transactions on Multimedia*, vol. 27, pp. 1362–1373, 2025.

[66] S. Wang, K. Schwarz, A. Geiger, and S. Tang, "Arah: Animatable volume rendering of articulated human sdfs," in *European conference on computer vision*. Springer, 2022, pp. 1–19.

[67] J. Long, K. Burns, and J. Yang, "Cloth modeling and simulation: a literature survey," in *Digital Human Modeling: Third International Conference, ICDHM 2011, Held as Part of HCI International 2011, Orlando, FL, USA July 9-14, 2011. Proceedings 3*. Springer, 2011, pp. 312–320.

[68] M. Atzmon and Y. Lipman, "Sal: Sign agnostic learning of shapes from raw data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2565–2574.

[69] L. Hu, H. Zhang, Y. Zhang, B. Zhou, B. Liu, S. Zhang, and L. Nie, "Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 634–644.

[70] Y. Xiu, Y. Ye, Z. Liu, D. Tzionas, and M. J. Black, "Puzzleavatar: Assembling 3d avatars from personal albums," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 6, pp. 1–15, 2024.

[71] Z. Weng, Z. Wang, and S. Yeung, "Zeroavatar: Zero-shot 3d avatar generation from a single image," *arXiv preprint arXiv:2305.16411*, 2023.

[72] Z. Dong, C. Guo, J. Song, X. Chen, A. Geiger, and O. Hilliges, "Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 470–20 480.

[73] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[74] Y. Li, H. Chen, E. Larionov, N. Sarafianos, W. Matusik, and T. Stuyck, "Diffavatar: Simulation-ready garment optimization with differentiable simulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4368–4378.

[75] Z. Su, L. Hu, S. Lin, H. Zhang, S. Zhang, J. Thies, Y. Liu, "Caphy: Capturing physical properties for animatable human avatars," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14150–14160.

[76] Y. Zheng et al., "PhysAvatar: Learning the Physics of Dressed 3D Avatars from Visual Observations," in *Proceedings of the European conference on computer vision (ECCV)*, vol. 15095, 2025, pp. 262–284.

[77] D. Vlasic, I. Baran, W. Matusik, and J. Popović, "Articulated mesh animation from multi-view silhouettes," *ACM Transactions on Graphics (TOG)*. vol. 27, no. 3, Art. no. 3, Aug. 2008.

[78] G. Pavlakos et al., "Expressive Body Capture: 3D Hands, Face, and Body From a Single Image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2019, pp. 10967–10977.