IF-Garments: Reconstructing Your Intersection-Free Multi-Layered Garments from Monocular Videos

Anonymous Authors

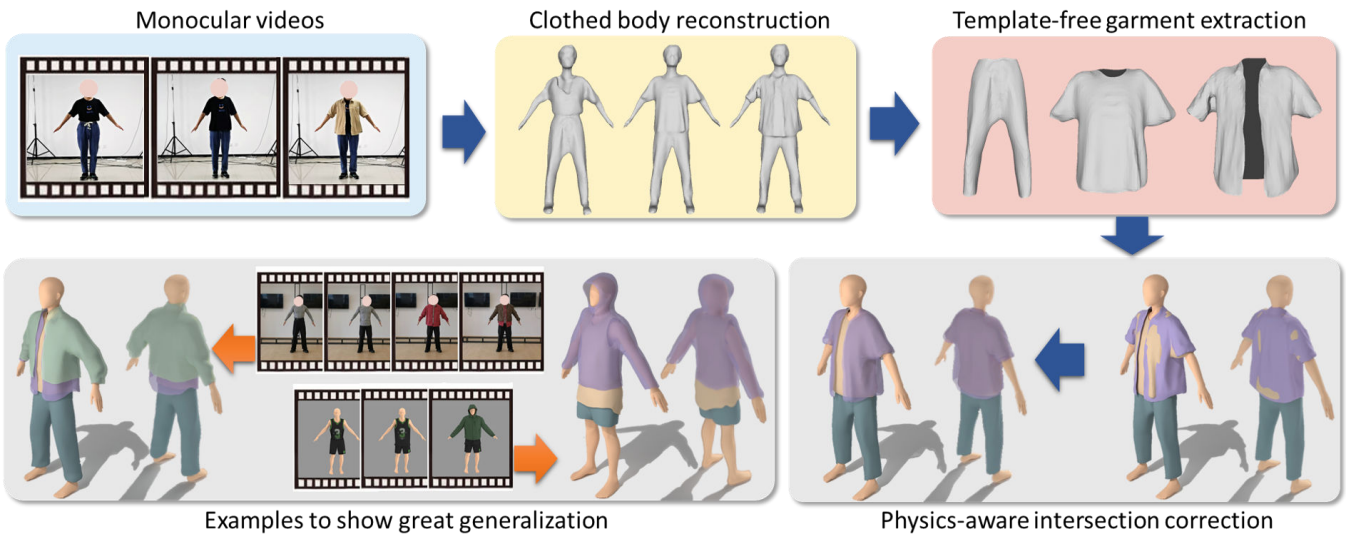


Figure 1: We propose IF-Garments to reconstruct Intersection-Free Garments from monocular videos. The template-free garment extraction shows great generalization while the physics-aware correction robustly eliminates inter-layer penetration.

ABSTRACT

Reconstructing garments from monocular videos has attracted considerable attention as it provides a convenient and low-cost solution for clothing digitization. In reality, people wear clothing with countless variations and multiple layers. Existing studies attempt to extract garments from a single video. They either behave poorly in generalization due to reliance on limited clothing templates or struggle to handle the intersections of multi-layered clothing leading to the lack of physical plausibility. Besides, there are inevitable and undetectable overlaps for a single video that hinder researchers from modeling complete and intersection-free multi-layered clothing. To address the above limitations, in this paper, we propose a novel method to reconstruct multi-layered clothing from multiple monocular videos sequentially, which surpasses existing work in generalization and robustness against penetration. For each video, neural fields are employed to implicitly represent the clothed body, from which the meshes with frame-consistent structures are explicitly extracted. Next, we implement a template-free method for extracting a single garment by back-projecting the image segmentation labels of different frames onto these meshes. In this way, multiple garments can be obtained from these monocular videos and then aligned to form the whole outfit. However, intersection always occurs due to overlapping deformation in the real world and perceptual errors for monocular videos. To this end, we innovatively introduce a physics-aware module that combines neural fields with a position-based simulation framework to fine-tune the penetrating vertices of garments, ensuring robustly intersection-free. Additionally, we collect a mini dataset with fashionable garments to evaluate the quality of clothing reconstruction comprehensively. The code and data will be open-sourced if this work is accepted.

CCS CONCEPTS

• **Information systems** → **Multimedia content creation**; • **Computing methodologies** → **Shape modeling**.

KEYWORDS

Clothing reconstruction, Clothing simulation

1 INTRODUCTION

Digitalization of clothing holds significant importance in livestream sales, virtual try-ons, and entertainment. Recent work[18, 24, 45] has shown progress in extracting garments from monocular images or videos, which becomes highly convenient and accessible for general commerce. In daily life, people wear multiple layers of garments varying in different styles, which indeed poses significant challenges for high-quality reconstruction.

Unpublished working draft. Not for distribution.

ACM MM, 2024, Melbourne, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

We start the discussion by reconstructing a single garment with a monocular image or video, which can be categorized into template-based and template-free approaches. Template-based approaches[12, 15, 24, 29, 45, 48, 51] choose and deform a pre-designed garment template to fit observations. They suffer from a limited set of templates, lacking generalization. Template-free methods[17, 18, 53] first reconstruct the 3D clothed body, then separate the clothing from the body using 2D segmentation of garments. To obtain the clothed body, one straightforward idea is to use scanning devices[31, 53, 56], which needs costly manual post-processing. With the parametric human body models[33, 43], some studies[4, 11, 17, 18, 23] consider clothing as the offset of body vertices, which is feasible for most garments as they usually adhere to the human body in rest pose. For multi-layered outfits, existing work aims to extract multiple pieces of clothing from a single monocular image or video simultaneously. They either treat all clothing as a single entity[17, 18] or support only two pieces of clothing[24, 29, 45, 51]. Those[2, 15, 39] attempting to overcome three or more layers of clothing suffer from distortions caused by overlap regarding inner garments. Besides, they suffer from occlusion in a single video so the inner layer of outfits is incomplete or in the wrong type.

Due to the inevitable occlusion, it is an ill-posed problem to recover multi-layered clothing from a single video, but is possible to extract an uncovered garment. Further, multiple garments can be obtained from corresponding videos and then aligned to form a multi-layered outfit. Therefore, we propose a novel methodology to sequentially reconstruct each layer of clothing from multiple monocular videos. As shown in Figure 2, we ask the actor to remove occlusion for the target garment in each video and rotate slowly to provide approximate multi-view information Our full approach is divided into three parts. First, a neural Signed Distance Field (SDF)[23, 42, 50] is leveraged to represent the clothed body and extract the meshes with consistent topology via marching cube[34]. Second, a template-free approach is implemented by back-projecting image segmentation labels of the garment from different frames onto these meshes' vertices. The labeled vertices are then gathered to form the garment, resulting in multiple garments from corresponding videos. Finally, it is crucial to combine these garments into a multi-layered outfit. However, it indeed introduces intersections due to the following reasons: i) in the real world, the movement and overlap of clothing can cause deformation; ii) for reconstructing, results obtained from different videos inevitably contain misaligned errors due to monocular depth ambiguity. To robustly eliminate inter-layer penetration, a physics-aware module is proposed with a novel pipeline to fine-tune garments from the outer to the inner. Concretely, given an SDF of the clothed body of the inner garment layer, we query the signed distance of the outer layer's vertices and push out those vertices with negative signs along the intersecting direction. To further ensure physically plausible deformation, the SDF-based penetration handling is implemented in a position-based simulation framework[36] with carefully devised physics constraints. In addition, existing datasets[4, 18, 23] have limited clothing variety, making it difficult to adequately evaluate generalization and also struggle to meet the requirements for multi-layered clothing reconstruction. Thus, we create *mini-IFG*, a small dataset with 23 videos collected from both physics simulation and the real world for comprehensive evaluation.

Table 1: Comparison among ours and existing works. Our method supports multi-layer clothing reconstruction without intersection and is independent of garment templates.

Research	Layers	Template-free	Intersection-free
PERGAMO[12]	1	✗	-
SCARF[18]	1	✓	-
DELTA[17]	1	✓	-
REC-MV[45]	2	✗	✗
BCNet[24]	2	✗	✗
Li *et al.*[29]	2	✗	✗
MulayCap[51]	2	✗	✓
SMPLict[15]	≥3	✗	✗
ClothWild[39]	≥3	✗	✗
LGN[2]	≥3	✗	✓
Ours	≥3	✓	✓

Briefly, this paper represents the first attempt to reconstruct intersection-free and complete multi-layered clothing from several monocular videos. In Table 1, we make a comparison of studies related to clothing reconstruction from monocular images or videos. Our method surpasses existing work in terms of the number of clothing layers, generalization, and robustness against penetration. Experiments demonstrate that our method can confidently reconstruct challenging garments and robustly eliminate intersections.

The contributions of this paper can be summarized as follows:

- A template-free approach with great generalization for extracting a complete garment.
- A physics-aware module ensuring multi-layered clothing intersection-free with excellent robustness and quality.
- A small dataset containing 23 self-rotating videos of actors wearing fashionable garments.

2 RELATED WORK

We briefly review work related to recovering clothing from images or videos from the following three aspects.

Clothed Body Reconstruction. Some statistical human body models have been proposed by fitting a function with shape and pose to real 3D scans[5, 33, 41, 43]. It has made great progress [26, 27, 30, 52, 57, 58] in predicting body parameters from images and videos, which lays the foundation for clothed body reconstruction. PIFu[46] and PIFuHD[47] have pioneered the extraction of pixel-aligned spatial features from images and mapping them to implicit fields to reconstruct people in arbitrary poses and clothing. Follow-up methods[19–21, 54, 55, 59] then introduce parametric models as 3D features to condition implicit fields. These methods all require expensive 3D scans for supervision. Unlike single images, videos contain richer perspectives. Some studies[3, 4, 16, 23] treat clothing as offsets of body vertices. They establish a clothed body in canonical space based on parameterized models[33, 43], then learn mappings from the canonical space to the posed space of video frames to recover clothing. In self-rotating videos, clothing has minimal deformation. Thus, we follow SelfRecon[23] to learn neural fields from self-rotating videos to obtain the clothed body.

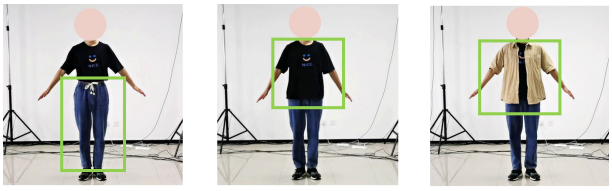


Figure 2: Videos with the corresponding uncovered garment. The target garment is marked with a green rectangle.

Garments Extraction. Researchers have already attempted to construct various clothing templates[8, 11, 24, 39, 51, 60]. Subsequent works fine-tune these templates to match observations. MGN[11] utilizes a large-scale clothing dataset to develop a per-category parametric model. BCNet[24] generates a rough template, which is subsequently enhanced with surface details through a displacement network. REC-MV[45] simultaneously optimizes the explicit feature curves and the implicit field of the garments, resulting in superior dynamic garment surfaces. However, the limited templates are insufficient to cover the vast diversity of clothing. Template-free methods extract garments from the clothed body, which is achieved by inverse mapping 2D segmentation to the 3D space. SCARF[18] represents the clothing using NeRF[38], which results in high-quality rendering but low-quality geometry. Similar to our template-free method, Xiang *et al.*[53] use 140 synchronized cameras to extract high-fidelity clothing. In contrast, our back-projection scheme leverages the geometric consistency of meshes across different frames, thus requiring only one camera.

Multi-Layered Clothing Reconstruction. Some studies learn a generative clothing model with neural distance field[13, 37, 42] from a 3D clothing dataset[15, 39]. They can decode each garment from the latent space and align them but lack consideration for penetration. LGN[2] leverages SDF[42] to propose a garment indication field to handle penetration but overlooks clothing deformation. In comparison, we implement a physics-aware module that combines SDF and physics constraints to eliminate inter-layer penetration while preserving natural non-rigid deformations.

3 METHODOLOGY

In Figure 3, we present an overview of our method, IF-Garments, which aims to reconstruct intersection-free multi-layered clothing from monocular videos faithfully. Our key insight lies in resourcefully leveraging the SDF's geometric and physical characteristics, employed in clothing extraction and inter-layer penetration correction respectively. Specifically, we first follow SelfRecon[23] to learn a neural SDF in canonical space to reconstruct the clothed body, which can be mapped to posed space by a pose-conditioned deformation field (Section 3.1). Next, we back-project the segmentation image labels from different viewpoints onto the vertices of the corresponding posed mesh to extract garments (Section 3.2). Finally, all of these garments are aligned in canonical space, and the physics-aware module solves the intersections between them (Section 3.3).

3.1 Clothed Body Reconstruction

Given a self-rotating video with N frames, we adopt VideoAvatar[4] to generate the camera intrinsic π, and SMPL[33] parameters of the initial shape β, and per-frame's pose $\{\theta_i | i \in 1, \ldots, N\}$ and translation $\{t_i | i \in 1, \ldots, N\}$.

Canonical Representation. The clothed body $\mathcal{S}_\eta$ in canonical space with an A-pose is represented as the zero-level-set of an neural SDF[42], which is parametrized by a Multi-Layer Perceptron (MLP) ϕ with learnable weights η:

$$\mathcal{S}_\eta = \{\mathbf{x} \in \mathbb{R}^3 | \phi(\mathbf{x}; \eta) = 0\}, \tag{1}$$

where the mesh of the clothed body $\mathcal{M}$ is extracted by marching cube[34].

Deformation Field. To map the clothed body in canonical space to posed space to match supervision from the video, we decompose the deformation field $\mathcal{D}$ into skinning transformation $\mathcal{W}$ and non-rigid deformation d. $\mathcal{W}$ ensures that the garment's surface deforms with the body's large-scale motions[20], which takes θ_i as the parameter and is pre-computed as described in [23]. An MLP d with learnable weights ψ models the fine-grained changes. In i-th frame, d is conditioned by an optimizable $\mathbf{h}_i$ variable to apply deformations to points in the canonical space. By compositing $\mathcal{W}$ and d, we get the final deformation field $\mathcal{D} = \mathcal{W}(d(\cdot))$. It takes $\mathbf{h}_i$ and θ_i as input and transforms canonical points to the i-th frame posed space. We train $\mathcal{S}$ and $\mathcal{D}$ in the same manner as SelfRecon[23].

For brevity of description, we use $\mathcal{D}_i$ to denote i-th frame's deformation field, $\mathcal{S}_i$ for i-th frame's zero-level-set $\mathcal{D}_i(\mathcal{S}_\eta)$, and $\mathcal{M}_i$ for the mesh extracted from $\mathcal{S}_i$ via marching cube[34].

3.2 Template-Free Garment Extraction

In monocular self-rotating videos, the actor rotates slowly, providing approximate multi-view information about the clothing. Thanks to the deformation field discussed in Section 3.1, we can obtain clothed body mesh $\mathcal{M}_i$ in posed space. Especially, $\mathcal{M}$ and $\mathcal{M}_i$ in all frames share the same topology. Inspired by this, we propose a template-free clothing extraction method by back-projecting the garment's segmentation labels of video frames onto corresponding $\mathcal{M}_i$. These labels are shared across all of the canonical and posed meshes and aggregate to form the clothing $\mathcal{G}$ (see Figure 3(a)).

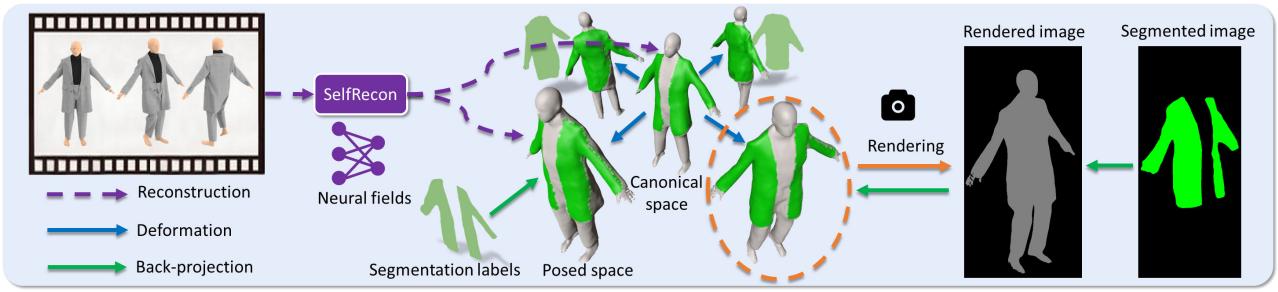
Back-Projection. Back-projection is realized by projecting 3D vertices forward onto 2D pixels to assign garment labels to $\mathcal{M}_i$. The parameters estimated from the video regarding π, θ_i, and t_i are referenced in the camera coordinate[4]. For i-th frame, we obtain $\mathcal{M}_i^{'}$ by applying the deformation field and translation to $\mathcal{M}$:

$$\mathcal{M}_i^{'} = \mathcal{D}_i(\mathcal{S}_\eta) + t_i = \mathcal{M}_i + t_i. \tag{2}$$

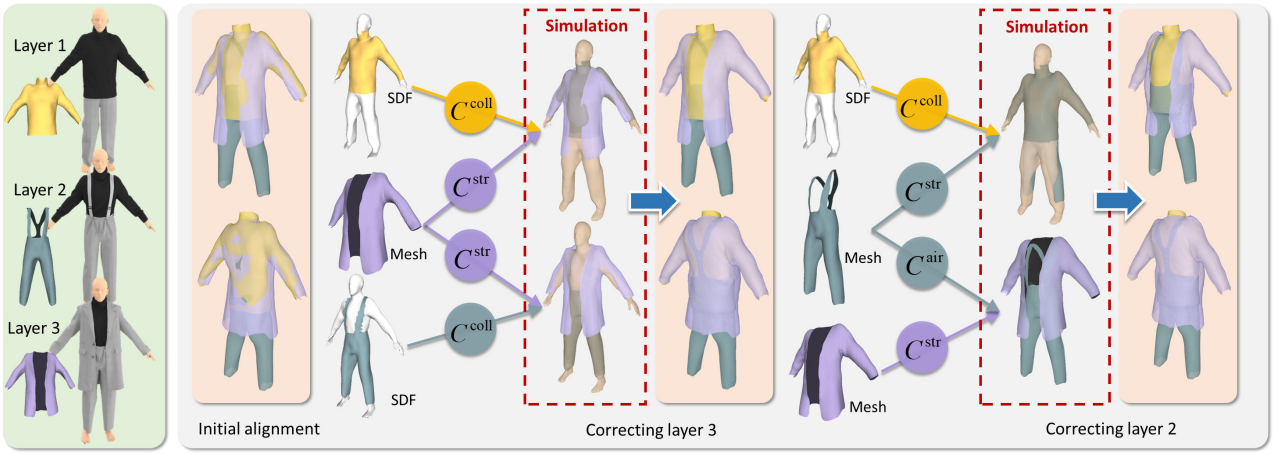
Given a vertex $P = [X, Y, Z]^T \in \mathcal{M}_i^{'}$, we render it onto the image plane as $p = [u, v]$ by the perspective projection:

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = KP = \begin{bmatrix} \alpha_u & 0 & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \tag{3}$$

where λ is the depth factor and equal to Z, $[u, v]$ are the 2D position in the image, $[u_0, v_0]$ and $[\alpha_u, \alpha_v]$ are the center and focal length of the camera intrinsic π respectively. However, Equation (3)



(a) Extracting an unoccluded garment from a monocular video.



(b) The physics-aware module to obtain intersection-free clothing.

Figure 3: Overview of IF-Garments. (a) Following SelfRecon[23], we employ neural fields to reconstruct the clothed body from a monocular video, which can be mapped into the posed space by the deformation field. The garment is extracted by back-projecting the segmentation labels to the posed clothed body's vertices from multiple viewpoints. (b) The physics-aware module involves a novel pipeline to eliminate penetration from outer to inner in canonical space, accomplished with carefully designed constraints of C^{str}, C^{coll}, and C^{air}.

ignores the visibility of vertices, leading to multiple vertices being projected onto the same pixel. To address this issue, we utilize the z-buffer algorithm from OpenGL[49], where the visibility of vertices is determined based on λ. Then, visible vertices will acquire the corresponding label on the segmentation image. Finally, $\mathcal{G}$ is composed of these labeled vertices. To mitigate the noise caused by segmentation errors, post-processing is employed, including isolated elements removal, hole filling, and Laplacian smoothing[14]. Depending on the complexity of the garment, we typically back-project 4 to 8 frames and fuse their labels.

3.3 Physics-aware Module

The clothed body in canonical space undergoes no rigid transformation, where all garments are aligned initially. However, penetration is inevitable due to the following reasons: i) the way clothing is worn varies slightly in each video to eliminate occlusion, and each additional garment will cause deformation to existing ones; b) perceptual errors arise due to the depth ambiguity of monocular videos. It poses a significant challenge for solving such penetration due to the lack of ground truth. Inspired by collision detection[6, 22, 35] in computer graphics, the neural SDF is employed to handle intersections. To ensure physical plausibility, we conscientiously design 3 physics constraints in a position-based simulation framework[7, 36, 40]. Due to the undetectable overlap, it is impossible to model multi-layered clothing that completely adheres to the real world. However, we strive to address severe penetration through deformation as physically as possible.

3.3.1 Simulation Framework. Following [36], we establish physics constraints between vertices and solve them with the Gauss-Seidel method. In each time step, the positions of the vertices are projected onto each constraint manifold along the constraint gradient. Due to space limitations, the solving pipeline can be found in the supplementary materials. Here, we primarily discuss the physics constraints devised in our work.

Given a vertex $\mathbf{p}$ and the vertex $\bar{\mathbf{p}}$ forming an edge with $\mathbf{p}$, we define the stretch constraint as:

$$C^{\text{str}}(\mathbf{p}, \bar{\mathbf{p}}) = \| \mathbf{p} - \bar{\mathbf{p}} \| - L, \tag{4}$$

where $\|\cdot\|$ means Euclidean norm, and L is the rest length. For a garment mesh $\mathcal{G}$, C^{str} accounts for the non-rigid deformation and is applied to all vertices.

3.3.2 SDF-Based Collision Detection. We query the penetration status of a point in the SDF, including the signed distance (penetration depth) and the gradient (penetration direction), which provides robustness for handling intersections. Given a query point $\mathbf{p}$ and an SDF ϕ, we define the collision constraint as:

$$C^{\text{coll}}(\mathbf{p}) = \phi(\mathbf{p}) - \alpha \geq 0, \tag{5}$$

where α is a small positive value to enhance robustness. When $C^{\text{coll}}(\mathbf{p})$ is not satisfied, penetration occurs. Then $\mathbf{p}$ is projected to $\mathbf{p}'$:

$$\mathbf{p}' = \mathbf{p} - \nabla\phi(\mathbf{p})\phi(\mathbf{p}). \tag{6}$$

3.3.3 Multi-Layerd Intersection Handling. The clothing with N layers of garments are aligned initially in canonical space and intersections may occur between any two layers. Fortunately, we have obtained the SDF ϕ for each clothed body in canonical space as described in Section 3.1, allowing us to leverage the SDF of a layer to correct penetrating vertices of other layer's mesh according to Equation (5) and Equation (6). Considering real-world scenarios, the innermost layer of clothing is usually closely fitted to the body, which is suitable to serve as the reference, hence penetration correction is performed from the outermost layer towards the inner layers. Since the goal is to eliminate penetration between garments, to avoid the influence of other parts of the clothed body, we apply a mask $\mathcal{R} \in \mathbb{R}$ to the SDF, which is determined by the bounding box $\mathcal{X} \in \mathbb{R}^{2\times3}$ of the corresponding garment extracted from this clothed body. If a query point is not in $\mathcal{X}$, we discard its collision constraint event it is penetrated. However, it is insufficient to rely solely on SDF to eliminate penetration. For a three-layer clothing, we first eliminate the penetration of the mesh of layer 3 based on the SDFs of layers 1 and 2. When handling the penetration of layer 2 based on the SDF of layer 1, penetration may occur again between layers 2 and 3. We discuss this further in Section 4.3.1. To address this problem, we devise the air constraint C^{air} to keep the gap between two adjacent intersection-free layers (represented as $\mathcal{G}$ and $\hat{\mathcal{G}}$). Given $\mathbf{p} \in \mathcal{G}$, we have:

$$\begin{aligned} C^{\text{air}}(\mathbf{p}, \hat{\mathcal{G}}) &= C^{\text{str}}(\mathbf{p}, \hat{\mathbf{q}}) \\ \hat{\mathbf{q}} &= \arg\min_{\mathbf{q}\in\hat{\mathcal{G}}} ||\mathbf{p} - \mathbf{q}||, \end{aligned} \tag{7}$$

where $\mathbf{p}$ and $\hat{\mathbf{q}}$ are uniquely corresponding and pre-computed at the beginning of the simulation.

The complete procedure is outlined in Algorithm 1, where N is the number of layers, $\mathcal{G}^i$ is the garment mesh of i-th layer, and ϕ^j is the clothed body SDF of j-th layer. Lines 2 to 7 correct the i-th layer, while lines 8 to 12 are implemented to ascertain that a gap is maintained between $(i+1)$-th layer and i-th layer.

Algorithm 1 Handling intersections from outside to inside.

```
1:  for i = N, N − 1, ..., 2 do
2:      for all vertices p ∈ G^i do
3:          for j = 1, 2 ..., i − 1 do
4:              solve collision R^j C^coll(p, φ^j)        ▷ Equation (5)
5:          end for
6:          solve stretch C^str(p, p̄)                    ▷ Equation (4)
7:      end for
8:      if i < N then
9:          for all vertices p ∈ G^{i+1} do
10:             solve air C^air(p, G^i)                  ▷ Equation (7)
11:         end for
12:     end if
13: end for
```

4 EXPERIMENTS

Datasets. People Snapshot[4] is a widely recognized dataset including monocular self-rotating videos[18, 23, 29, 45]. However, it only contains a few types of tight-fitting clothing, which is insufficient to support a comprehensive evaluation. Therefore, we create *mini-IFG* that involves self-captured sequences (*mini-IFG-real*) and synthetic data (*mini-IFG-sim*) in a popular clothing design software, Style3D[1]. *mini-IFG* consists of 23 videos of 8 subjects with fashionable clothing. In each video, the actor rotates slowly in front of the camera while ensuring an uncovered garment as the target. Such measures allow for the evaluation of reconstructing both single-garment and multi-layered clothing.

Baselines. We compare with state-of-the-art (SOTA) works including video-based methods of SCARF[18] and REC-MV[45] and image-based of BCNet[24], SMPLicit[15], and ClothWild[39]. Since REC-MV doesn't release the model for detecting garment feature lines, we only reproduce *People Snapshot* for it.

Metrics. For quantitative comparison, we first align the estimated mesh to ground truth (synthetic data) by Iterative Closest Point (ICP) and then compute the Chamfer Distance (CD)[37] between them, where lower is better.

4.1 Single Garment Recontruction

Here, we aim to compare our approach with SOTA methods on the accuracy and generalization in single garment reconstruction. To avoid the impact of incomplete observations, we only reconstruct a single uncovered garment in each video and obtain the result of the first frame. For image-based methods[15, 24, 39], we input the first frame of the video. Similarly, for video-based methods[18, 45], we also compute the results of the first frame.

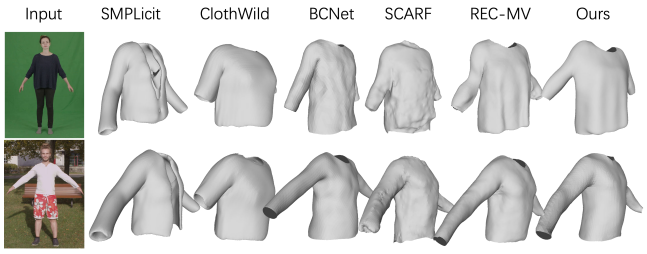
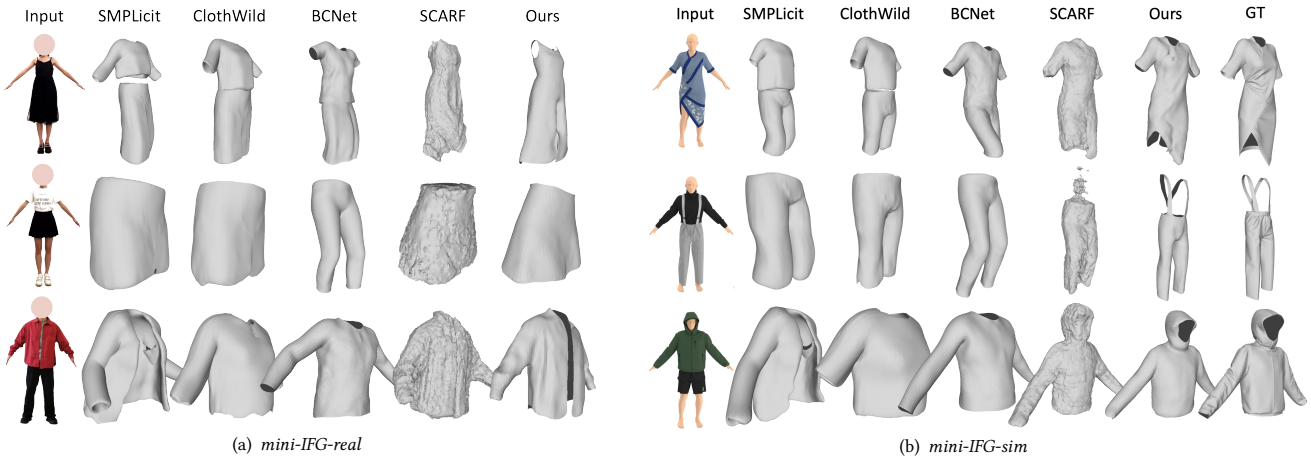


Figure 4: Qualitative comparison on *People Snapshot*. Most methods can model such garments with simple topology.

Table 2: Quantitative results on synthetic sequences (*mini-IFG-sim*). We compare the Chamfer Distance (CD) between the ground-truth and reconstructed surfaces (in *cm*).

Method	Female-a		Female-b			Male-a			Male-b		
	Layer-1	Layer-2	Layer-1	Layer-2	Layer-3	Layer-1	Layer-2	Layer-3	Layer-1	Layer-2	Layer-3
SMPLicit[15]	2.9552	3.1213	1.9412	2.9713	2.5360	3.5042	4.0089	4.7301	1.7433	2.7373	3.4295
ClothWild[39]	2.5154	2.7793	3.1969	1.6661	4.0820	3.6386	4.2489	5.6636	2.4353	3.4968	4.9829
BCNet[24]	3.5378	1.9966	1.0470	1.0520	1.6689	3.6667	2.8670	3.4494	**0.9508**	2.5136	2.8745
SCARF[18]	2.3716	2.7509	4.2129	4.0528	3.5365	3.5762	4.4049	4.5656	2.4269	2.6494	2.9991
Ours	**1.2689**	**1.9152**	**1.0215**	**1.0281**	**1.2331**	**1.5865**	**0.9943**	**1.6717**	1.2363	**1.0536**	**1.1297**



(a) *mini-IFG-real* (b) *mini-IFG-sim*

Figure 5: Qualitative comparison on *mini-IFG*. The goal is to reconstruct the uncovered garment in the image/video. SCARF[18] exhibits good generalization but low quality. Template-based methods[15, 24, 39] lack both detail and generalization. Our template-free approach demonstrates outstanding generalization, confidently handling extremely challenging garments.

4.1.1 Evaluation on Real-world Videos. We show only two typical sequences from *People Snapshot* due to the limited variety of available garments. In Figure 4, most methods can handle such simple clothing. Among them, the template-based methods of REC-MV[45] and BCNet[24] behave close to ours. *mini-IFG-real* includes videos of humans wearing fashionable clothing. As shown in Figure 5(a), our method significantly outperforms others, thanks to the template-free extraction approach. Given the almost infinite variety of clothing styles, template-based methods[15, 24, 39, 45] struggle to be applicable in real life. Although SCARF[18] is also template-free, it suffers from poor geometric quality. In contrast, we achieve excellent generalization and high quality by effectively combining 2D segmentation with 3D implicit neural fields.

4.1.2 Evaluation on Synthesis Videos. Table 2 presents the quantitative results testing on *mini-IFG-sim*. We visually compare the reconstruction quality in Figure 5(b). For simple clothing, BCNet[24] and our method perform comparably. However, for challenging clothing, only our method can obtain correct results. SCARF[18] achieves shape similarity but exhibits significant noise, while others[15, 24, 39] deviate significantly from the ground truth. We demonstrate impressive results by reconstructing challenging clothing such as cheongsams and dungarees, distinctively.

4.1.3 Dynamic Reconstruction. As described in Section 3.2, clothed bodies in both the canonical space and posed space share a consistent geometric topology, which allows us to support dynamic reconstruction as well. We provide examples in Figure 6.

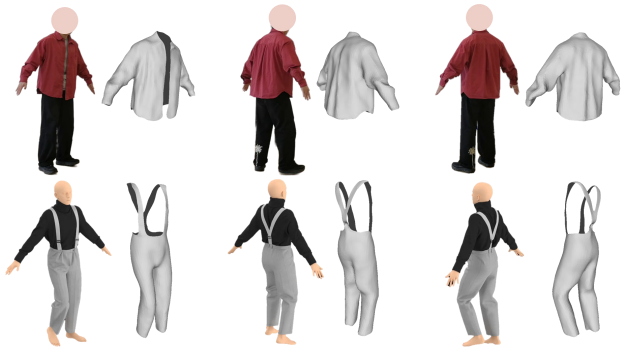


Figure 6: Examples of dynamic reconstruction. For a single garment, since the mesh structure is shared across frames, we can accomplish dynamic capture.

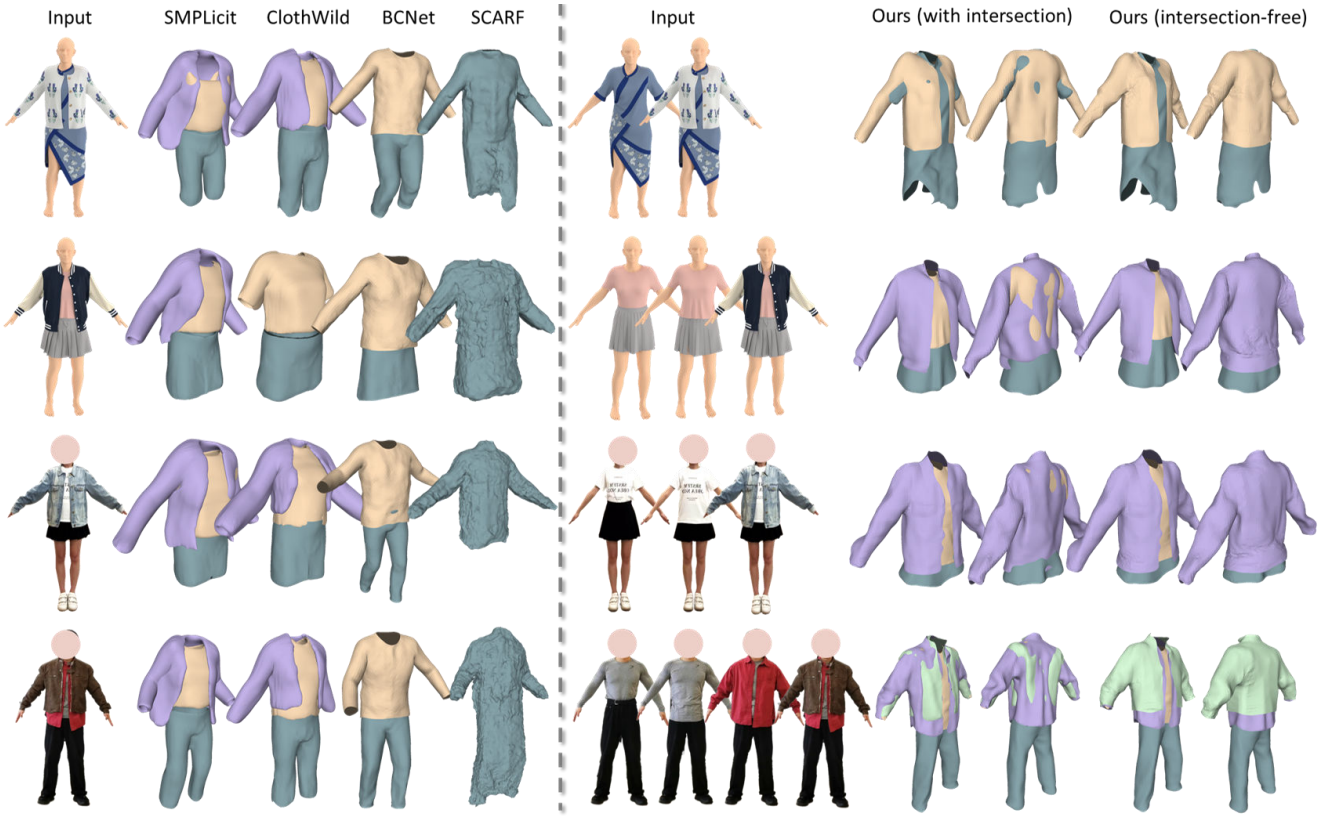


Figure 7: Qualitative comparison of multi-layered clothing reconstruction on *mini-IFG*. SCARF[18] treats clothing as a single entity. The template of BCNet[24] only supports a pair of tops and bottoms. SMPLicit[15] and ClothWild[39] can handle multi-layered clothing but lack detail. In contrast, we advocate for reconstructing multi-layered outfits from several videos. We are the only method that faithfully reconstructs multi-layered clothing while ensuring high quality and no penetration.

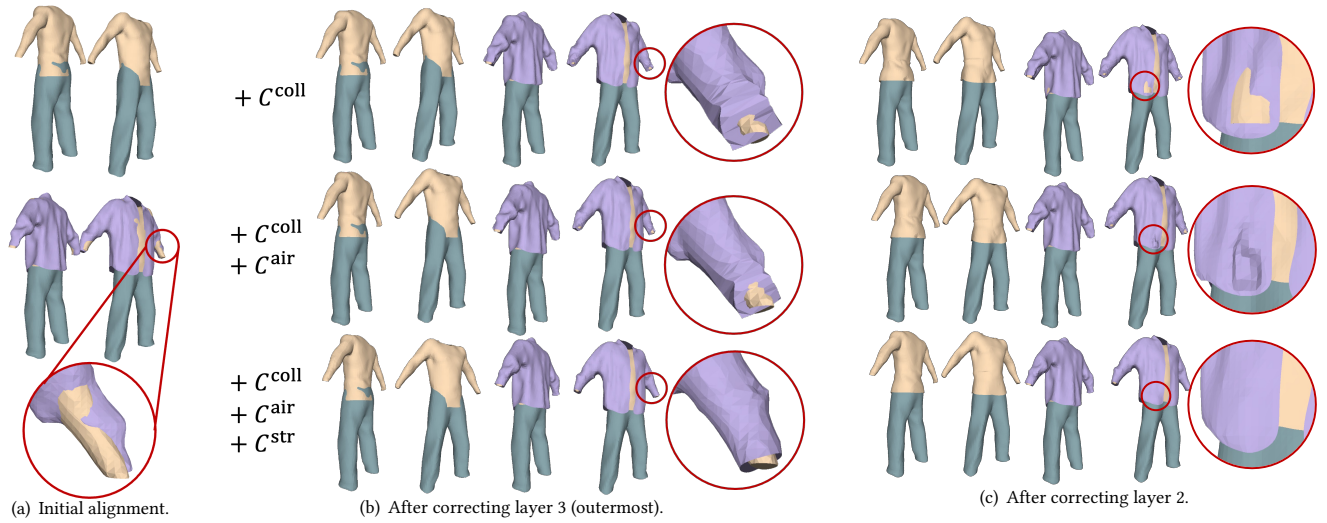
4.2 Multi-Layered Clothing Reconstruction

For each subject, our method inputs multiple videos, while baselines only input one video containing the outermost layer. This is because the baselines focus on the reconstruction from a single video but do not support the combination of clothing from multiple videos. Here, we aim to demonstrate the superiority of reconstructing multi-layered clothing based on multiple videos through comparison with baselines. Moreover, the robustness of penetration handling will also be validated.

Figure 7 provides qualitative comparisons. SCARF[18] exhibits the generality of the template-free method but regards clothing as a whole. The template in [24] is insufficient to meet fashion garments' requirements. Though layered clothing is available in SMPLicit[15] and ClothWild[39], the results lack detail and fail in the case of 4 layers. Fundamentally, since the mutual occlusion of clothing within a single video leads to incomplete observations, they can not obtain multi-layered clothing with great completeness. Therefore, additional videos are necessary, which hardly increases the usage difficulty. However, simply aligning the results from multiple videos inevitably causes penetration. We overcome this issue through a physics-aware module that robustly eliminates intersections, as depicted in the right of Figure 7. For more detailed visualization, please refer to the supplementary material.

4.3 Ablation Study

4.3.1 Constraints in Physics-Aware Module. Here, we adequately illustrate the influence of constraints in the physics-aware module. First, we align the uncovered garments extracted from multiple videos in the canonical space (see Figure 8(a)). There are severe penetrations due to perceptual errors and overlapping deformations. Then, Figure 8(b) and Figure 8(c) depict the penetration correction strategy from the outer to the inner, as discussed in Section 3.3. Obviously, the intersection between layer 3 and layer 2 is alleviated in Figure 8(b) and the same occurs for layer 2 and layer 1 in Figure 8(c). For Figure 8(b) and Figure 8(c), we show constraints activated sequentially from top to bottom. For the top row of Figure 8(b), with only C^{coll}, the penetration between layer 3 and layer 2 is noticeably improved but not completely eliminated (as seen at the cuff). This is because the simulation objects are discrete mesh vertices, and the sign distance values of vertices cannot reliably represent the penetration state of triangle faces. For the top row of Figure 8(c), some vertices of layer 2 are pushed out from the interior of layer 1 and then intersect with layer 3 again (at the hem of the top layer).



(a) Initial alignment. (b) After correcting layer 3 (outermost). (c) After correcting layer 2.

Figure 8: Ablation study on the physics-aware module. In (a), there are penetrations of the initial alignment of three pieces of garments. In (b) and (c), we first correct the intersection of the outermost layer and follow with layer 2. From top to bottom of (b) and (c), different constraints are enabled in order. With C^{coll}, C^{air}, and C^{srt}, our proposed physics-aware module demonstrates outstandingly robust penetration handling capability.

Fortunately, with C^{air} to maintain gaps, the repeated inter-layer intersections are resolved (middle row of Figure 8(c)). However, the corrections brought by C^{air} result in non-smoothness compared to areas where no penetration occurred around. Subsequently, after enabling C^{str}, non-rigid deformation eliminates this distortion (bottom row of Figure 8(c)). It is interesting that the penetration at the cuffs also disappears (bottom row of Figure 8(b)). This is attributed to our original pipeline in Algorithm 1. After solving C^{coll}, some vertices are projected to non-penetrating positions, accompanied by excessive stretching of edges. Then, C^{str} pulls them closer by a certain distance. Since C^{coll} and C^{air} are solved once while C^{str} is iteratively satisfied, they reach a balance as the simulation progresses: no penetration and no excessive deformation.

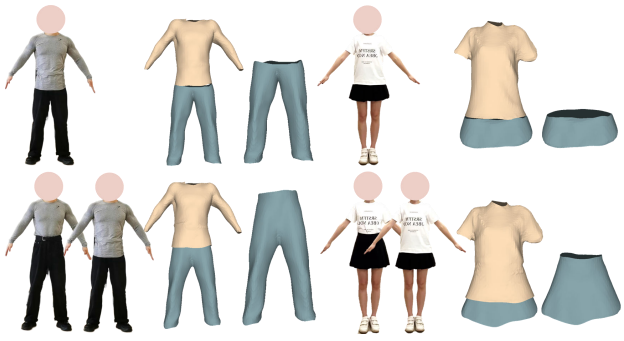


Figure 9: With back-projection described in Section 3.2, we can also extract two garments from a single video (top row), but it is incomplete compared to the multiple videos way that we actually adopted (bottom row).

4.3.2 Reconstruction Two Garments from Single Video. In Figure 9, we claim that IF-Garments can reconstruct two pieces of clothing from one video. However, due to occlusion, the lower lacks completeness. Instead, with multiple videos, all garments are complete.

5 LIMITATIONS

Segmentation. While our proposed template-free clothing extraction method achieves impressive generalization, the segmentation errors negatively impact mesh quality. With the rapid development of research on automated and semi-automated segmentation[25, 28, 32, 44], we believe this issue will be alleviated.

Animation. Currently, we are reconstructing multi-layered clothing in the canonical space. Since the deformation fields corresponding to each layer of clothing are independent, penetration occurs again when garments are mapped to the posed space. So IF-Garments is unsuitable for direct usage in animation. Recently, some research related to clothing simulation has contributed to multi-layered clothing animation[9, 10]. It is possible to achieve penetration-free posed meshes by leveraging their results.

6 CONCLUSION

We have presented IF-Garments, a novel framework for multi-layered intersection-free clothing reconstruction from monocular videos. Our core innovation lies in ingeniously combining neural SDFs with back-projection and physics simulation to accomplish both remarkable generalization of clothing extraction and robustness of handling intersections. Sufficient experiments thoroughly demonstrate the superiority and effectiveness of our method. Due to convenience and high quality, we believe that IF-Garments can benefit downstream multimedia tasks such as human performance capture, personalized avatar modeling, and virtual try-ons.

REFERENCES

[1] 2024. Style3D. https://www.linctex.com/.

[2] Alakh Aggarwal, Jikai Wang, Steven Hogue, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. 2022. Layered-garment net: Generating multiple implicit garment layers from a single image. In *Proceedings of the Asian Conference on Computer Vision*. 3000–3017.

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*. IEEE, 98–109.

[4] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8387–8397.

[5] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*. 408–416.

[6] Kinjal Basu and Art B Owen. 2015. Low discrepancy constructions in the triangle. *SIAM J. Numer. Anal.* 53, 2 (2015), 743–761.

[7] Jan Bender, Matthias Müller, and Miles Macklin. 2015. Position-Based Simulation Methods in Computer Graphics.. In *Eurographics (tutorials)*. 8.

[8] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. 2020. Cloth3d: clothed 3d humans. In *European Conference on Computer Vision*. Springer, 344–359.

[9] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. 2021. PBNS: physically based neural simulation for unsupervised garment pose space deformation. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–14.

[10] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. 2022. Neural cloth simulation. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–14.

[11] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. 2019. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5420–5430.

[12] Andrés Casado-Elvira, Marc Comino Trinidad, and Dan Casas. 2022. PERGAMO: Personalized 3d garments from monocular video. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 293–304.

[13] Julian Chibane, Gerard Pons-Moll, et al. 2020. Neural unsigned distance fields for implicit function learning. *Advances in Neural Information Processing Systems* 33 (2020), 21638–21652.

[14] Paolo Cignoni, Marco Callieri, Massimiliano Corsini, Matteo Dellepiane, Fabio Ganovelli, Guido Ranzuglia, et al. 2008. Meshlab: an open-source mesh processing tool.. In *Eurographics Italian chapter conference*, Vol. 2008. Salerno, Italy, 129–136.

[15] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11875–11885.

[16] Junting Dong, Qi Fang, Yudong Guo, Sida Peng, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2022. Totalselfscan: Learning full-body avatars from self-portrait videos of faces, hands, and bodies. *Advances in Neural Information Processing Systems* 35 (2022), 13654–13667.

[17] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J Black. 2023. Learning disentangled avatars with hybrid 3d representations. *arXiv preprint arXiv:2309.06441* (2023).

[18] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J Black, and Timo Bolkart. 2022. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*. 1–9.

[19] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. 2020. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *Advances in Neural Information Processing Systems* 33 (2020), 9276–9287.

[20] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. 2021. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11046–11056.

[21] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. 2020. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3093–3102.

[22] Inigo Quilez. 2010. Inigo Quilez, distance functions. https://iquilezles.org/,.

[23] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. 2022. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5605–5615.

[24] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. 2020. Bcnet: Learning body and cloth shape from a single image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 18–35.

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

[26] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. 2020. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5253–5263.

[27] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. 2021. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3383–3393.

[28] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. 2020. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 6 (2020), 3260–3271.

[29] Xiongzheng Li, Jinsong Zhang, Yu-Kun Lai, Jingyu Yang, and Kun Li. 2023. High-quality animatable dynamic garment reconstruction from monocular videos. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[30] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. 2022. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*. Springer, 590–606.

[31] Zhe Li, Zerong Zheng, Hongwen Zhang, Chaonan Ji, and Yebin Liu. 2022. Avatarcap: Animatable avatar conditioned monocular human volumetric capture. In *European Conference on Computer Vision*. Springer, 322–341.

[32] Yi Liu, Lutao Chu, Guowei Chen, Zewu Wu, Zeyu Chen, Baohua Lai, and Yuying Hao. 2021. Paddleseg: A high-efficient development toolkit for image segmentation. *arXiv preprint arXiv:2101.06175* (2021).

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–16.

[34] William E Lorensen and Harvey E Cline. 1998. Marching cubes: A high resolution 3D surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*. 347–353.

[35] Miles Macklin, Kenny Erleben, Matthias Müller, Nuttapong Chentanez, Stefan Jeschke, and Zach Corse. 2020. Local optimization for robust signed distance field collision. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 3, 1 (2020), 1–17.

[36] Miles Macklin, Matthias Müller, and Nuttapong Chentanez. 2016. XPBD: position-based simulation of compliant constrained dynamics. In *Proceedings of the 9th International Conference on Motion in Games*. 49–54.

[37] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4460–4470.

[38] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

[39] Gyeongsik Moon, Hyeongjin Nam, Takaaki Shiratori, and Kyoung Mu Lee. 2022. 3D clothed human reconstruction in the wild. In *European conference on computer vision*. Springer, 184–200.

[40] Matthias Müller, Bruno Heidelberger, Marcus Hennix, and John Ratcliff. 2007. Position based dynamics. *Journal of Visual Communication and Image Representation* 18, 2 (2007), 109–118.

[41] Ahmed AA Osman, Timo Bolkart, and Michael J Black. 2020. Star: Sparse trained articulated human body regressor. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 598–613.

[42] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 165–174.

[43] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. 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*. 10975–10985.

[44] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. 2020. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern recognition* 106 (2020), 107404.

[45] Lingteng Qiu, Guanying Chen, Jiapeng Zhou, Mutian Xu, Junle Wang, and Xiaoguang Han. 2023. Rec-mv: Reconstructing 3d dynamic cloth from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4637–4646.

[46] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2304–2314.

[47] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 84–93.

[48] Igor Santesteban, Miguel A Otaduy, and Dan Casas. 2019. Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 355–366.

[49] Mark Segal and Kurt Akeley. 2022. The OpenGL® Graphics System: A Specification (Version 4.6 (Core Profile)-May 5, 2022).

[50] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. 2020. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems* 33 (2020), 7462–7473.
[51] Zhaoqi Su, Weilin Wan, Tao Yu, Lingjie Liu, Lu Fang, Wenping Wang, and Yebin Liu. 2020. Mulaycap: Multi-layer human performance capture using a monocular video camera. *IEEE Transactions on Visualization and Computer Graphics* 28, 4 (2020), 1862–1879.
[52] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. 2023. Recovering 3d human mesh from monocular images: A survey. *IEEE transactions on pattern analysis and machine intelligence* (2023).
[53] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. 2021. Modeling clothing as a separate layer for an animatable human avatar. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–15.
[54] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. 2023. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 512–523.
[55] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. 2022. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 13286–13296.
[56] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. 2018. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (ToG)* 37, 2 (2018), 1–15.
[57] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. 2023. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
[58] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. 2021. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11446–11456.
[59] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. 2021. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence* 44, 6 (2021), 3170–3184.
[60] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. 2020. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 512–530.