Reconstruction of Manipulated Garment with Guided Deformation Prior

Ren Li **Corentin Dumery** **Zhantao Deng** **Pascal Fua**
Computer Vision Lab, EPFL
Lausanne, Switzerland
ren.li@epfl.ch corentin.dumery@epfl.ch zhantao.deng@epfl.ch pascal.fua@epfl.ch

Abstract

Modeling the shape of garments has received much attention, but most existing approaches assume the garments to be worn by someone, which constrains the range of shapes they can assume. In this work, we address shape recovery when garments are being manipulated instead of worn, which gives rise to an even larger range of possible shapes. To this end, we leverage the implicit sewing patterns (ISP) model for garment modeling and extend it by adding a diffusion-based deformation prior to represent these shapes. To recover 3D garment shapes from incomplete 3D point clouds acquired when the garment is folded, we map the points to UV space, in which our priors are learned, to produce partial UV maps, and then fit the priors to recover complete UV maps and 2D to 3D mappings. Experimental results demonstrate the superior reconstruction accuracy of our method compared to previous ones, especially when dealing with large non-rigid deformations arising from the manipulations.

1 Introduction

Garments play an important role in our daily lives, as we interact with them through wearing, folding, and manipulating them. Therefore, the ability to recover their 3D shape is important in many fields, including virtual try-on, VR/AR, and robotic manipulation. However, since garments are non-rigid thin structures with a near-infinite number of degrees of freedom, accurate reconstruction remains a challenge, especially in the presence of massive self-occlusions caused by folding or crumpling.

Most existing techniques focus on reconstructing garments being worn by someone and therefore constrained by the body shape. This limits the amount of crumpling and provides a shape prior that can be exploited. In this paper, we address the even more challenging problem of recovering the shape of garments not being worn and therefore possibly assuming arbitrary shapes, such as those of Fig. 1.

To this end, we start from the Implicit Sewing Patterns (ISP) model [1]. As in models used by clothing designers, each garment consists of individual 2D panels. Their 2D shape is defined by a Signed Distance Function and 3D shape by a 2D to 3D mapping. We chose this formalism because it can handle complex garments with various geometries, while preserving differentiability with respect to observations. However, its 3D parameterization is limited to a single rest state for each garment and is designed to be draped on a human body.

To handle garments unconstrained by the wearer's body, we introduce a prior to represent the many plausible deformations, including folding and crumpling. It is learned using a generative diffusion model that generates 2D positional UV maps that are applicable to many different garments. We use it to recover 3D garment shapes from incomplete 3D point clouds, such as those that are acquired using a laser scanner when the garment is folded.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-1869

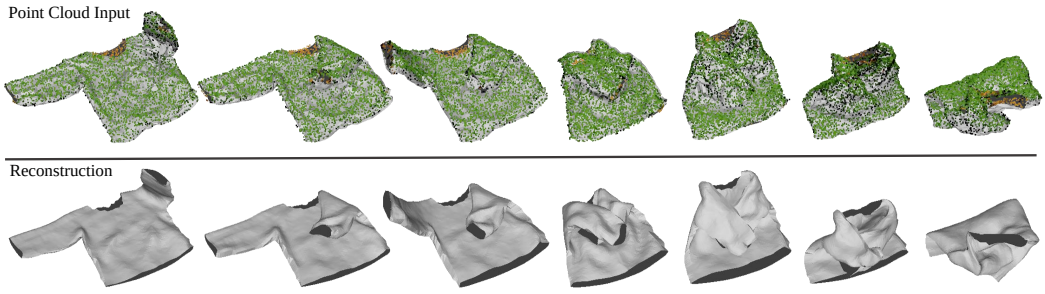


Figure 1: Recovering the 3D shape of folded and crumpled garments from incomplete point clouds. **Top**: The input point clouds (green) overlaid on the ground truth meshes (gray). **Bottom**: Our reconstructions.

In practice, given that every panel in the ISP model is registered to a unified 2D space parameterized by its UV coordinates, we train a *UV mapper* to assign the points to individual panels and to project them into the corresponding UV-space, as shown at the top of Fig. 2. This yields partial UV maps for each panel. We then fit the 2D panels and use a guided reverse diffusion process to generate complete UV maps from the partial ones.

We validate our approach on the data from the VR-Folding dataset [2], where point clouds are generated from multi-view RGBD images. As shown in Fig. 1, our approach accurately reconstructs 3D garment meshes under high levels of deformation and self-occlusion. We also demonstrate that our algorithm can handle real point cloud data. Notably, our method achieves this without requiring explicit prior knowledge of the garment geometry, further demonstrating its practical applicability. This goes well-beyond prior diffusion-based work [3] that can only model the deformations of a single specific garment worn on the human body. Our implementation and model weights are available at `https://github.com/liren2515/GarmentFolding`.

2 Related Work

Non-Rigid 3D Reconstruction. Reconstructing non-rigid deforming objects has been a longstanding research area in computer vision and graphics. One line of work [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16], known as Shape from Template (SfT) and 3D registration, assumes the availability of 3D surface templates. These methods aim to minimize the difference between the observations, such as captured images or point clouds, and the given template surface to recover the deformed state of the objects with geometric constraints or differentiable physics simulators. Unfortunately 3D templates are rarely available. Thus, there are many approaches that rely on free-form techniques without any geometric prior [17, 18, 19, 20], which warp and accumulate different observations across frames into a 3D volume. However, these methods face challenges when reconstructing regions that remain occluded throughout the sequence. While the algorithms of [21, 22, 23, 24, 25, 26] can recover full object geometry from RGB-D videos, the shape representations they use are designed for watertight surfaces, and thus not suitable for garments that are thin open surfaces.

In contrast, GarmentNets [27] recovers the full 3D surfaces of previously unseen garments from point clouds. It leverages the Normalized Object Coordinate Space (NOCS) [28] as a category-specific canonical representation for garments. The garment mesh is reconstructed by mapping the predicted canonical mesh to the deformed one. However, GarmentNets only handles garments being grasped. To cover the much wider range of possible states that garments can be in, a large-scale garment manipulation dataset is introduced in [2]. It relies on a VR system and uses it to learn a tracking model for estimating the complete pose of a given garment. Despite promising results, requiring a canonical geometry for garment and an initial shape estimate in the first frame imposes limitations similar to those of template-based methods. In contrast, our proposed method is not subject to these and can recover 3D meshes of garments with unknown geometry.

On-Body Garment Reconstruction. Clothed human reconstruction has received significant attention in recent years. However, the majority of methods primarily focus on clothing that tightly adheres to the body. In these, garments are represented either explicitly using template meshes

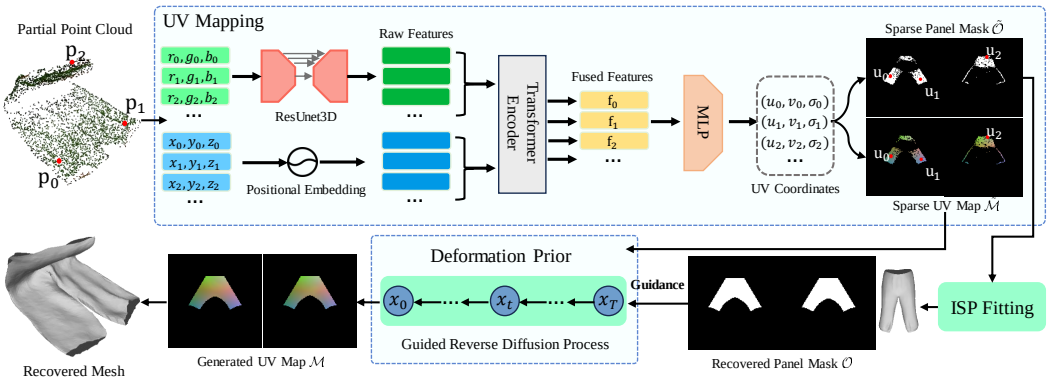


Figure 2: **Our framework**. Given a point cloud, we first map it to UV space to obtain sparse UV maps $\tilde{\mathcal{M}}$ and panel masks $\tilde{\mathcal{O}}$. We recover complete UV maps $\mathcal{M}$ and panel masks $\mathcal{O}$ from them using ISP and a deformation prior, enabling the reconstruction of the deformed garment's 3D mesh.

[29, 30, 31, 32] or implicitly by signed and unsigned distance functions [33, 34, 35]. To handle loose-fitting garments such as skirts and dresses, the methods of [36, 37, 38] leverage complex physics simulation steps or feature line estimation to align the garment surface with the input image. The one of [39] reconstructs garments from point clouds by predicting displacement and principal component analysis (PCA) coefficients for the mesh registered to base templates. While effective, these methods' reliance on garment templates limits their generality. To address this limitation, [40] uses the Implicit Sewing Patterns (ISP) model from [1] as the garment representation and fits an image-conditioned deformation model to the normal estimation of the garments.

However, all these methods rely on the articulated body shape model [41] as a prior. This makes them unsuitable for the perception task in the context of robot manipulations where no body is involved and garments can exhibit higher levels of crumpled deformation and occlusions, which is the focus of our work.

Diffusion Model. Diffusion models [42, 43] are a class of generative models that excel at learning complex data distributions through score matching. By iteratively denoising the data, these models can generate high-quality samples. They have achieved state-of-the-art performance in a wide range of image-based generative tasks [44, 45, 46, 47]. Additionally, diffusion models have found application in various 3D tasks, such as text-to-3D generation [48, 49, 48], image-to-3D generation [50, 48, 51, 52], and point cloud synthesis [53, 54]. Recently, [3] introduced a diffusion-based shape prior for on-body garment registration, employing UV maps to parameterize the garment. However, its prior is specific to a single garment piece and requires coarse registration of the input point clouds. In contrast, our proposed method is capable of handling garments with diverse geometries and does not impose any registration requirements.

3 Method

Given a point cloud $\mathbf{P} \in \mathbb{R}^{N\times 3}$ that partially represents a previously unknown garment, we want to reconstruct an accurate 3D model that captures both its geometry and deformation. Instead of doing this directly in 3D space, we first use a trained UV mapper to map 3D points to the unified 2D UV space in which individual ISP [1] garment panels are represented. In practice, we typically use two such panels, one for the front, and one for the back. In the resulting *UV maps*, each pixel can either be empty or contain the 3D location of a point. We then fit the 2D panels to these and use a reverse diffusion process to fill the potentially large holes in the UV maps, which yields a complete reconstruction. Fig. 2 depicts this process.

In this section, we first briefly describe the Implicit Sewing Pattern (ISP) garment model [1], which represents the garment geometry using one UV map for each 2D panel and which we extend by adding a diffusion-based [42, 43] deformation model. This makes it possible to model plausible and potentially large garment deformations. We then introduce the UV mapper that maps the point

https://doi.org/10.52202/079017-1869

cloud data to UV space of the individual panels and discuss our approach to fitting the augmented ISP model to the resulting UV maps.

3.1 ISP Garment Model

Formalization. ISP [1] is a garment model inspired by the sewing patterns used in the fashion industry to design and manufacture clothes. Such a pattern consists of several 2D panels along with stitch information for assembling them. They are implicitly modeled using a 2D signed distance field (SDF) and a 2D label field, respectively. For a specific garment, its corresponding latent code $\mathbf{z}$, and a point $\mathbf{u}$ in the 2D UV space $\Omega = [-1, 1]^2$, the ISP model outputs the signed distance s to the panel boundary and a label l using a fully connected network $\mathcal{I}_{\mathbf{\Theta}}$ as

$$(s, l) = \mathcal{I}_{\mathbf{\Theta}}(\mathbf{u}, \mathbf{z}) \ . \tag{1}$$

The zero crossing of the SDF defines the shape of the panel, with $s < 0$ indicating that $\mathbf{u}$ is within the panel and $s > 0$ indicating that $\mathbf{u}$ is outside the panel. The label l encodes the stitch information, indicating which panel boundaries should be stitched together. To map the 2D sewing patterns to 3D surfaces, a UV parameterization function $\mathcal{A}_\Phi$ is learned to perform the 2D-to-3D mapping. It is written as

$$\mathbf{X} = \mathcal{A}_\Phi(\mathbf{u}, \mathbf{z}) \ , \tag{2}$$

where $\mathbf{X} \in \mathbb{R}^3$ represents the 3D position of $\mathbf{u}$. ISP effectively registers different garments onto a unified UV space and establishes the mapping functions between points in UV space and the garments' 3D surfaces. Crucially, this is a differentiable representation. Given masks or contours of the panels, we can easily fit the latent code $\mathbf{z}$ to recover the corresponding garment geometry.

Training. Training ISP requires the 2D sewing patterns of 3D garments in a rest state, which are not available in most garment datasets, such as CLOTH3D [55]. Following the garment flattening approach described in [40, 56], we cut the garment mesh into a front and a back piece according to predefined cutting rules and then unfold them into 2D panels by minimizing an as-rigid-as-possible energy [57] to ensure local area preservation. For each garment in the dataset, we generate a front and a back panel as its sewing pattern. By pairing these 2D sewing patterns with their corresponding 3D meshes, we follow the training procedure of [1] to learn the weights of the ISP model $(\mathcal{I}_{\mathbf{\Theta}}, \mathcal{A}_\Phi)$.

3.2 Modeling Deformations

Although the UV parameterization described above is good at representing garments in their rest state, it does not capture the various deformations that can occur when the garment is subjected to external forces, such as folding or creasing. To address this and model the possibly large deformations of garments, we incorporate a deformation prior into ISP.

Given a set of deformed garments whose rest states are modeled by ISP, we write the corresponding UV maps as

$$\mathcal{M}[u, v] = \begin{cases} \mathbf{V}, & \text{if } s_{\mathbf{u}} \leq 0 \\ \varnothing, & \text{if } s_{\mathbf{u}} > 0 \end{cases} \ , \tag{3}$$

where $\mathbf{V} \in \mathbb{R}^3$ is the corresponding position on the deformed mesh surface for the UV point $\mathbf{u} = (u, v)$, $s_{\mathbf{u}}$ is the SDF value of $\mathbf{u}$, $[\cdot, \cdot]$ denotes standard array addressing and $\varnothing = (-1, -1, -1)$. Note that $\varnothing$ indicates that $\mathbf{u}$ is out of the panel and has no corresponding 3D point. Each $\mathcal{M}$ represents a specific deformed state for a particular garment. To capture the distribution of plausible deformations represented in this way, we learn a deformation prior using a standard diffusion model [42, 43].

Diffusion. The popular Denoising Diffusion Probabilistic Model (DDPM) framework [42] comprises a forward and a reverse process. The forward process perturbs the clean data $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ by iteratively adding Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ to it. This is written as

$$\mathbf{x}_t = \sqrt{1 - \beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon} \ , \tag{4}$$

where $\mathbf{x}_t$ is the noised intermediate state at step $t = 1, 2, .., T$, and $\beta_t \in (0, 1)$ denotes the variance schedule. The reverse process recovers the clean data from random noise with a trained neural network $\boldsymbol{\epsilon}_\theta$

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z} \ , \tag{5}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ and $\sigma_t = \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}$.

Training. In our context, we concatenate the UV map $\mathcal{M}$ generated by Eq. 3 with the panel mask $\mathcal{O}$ along the channel dimension to form the training samples $\mathbf{x}_0 = [\mathcal{M}, \mathcal{O}]$ where

$$\mathcal{O}[u, v] = \begin{cases} 1, & \text{if } s_{\mathbf{u}} \leq 0 \\ 0, & \text{if } s_{\mathbf{u}} > 0 \end{cases} . \tag{6}$$

The panel mask $\mathcal{O}$ encodes the shape of the panels as well as the 3D geometry of the canonical garment. The network $\boldsymbol{\epsilon}_\theta$ is trained on corrupted $\mathbf{x}_0$ to predict the noise by minimizing the loss

$$L = \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta \left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\right)\|_2 . \tag{7}$$

Once the diffusion model is trained, it learns the deformation prior, enabling it to generate or recover realistic and diverse deformations for different garments.

3.3 Mapping Point Cloud to UV Space

To relate an input 3D point cloud $\mathbf{P}$ of the garment to the UV space in which the deformation prior is learned, we rely on the UV mapper $\mathcal{G}$ shown at the top of Fig. 2. For each 3D point, it predicts σ, the probability of belonging to either the front or back panel, along with the u, v coordinates of the pixel where the 3D point should be stored in the UV map. As in [2], we use a sparse 3D convolution network [58] to extract raw features for each point $\mathbf{p}_i$ in $\mathbf{P}$. These raw features are then passed through a transformer encoder with self-attention, producing fused per-point features $\mathbf{f}_i$ that capture relationships across points. These are fed to an MLP that predicts the per-point UV coordinates. It outputs probability distributions $\phi_u \in \mathbb{R}^K$ and $\phi_v \in \mathbb{R}^K$ over K discrete values for the u- and v-axes, along with σ. $\mathcal{G}$ is trained by minimizing

$$L_{\mathcal{G}} = CE(\phi_u, \hat{\phi}_u) + CE(\phi_v, \hat{\phi}_v) + BE(\sigma, \hat{\sigma}) , \tag{8}$$

where $\hat{\cdot}$ denotes the ground-truth values, and CE and BE are the cross entropy and the binary cross entropy, respectively.

Once trained, $\mathcal{G}$ assigns to each point $\mathbf{p}_i$ a UV coordinate $\mathbf{u}_i = (u_i, v_i)$ in the front ($\sigma_i \geq 0.5$) or the back ($\sigma_i < 0.5$) panel with

$$u_i = -1 + \frac{2k_u}{K-1},\ v_i = -1 + \frac{2k_v}{K-1} , \tag{9}$$

where $k_u = \underset{k \in \{0,\ldots,K-1\}}{\operatorname{argmax}} \phi_u^k$ and $k_v = \underset{k \in \{0,\ldots,K-1\}}{\operatorname{argmax}} \phi_v^k$.

We then combine these predictions with $\tilde{\mathcal{M}}[u_i, v_i] = \mathbf{p}_i$ and $\tilde{\mathcal{O}}[u_i, v_i] = 1$ at pixels where a 3D point is projected, and $\tilde{\mathcal{M}}[u_i, v_i] = \varnothing$ and $\tilde{\mathcal{O}}[u_i, v_i] = 0$ elsewhere, producing the assembled UV map $\tilde{\mathcal{M}}$ and the panel mask $\tilde{\mathcal{O}}$.

3.4 Fitting the Model

When the garment deformations are severe, there are many occlusions, and $\tilde{\mathcal{M}}$ and $\tilde{\mathcal{O}}$ are typically sparse. Nevertheless, we can use the deformation model of Section 3.2 to fill-in the holes and recover complete UV maps. To this end, we first recover the 2D shape of the 2D panels and then their individual 3D surfaces, as shown at the bottom of Fig. 2.

Panel Recovery. To recover the 2D shape of the panels, we find the latent code $\mathbf{z}$ of Eq. 1 that yields patterns matching $\tilde{\mathcal{O}}$ as well as possible. We take it to be

$$\mathbf{z}^* = \underset{\mathbf{z}}{\operatorname{argmin}} \sum_{\mathbf{u} \in \mathcal{O}_+} R(-s_{\mathbf{u}}(\mathbf{z})) - \lambda_{area} \sum_{\mathbf{u} \in \Omega} s_{\mathbf{u}}(\mathbf{z}) + \lambda_{\mathbf{z}}||\mathbf{z}||_2 , \tag{10}$$

where $\mathcal{O}_+ = \{\mathbf{u} | \tilde{\mathcal{O}}_{\mathbf{u}} = 1, \mathbf{u} \in \Omega\}$, $R(\cdot)$ is the ReLU function, $s_{\mathbf{u}}(\mathbf{z})$ is the SDF value of $\mathbf{u}$ computed by ISP, and λ_{area} and $\lambda_{\mathbf{z}}$ are the weighting constants. Since the second item of the objective function

https://doi.org/10.52202/079017-1869

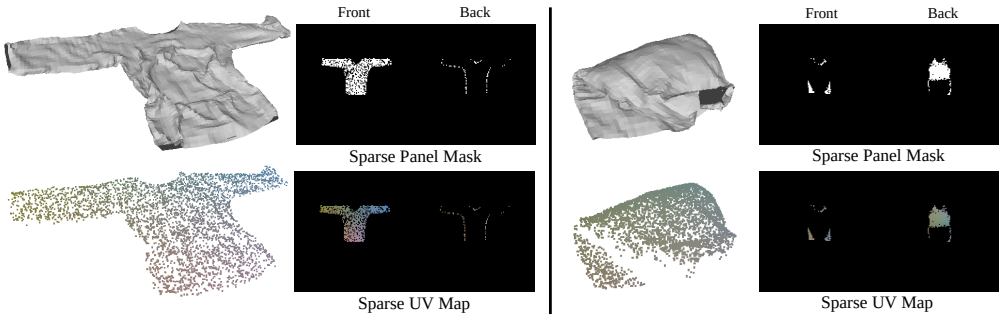


Figure 3: **The projected sparse masks $\tilde{\mathcal{O}}$ and UV maps $\tilde{\mathcal{M}}$** of the point clouds with (a) the maximum volume and (b) the minimum volume. The point clouds are color coded by their 3D positions.

in Eq. 10 penalizes large panel area, this yields panels $\mathcal{O}$ whose contours—the zero crossings of the SDF—surround the non-zero point of $\tilde{\mathcal{O}}$ as closely as possible, as shown in the bottom right of Fig. 2.

Given a point-cloud sequence, we solve Eq. 10 only once, specifically at the frame where the point cloud occupies the maximum volume in 3D space. This choice is motivated by the fact that a large volume of the point cloud indicates less deformation and occlusion of the garment, resulting in a more visible and informative mask $\tilde{\mathcal{O}}$, as shown in Fig. 3. Using this frame for the optimization of $\mathbf{z}$ leads to more accurate fitting results as demonstrated in the Appendix.

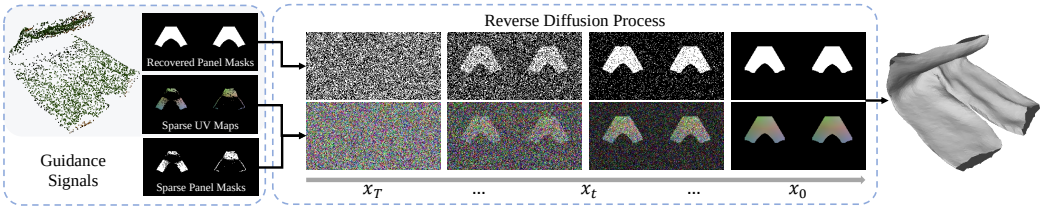


Figure 4: **The guided reverse diffusion process**. The UV maps of the deformed garment are generated by using the observations from the input point cloud as guidance to direct the reverse diffusion process.

Deformation Recovery. The deformation model of Section 3.2 has been trained to generate UV maps representing plausible deformations of garments of many different geometries. To align the generation process with the observation of the sparse UV map $\tilde{\mathcal{M}}$ and the ISP-recovered mask $\mathcal{O}$ introduced in the previous paragraph, we use them as manifold guidance [45, 46] in the reverse diffusion process. We write

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\tilde{\mathcal{M}}, \tilde{\mathcal{O}}, \mathcal{O}) \simeq -\frac{\epsilon_\theta(\mathbf{x}_t, t)}{\sigma_t} - \rho \nabla_{\mathbf{x}_t} d(\hat{\mathbf{x}}_0, \tilde{\mathcal{M}}, \tilde{\mathcal{O}}, \mathcal{O}) \ , \tag{11}$$

$$\hat{\mathbf{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\mathbf{x}_t - \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t) \ , \tag{12}$$

where ρ is the guidance step size. The function d measures the difference between the generated result and the observations

$$d(\hat{\mathbf{x}}_0, \tilde{\mathcal{M}}, \tilde{\mathcal{O}}, \mathcal{O}) = \|\tilde{\mathcal{O}} * (\hat{\mathbf{x}}_{0,\mathcal{M}} - \tilde{\mathcal{M}})\|_2 + \|\hat{\mathbf{x}}_{0,\mathcal{O}} - \mathcal{O}\|_1 \ , \tag{13}$$

where $\hat{\mathbf{x}}_{0,\mathcal{M}}$ and $\hat{\mathbf{x}}_{0,\mathcal{O}}$ refer to the generated UV map and panel mask respectively, and $*$ denotes the elementwise multiplication. When sequential information is available, we additionally refine our method by using the previous prediction $\mathcal{M}_{prev}$ as the regularization for the unobserved part

$$d(\hat{\mathbf{x}}_0, \tilde{\mathcal{M}}, \tilde{\mathcal{O}}, \mathcal{O}) = \|\tilde{\mathcal{O}} * (\hat{\mathbf{x}}_{0,\mathcal{M}} - \tilde{\mathcal{M}})\|_2 + \|\hat{\mathbf{x}}_{0,\mathcal{O}} - \mathcal{O}\|_1 + \lambda\|(1-\tilde{\mathcal{O}}) * (\hat{\mathbf{x}}_{0,\mathcal{M}} - \mathcal{M}_{prev})\|_2. \tag{14}$$

where λ is a weighting constant. Finally, the garment mesh is inferred from the generated UV map using the mapping function of ISP. As illustrated in Fig. 4, this process finally produces a garment mesh that aligns with the point cloud observation.

4 Experiments

4.1 Dataset, Evaluation Metrics, and Baseline

We train our models using data from the VR-Folding [2] and CLOTH3D [55] datasets. The VR-Folding dataset is collected using a VR-based recording system, where participants manipulate garments (i.e., folding and flattening) in a simulator through a VR interface. The dataset features pants, shirts, tops, and skirts selected from the CLOTH3D, and each category covers a wide shape range. VR-Folding comprises 9767 manipulation videos and 790K multi-view RGB-D frames, which are used to generate point clouds. For the training of ISP, we use the same garments from CLOTH3D as those selected in VR-Folding and generate their sewing patterns as described in Sec. 3.1. We generate UV maps of deformed garments and the corresponding UV coordinates of point clouds using the mapping function of ISP. These UV maps and coordinates serve to train the diffusion model and the UV mapper, respectively. For each garment category, we train a separate set of models, using the same training and test splits as [2].

As in [2, 27], we employ the Chamfer Distance D_{cf} and the Correspondence Distance (D_{cr}, A_d) as evaluation metrics. D_{cf} measures the surface reconstruction quality by calculating the Chamfer distance in centimeters from the reconstructed mesh to the ground truth. D_{cr} represents the point-wise L2 distance in centimeters between the reconstruction and the ground truth, which evaluates the accuracy of garment pose estimation. Note that the correspondences of D_{cr} are established by finding the closest point of the ground truth in canonical space instead of the deformed one as D_{cf}. Finally, we take A_d to be the ratio of frames with $D_{cr} < d$ cm.

We compare our method against state-of-the-art approaches: GarmentNets [27] and GarmentTracking [2]. GarmentTracking estimates the per-vertex garment pose based on the given canonical garment mesh and the initialization of the first frame. GarmentNets is a single-frame garment shape estimation method that utilizes the winding number field for garment meshing. Like our method, GarmentNets does not require ground truth geometry.

4.2 Quantitative Results

Table 1: **Quantitative comparisons** of our method to GarmentNets and GarmentTracking on VR-Folding dataset.

Type	Method	Init.	Folding				Flattening			
			$A_3 \uparrow$	$A_5 \uparrow$	$D_{cr} \downarrow$	$D_{cf} \downarrow$	$A_5 \uparrow$	$A_{10} \uparrow$	$D_{cr} \downarrow$	$D_{cf} \downarrow$
Shirt	GarmentNets [27]	N/A	0.8%	21.5%	6.40	1.58	13.2%	59.4%	10.54	3.54
	GarmentTracking [2]	GT	29.8%	85.8%	3.88	1.16	30.7%	83.4%	8.63	1.78
	GarmentTracking [2]	Pert.	29.0%	85.9%	3.88	1.18	25.4%	81.6%	8.94	1.85
	GarmentTracking [2]	GN.	25.4%	78.9%	4.04	1.18	-	-	-	-
	Ours	N/A	**84.7%**	**97.9%**	**2.36**	**0.77**	**78.8%**	**95.2%**	**4.19**	**1.08**
Pants	GarmentNets [27]	N/A	16.2%	69.5%	4.43	1.30	1.5%	42.4%	12.54	4.19
	GarmentTracking [2]	GT	47.3%	94.0%	3.26	1.07	31.3%	78.2%	8.97	1.64
	GarmentTracking [2]	Pert.	42.8%	93.6%	3.35	1.10	30.7%	76.9%	9.55	2.71
	GarmentTracking [2]	GN	45.1%	92.2%	3.33	1.16	-	-	-	-
	Ours	N/A	**75.9%**	**97.9%**	**2.69**	**0.70**	**60.8%**	**91.4%**	**5.32**	**1.16**
Top	GarmentNets [27]	N/A	10.3%	53.8%	5.19	1.51	13.1%	42.5%	12.11	2.85
	GarmentTracking [2]	GT	37.9%	85.9%	3.75	0.99	54.6%	82.8%	6.59	1.15
	GarmentTracking [2]	Pert.	36.6%	86.1%	3.76	1.00	54.2%	82.6%	7.80	2.59
	GarmentTracking [2]	GN	21.1%	61.9%	4.82	1.11	-	-	-	-
	Ours	N/A	**71.2%**	**93.5%**	**2.65**	**0.74**	**70.2%**	**86.2%**	**5.24**	**1.08**
Skirt	GarmentNets [27]	N/A	1.1%	30.3%	6.95	1.89	0.1%	7.9%	18.48	5.99
	GarmentTracking [2]	GT	23.5%	71.3%	**4.61**	1.33	**5.4%**	**39.4%**	16.09	2.02
	GarmentTracking [2]	Pert.	22.8%	70.6%	4.72	1.36	2.3%	35.5%	16.55	2.15
	GarmentTracking [2]	GN	14.7%	65.9%	5.36	1.46	-	-	-	-
	Ours	N/A	**32.5%**	**76.5%**	4.70	**1.04**	5.1%	33.1%	**14.26**	**1.75**

Table 1 shows the quantitative results obtained on the VR-Folding dataset. In the third column, the abbreviation "N/A" indicates the absence of any initialization, while "GT", "Pert." and "GN" represent

https://doi.org/10.52202/079017-1869

the results of GarmentTracking using the ground-truth mesh, the noise-perturbed ground-truth mesh, and the estimation of GarmentNets as the initialization, respectively. Our method outperforms the baselines by a large margin for both the Folding and Flattening sets. GarmentNets can handle garments without prior knowledge of their geometry but has the lowest reconstruction accuracy. GarmentTracking is more accurate and benefits from using the given garment mesh as a prior, but its performance is highly influenced by the choice of initialization. When using the network-predicted result (GN) instead of the ground truth, its performance drops substantially, particularly for Shirt, Skirt and Top. This greatly restricts its applicability in real-world scenarios where obtaining the ground truth garment mesh in advance is rarely possible. In contrast, our method has no such limitation while still achieving the highest reconstruction accuracy. The performance disparity between our method and the baselines is particularly significant for challenging metrics such as A_3 (on Folding) and A_5 (on Flattening), for instance, 84.7% vs 29.8% and 78.0% vs 24.5% on Shirt.

We also notice that both our method and the baseline models achieve relatively higher D_{cr} and lower A_d values for Skirt compared to other categories. This discrepancy arises from the ambiguous definition of skirt sides due to its rotational symmetry. When the skirt is rotated by a specific amount around the medial axis, the resulting shape is nearly identical to the original one, which can yield high D_{cr} and low A_d values because they are computed using correspondence between the estimated canonical mesh and the ground truth. Consequently, this symmetrical ambiguity makes these metrics unsuitable for assessing the reconstruction quality of skirts, whereas the Chamfer distance D_{cf} does not have this issue, on which we obtain the lowest values. An illustrative example is provided in the Appendix.

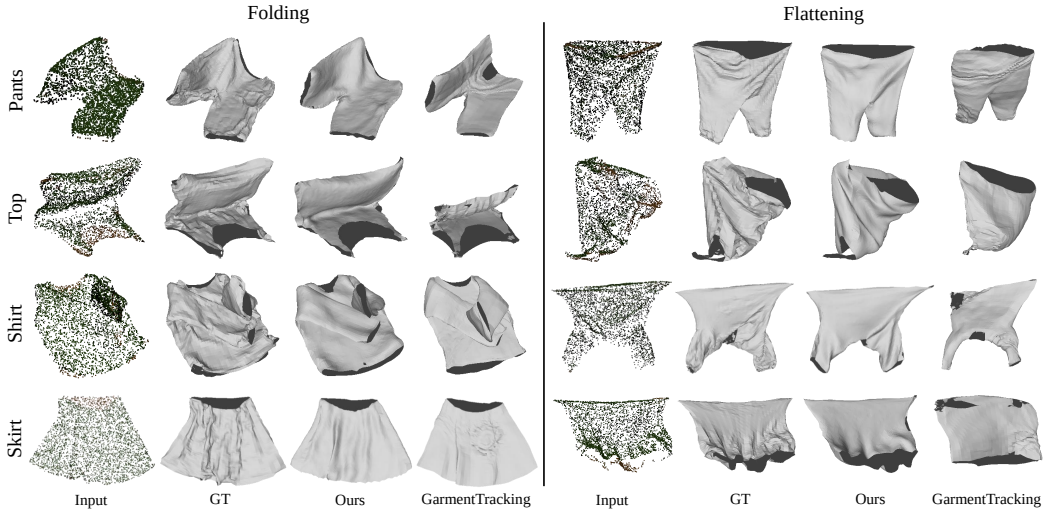


Figure 5: **Qualitative comparisons** of our method to GarmentTracking (initialized with ground truth meshes) on VR-Folding dataset for the categories of Pants, Top, Shirt and Skirt.

4.3 Qualitative Results

In Fig. 5, we show the qualitative comparison with GarmentTracking which uses the ground truth garment mesh as the initialization. GarmentTracking produces results with inaccurate size and deformation, and unrealistic artifact can show up on the reconstructed surfaces. In contrast, our method can recover garment meshes from input point clouds faithfully with correct shape and deformations. In Fig. 6, we further show the reconstructed results for a folding and a flattening sequences, which demonstrates our method can consistently produce accurate results compared with GarmentTracking. More qualitative comparisons can be found in the Appendix.

4.4 Evaluation on Real-World Data

To evaluate our method on real-world data, we capture RGB images of a pair of pants and a sweater, and compute dense point clouds using the *nerfstudio* library [59] for them, as illustrated in Fig. 7 (a). We remove background points and use the resulting point cloud downsampled as input to our method.

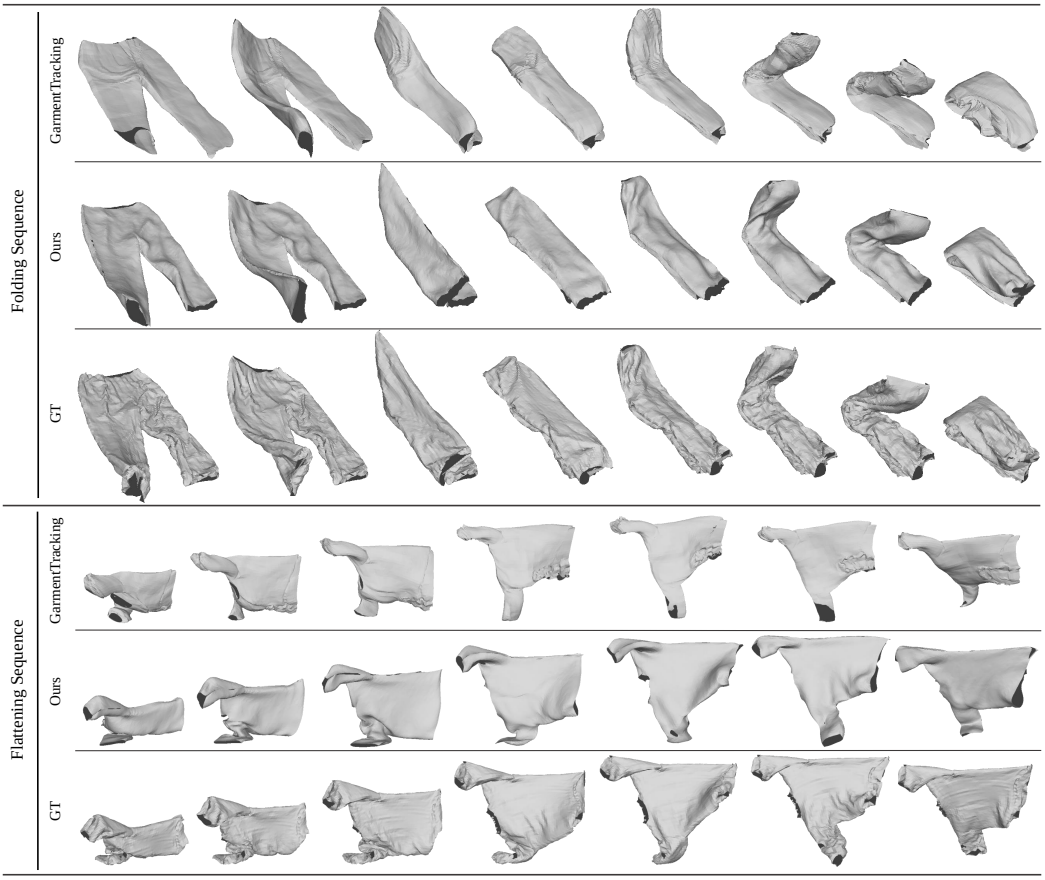


Figure 6: **Qualitative comparisons** of our method to GarmentTracking (initialized with ground truth meshes) on VR-Folding dataset for the sequences of Folding and Flattening.

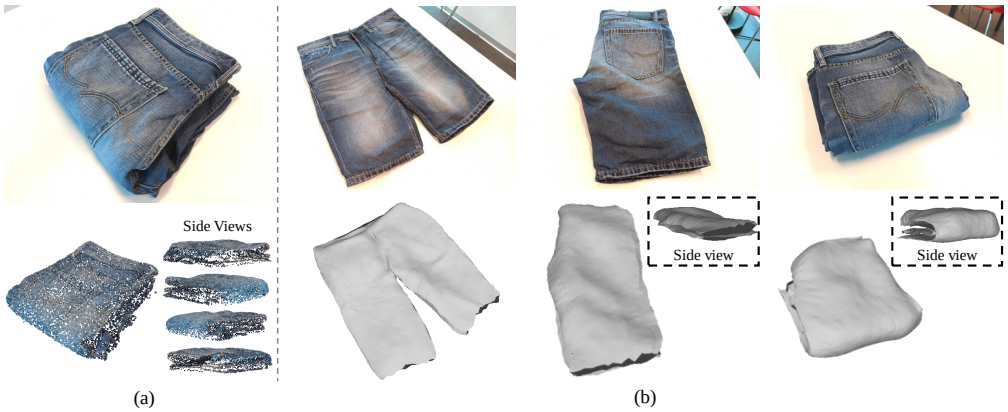


Figure 7: **Real-world evaluation**. (a) The captured image and point cloud of the pants. (b) Our reconstructed results.

Fig. 7 (b) shows the qualitative results for the pants (the results of the sweater are included in the Appendix). Despite being trained on simulated data, our method is able to reconstruct 3D meshes for both flat and folded garments in real-world scenarios.

https://doi.org/10.52202/079017-1869

5 Conclusion

We have proposed a method that addresses the challenges of reconstructing garment that is not being worn and can be manipulated in complex ways. It leverages the Implicit Sewing Patterns (ISP) model for geometry modeling, a generative diffusion model for learning deformation prior, and a UV mapping network to relate the 3D point cloud observations to the UV space where the priors are learned. We have demonstrated the effectiveness of our fitting approach in accurately reconstructing garment meshes in the presence of severe self-occlusion and unknown garment geometries. In future work, we will incorporate accumulated point-cloud information across time to improve the accuracy of UV mapping and mesh reconstruction.

Acknowledgement. This project was supported in part by the Swiss National Science Foundation.

References

[1] R. Li, B. Guillard, and P. Fua. ISP: Multi-Layered Garment Draping with Implicit Sewing Patterns. In *Advances in Neural Information Processing Systems*, 2023.

[2] H. Xue, W. Xu, J. Zhang, T. Tang, Y. Li, W. Du, R. Ye, and C. Lu. Garmenttracking: Category-Level Garment Pose Tracking. In *Conference on Computer Vision and Pattern Recognition*, pages 21233–21242, 2023.

[3] J. Guo, F. Prada, D. Xiang, J. Romero, C. Wu, H. S. Park, T. Shiratori, and S. Saito. Diffusion Shape Prior for Wrinkle-Accurate Cloth Registration. In *International Conference on 3D Vision*, 2024.

[4] D. Ngo, S. Park, A. Jorstad, A. Crivellaro, C. Yoo, and P. Fua. Dense Image Registration and Deformable Surface Reconstruction in Presence of Occlusions and Minimal Texture. In *International Conference on Computer Vision*, 2015.

[5] M. Perriollat, R. Hartley, and A. Bartoli. Monocular Template-Based Reconstruction of Inextensible Surfaces. *International Journal of Computer Vision*, 95(2):124–137, 2011.

[6] M. Salzmann, J. Pilet, S. Ilić, and P. Fua. Surface Deformation Models for Non-Rigid 3D Shape Recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1481–1487, February 2007.

[7] R. Yu, C. Russell, N. Campbell, and L. Agapito. Direct, Dense, and Deformable: Template-Based Non-Rigid 3D Reconstruction from RGB Video. In *International Conference on Computer Vision*, 2015.

[8] A. Pumarola, A. Agudo, L. Porzi, A. Sanfeliu, V. Lepetit, and F. Moreno-Noguer. Geometry-Aware Network for Non-Rigid Shape Prediction from a Single View. In *Conference on Computer Vision and Pattern Recognition*, June 2018.

[9] N. Kairanda, E. Tretschk, M. Elgharib, C. Theobalt, and V. Golyanik. F-Sft: Shape-From-Template with a Physics-Based Deformation Model. In *Conference on Computer Vision and Pattern Recognition*, pages 3948–3958, 2022.

[10] B. Amberg, S. Romdhani, and T. Vetter. Optimal Step Nonrigid ICP Algorithms for Surface Registration. In *Conference on Computer Vision and Pattern Recognition*, 2007.

[11] Y. Furukawa and J. Ponce. Dense 3D Motion Capture from Synchronized Video Streams. In *Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[12] M. Salzmann and P. Fua. Linear Local Models for Monocular Reconstruction of Deformable Surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):931–944, 2011.

[13] G. Pons-Moll, S. Pujades, S. Hu, and M.J. Black. Clothcap: Seamless 4D Clothing Capture and Retargeting. *ACM SIGGRAPH*, 36(4):731–7315, July 2017.

[14] D. Xiang, F. Prada, C. Wu, and J. Hodgins. Monoclothcap: Towards Temporally Coherent Clothing Capture from Monocular RGB Video. In *arXiv Preprint*, 2020.

[15] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon. The Vitruvian Manifold: Inferring Dense Correspondences for One-Shot Human Pose Estimation. In *Conference on Computer Vision and Pattern Recognition*, pages 103–110, 2012.

[16] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry. 3D-CODED : 3D Correspondences by Deep Deformation. In *European Conference on Computer Vision*, 2018.

[17] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and Tracking of Non-Rigid Scenes in Real-Time. In *Conference on Computer Vision and Pattern Recognition*, 2015.

[18] A. Bozic, M. Zollhofer, C. Theobalt, and M. Nießner. Deepdeform: Learning Non-Rigid Rgb-D Reconstruction with Semi-Supervised Data. In *Conference on Computer Vision and Pattern Recognition*, pages 7002–7012, 2020.

[19] A. Bozic, P. Palafox, M. Zollhöfer, A. Dai, J. Thies, and M. Nießner. Neural Non-Rigid Tracking. In *Advances in Neural Information Processing Systems*, pages 18727–18737, 2020.

[20] S. Parashar, Y. Long, M. Salzmann, and P. Fua. A Closed-Form, Pairwise Solution to Local Non-Rigid Structure-From-Motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[21] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4D: Real-Time Performance Capture of Challenging Scenes. *ACM Transactions on Graphics*, 35(4):1–13, 2016.

[22] M. Dou, P. Davidson, S. R. Fanello, S. Khamis, A. Kowdle, C. Rhemann, V. Tankovich, and S. Izadi. Motion2fusion: Real-Time Volumetric Performance Capture. *ACM Transactions on Graphics*, 36(6):1–16, 2017.

[23] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. Occupancy Flow: 4D Reconstruction by Learning Particle Dynamics. In *International Conference on Computer Vision*, 2019.

[24] P. Palafox, A. Božič, J. Thies, M. Nießner, and A. Dai. NPMs: Neural Parametric Models for 3D Deformable Shapes. In *International Conference on Computer Vision*, 2021.

[25] Y. Li, H. Takehara, T. Taketomi, B. Zheng, and M. Nießner. 4Dcomplete: Non-Rigid Motion Estimation Beyond the Observable Surface. In *International Conference on Computer Vision*, pages 12706–12716, 2021.

[26] W. Lin, C. Zheng, J.-H. Yong, and F. Xu. Occlusionfusion: Occlusion-Aware Motion Estimation for Real-Time Dynamic 3D Reconstruction. In *Conference on Computer Vision and Pattern Recognition*, pages 1736–1745, 2022.

[27] C. Chi and S. Song. Garmentnets: Category-Level Pose Estimation for Garments via Canonical Space Shape Completion. In *International Conference on Computer Vision*, pages 3324–3333, 2021.

[28] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas. Normalized Object Coordinate Space for Category-Level 6d Object Pose and Size Estimation. In *Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019.

[29] R. Danerek, E. Dibra, C. öztireli, R. Ziegler, and M. Gross. Deepgarment : 3D Garment Shape Estimation from a Single Image. *Eurographics*, 2017.

[30] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multi-Garment Net: Learning to Dress 3D People from Images. In *International Conference on Computer Vision*, 2019.

[31] B. Jiang, J. Zhang, Y. Hong, J. Luo, L. Liu, and H. Bao. Bcnet: Learning Body and Cloth Shape from a Single Image. In *European Conference on Computer Vision*, 2020.

 https://doi.org/10.52202/079017-1869

[32] X. Liu, J. Li, and G. Lu. Modeling Realistic Clothing from a Single Image Under Normal Guide. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[33] E. Corona, A. Pumarola, G. Alenyà, and F. Moreno-Noguer. Context-Aware Human Motion Prediction. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[34] R. Li, B. Guillard, E. Remelli, and P. Fua. DIG: Draping Implicit Garment over the Human Body. In *Asian Conference on Computer Vision*, 2022.

[35] L. DeLuigi, R. Li, B. Guillard, M. Salzmann, and P. Fua. Drapenet: Generating Garments and Draping Them with Self-Supervision. In *Conference on Computer Vision and Pattern Recognition*, 2023.

[36] S. Yang, Z. Pan, T. Amert, K. Wang, L. Yu, T. Berg, and M. Lin. Physics-Inspired Garment Recovery from a Single-View Image. *ACM Transactions on Graphics*, 37(5):1–14, 2018.

[37] H. Zhu, Y. Cao, H. Jin, W. Chen, D. Du, Z. Wang, S. Cui, and X. Han. Deep Fashion3D: A Dataset and Benchmark for 3d Garment Reconstruction from Single Images. In *European Conference on Computer Vision*, pages 512–530, 2020.

[38] H. Zhu, L. Qiu, Y. Qiu, and X. Han. Registering Explicit to Implicit: Towards High-Fidelity Garment Mesh Reconstruction from Single Images. In *Conference on Computer Vision and Pattern Recognition*, pages 3845–3854, 2022.

[39] F. Hong, L. Pan, Z. Cai, and Z. Liu. Garment4D: Garment Reconstruction from Point Cloud Sequences. In *Advances in Neural Information Processing Systems*, pages 27940–27951, 2021.

[40] R. Li, C. Dumery, B. Guillard, and P. Fua. Garment Recovery with Shape and Deformation Priors. In *Conference on Computer Vision and Pattern Recognition*, 2024.

[41] M. Loper and M.J. Black. Opendr: An Approximate Differentiable Renderer. In *European Conference on Computer Vision*, pages 154–169, 2014.

[42] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[43] J. Song, C. Meng, and S. Ermon. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*, 2021.

[44] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021.

[45] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *International Conference on Learning Representations*, 2022.

[46] H. Chung, B. Sim, D. Ryu, and J. C. Ye. Improving Diffusion Models for Inverse Problems Using Manifold Constraints. In *Advances in Neural Information Processing Systems*, pages 25683–25696, 2022.

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[48] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. Dmv3D: Denoising Multi-View Diffusion Using 3d Large Reconstruction Model. In *International Conference on Learning Representations*, 2024.

[49] Ben Poole, A. Jain, J. T. Barron, and Ben Mildenhall. Dreamfusion: Text-To-3D Using 2D Diffusion. In *International Conference on Learning Representations*, 2022.

[50] N. Müller, Y. Siddiqui, L. Porzi, S. R. Bulo, P. Kontschieder, and M. Nießner. Diffrf: Rendering-Guided 3D Radiance Field Diffusion. In *Conference on Computer Vision and Pattern Recognition*, pages 4328–4338, 2023.

[51] T. Anciukevičius, Z. Xu, M. Fisher, P. Henderson, H. Bilen, N. J. Mitra, and P. Guerrero. Renderdiffusion: Image Diffusion for 3D Reconstruction, Inpainting and Generation. In *Conference on Computer Vision and Pattern Recognition*, pages 12608–12618, 2023.

[52] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-To-3: Zero-Shot One Image to 3D Object. In *International Conference on Computer Vision*, pages 9298–9309, 2023.

[53] M. Tyszkiewic, P. Fua, and E. Trulls. GECCO: Geometrically-Conditioned Point Diffusion Models. In *International Conference on Computer Vision*, 2023.

[54] L. Melas-Kyriazi, C. Rupprecht, and A. Vedaldi. Pc2: Projection-Conditioned Point Cloud Diffusion for Single-Image 3D Reconstruction. In *Conference on Computer Vision and Pattern Recognition*, pages 12923–12932, 2023.

[55] H. Bertiche, M. Madadi, and S. Escalera. CLOTH3D: Clothed 3D Humans. In *European Conference on Computer Vision*, pages 344–359, 2020.

[56] N. Pietroni, C. Dumery, R. Falque, M. Liu, T. Vidal-Calleja, and O. Sorkine-Hornung. Computational Pattern Making from 3D Garment Models. *ACM Transactions on Graphics*, 41(4):1–14, 2022.

[57] L. Liu, L. Zhang, Y. Xu, C. Gotsman, and S. J. Gortler. A Local/global Approach to Mesh Parameterization. In *Proceedings of the Symposium on Geometry Processing*, pages 1495–1504, 2008.

[58] C. Choy, J. Park, and V. Koltun. Fully Convolutional Geometric Features. In *International Conference on Computer Vision*, pages 8958–8966, 2019.

[59] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. McAllister, and A. Kanazawa. Nerfstudio: A Modular Framework for Neural Radiance Field Development. In *ACM SIGGRAPH*, 2023.

[60] Blender, 2018. https://www.blender.org/.

[61] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Conference on Medical Image Computing and Computer Assisted Intervention*, pages 234–241, 2015.

[62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is All You Need. In *Advances in Neural Information Processing Systems*, 2017.

[63] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimisation. In *International Conference on Learning Representations*, 2015.

 https://doi.org/10.52202/079017-1869

Appendix

A Additional Results

A.1 Qualitative Comparisons

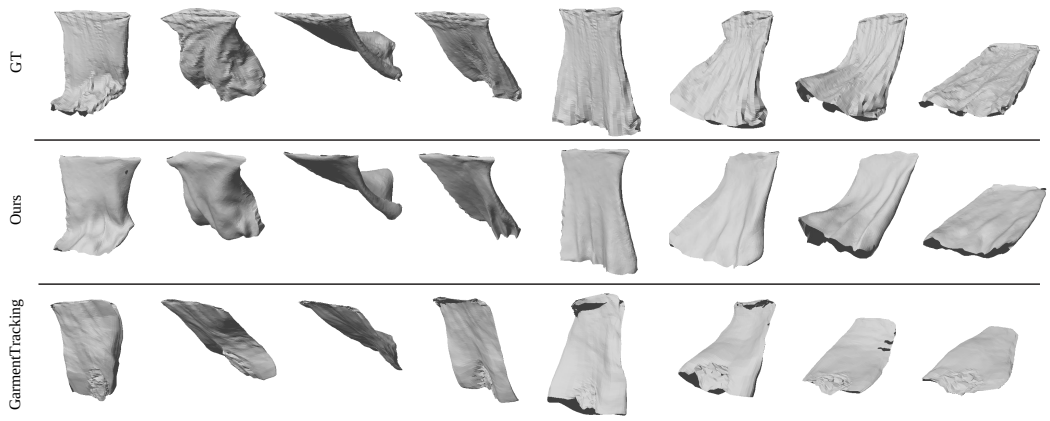


Figure 8: The comparison of reconstructed results for the flattening sequence of Skirt.

In Fig. 8 to 15, we provide additional qualitative comparisons between the results of our method and GarmentTracking [2] initialized with the ground truth garment mesh. Our reconstructions demonstrate higher accuracy and greater fidelity to the ground truth.

A.2 Evaluation on Real-World Data

Fig. 16 shows the qualitative results for the sweater. Similar to the results of the pants shown in Fig. 7 of the main paper, our method is able to reconstruct 3D meshes for both the flat and folded sweaters. However, some discrepancies still exist between the input and our reconstruction, which is attributed to the sim-to-real gap. Closing this domain gap will be an important direction of future work.

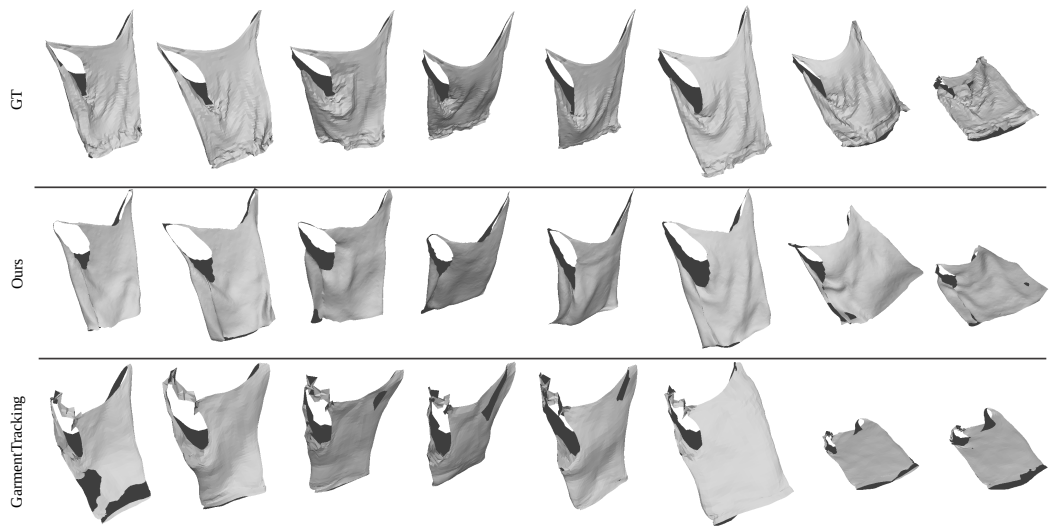


Figure 9: The comparison of reconstructed results for the flattening sequence of Top.

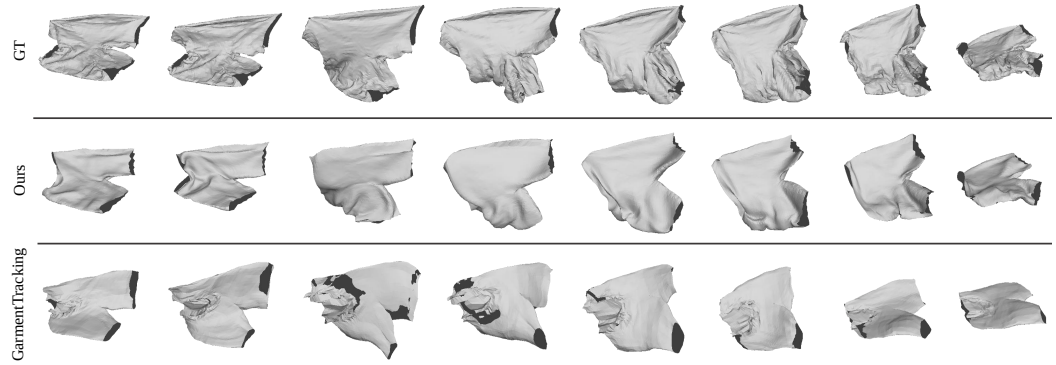


Figure 10: The comparison of reconstructed results for the flattening sequence of Pants.

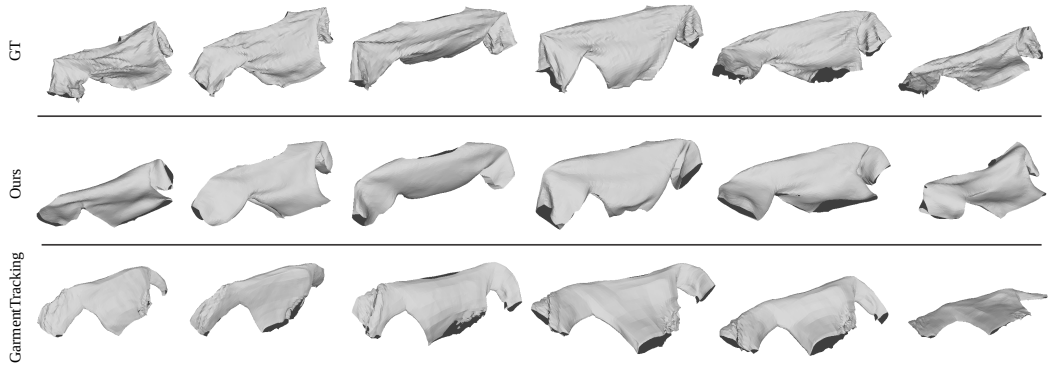


Figure 11: The comparison of reconstructed results for the flattening sequence of Shirt.

A.3 Evaluation of Intersections

In Table 2, we evaluate the intersections of our reconstructions and compare them with those of GarmentTracking [2] using the ground-truth initialization. We compute the average ratio of faces with intersection as the evaluation metric. Notably, our results exhibit fewer intersections compared to GarmentTracking on Pants, Top and Skirt.

https://doi.org/10.52202/079017-1869

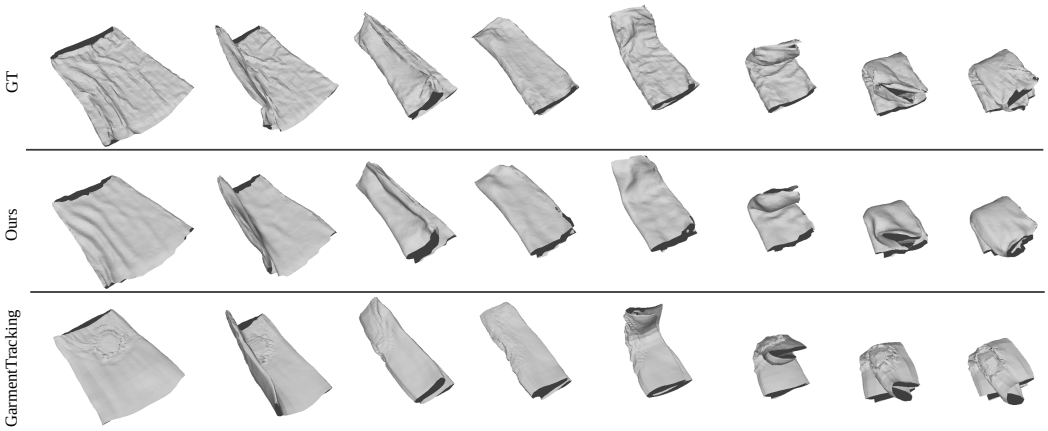


Figure 12: The comparison of reconstructed results for the folding sequence of Skirt.

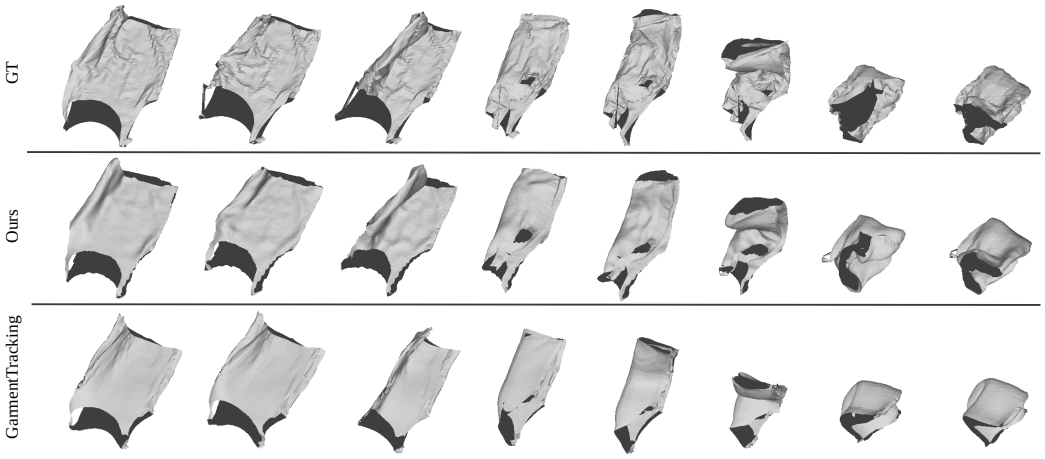


Figure 13: The comparison of reconstructed results for the folding sequence of Top.

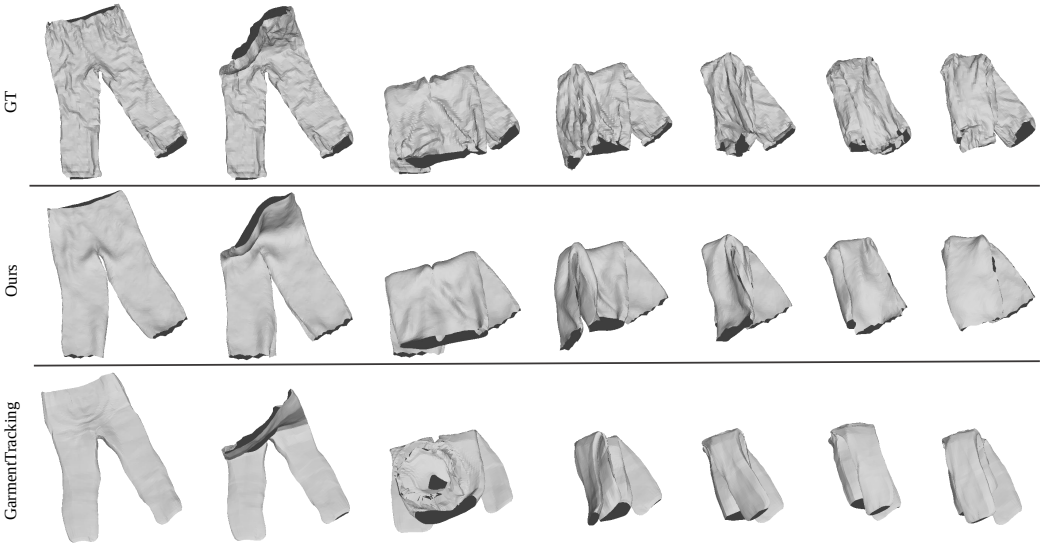


Figure 14: The comparison of reconstructed results for the folding sequence of Pants.

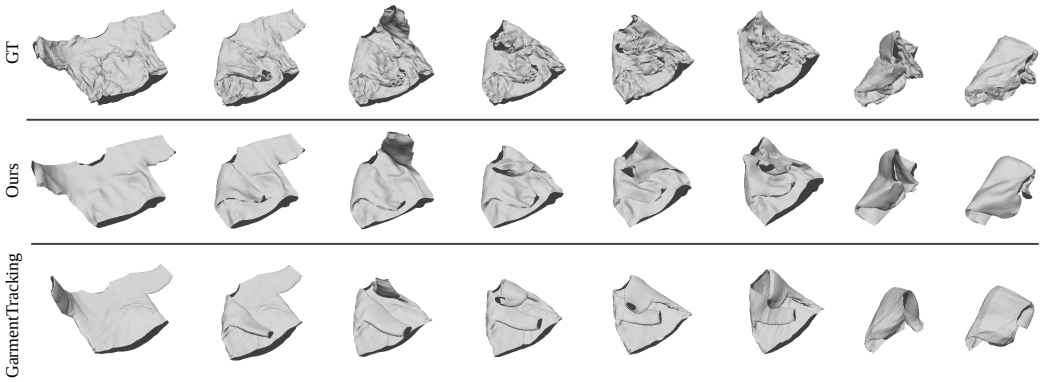


Figure 15: The comparison of reconstructed results for the folding sequence of Shirt.

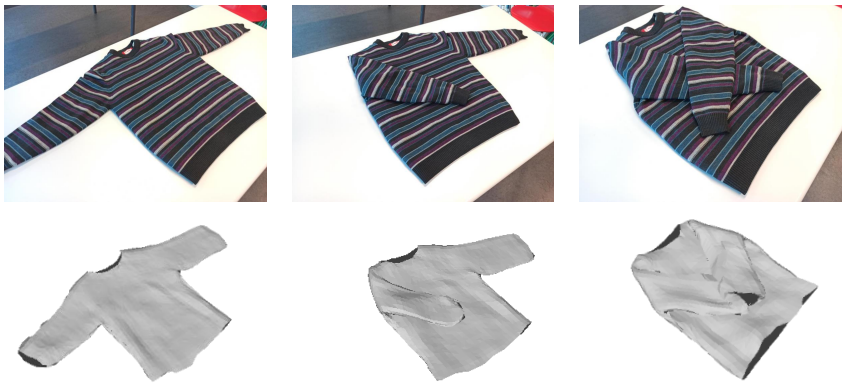


Figure 16: Qualitative results for folding a real sweater. Top: the images of the sweater. Bottom: our reconstructed results.

Table 2: The average intersection ratio of faces intersecting another face of our reconstructions compared to GarmentTracking [2]. GarmentTracking is initialized with the ground-truth mesh.

	Ours	GarmentTracking [2]
Pants	**1.9%**	3.8%
Shirt	2.2%	**1.6%**
Top	**1.8%**	6.4%
Skirt	**1.5%**	2.7%

A.4 Ablation Study

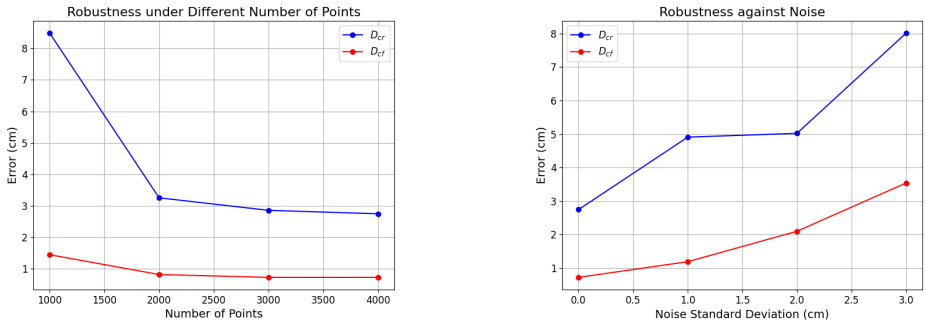
Table 3: **Ablation study**. +$\tilde{\mathcal{M}}$, +$\mathcal{O}$ and +$\mathcal{M}_{prev}$ denote using the guidance of the sparse UV maps, the recovered panel mask and the recovery of previous frames, respectively.

	$A_3 \uparrow$	$A_5 \uparrow$	$D_{cr} \downarrow$	$D_{cf} \downarrow$
+$\tilde{\mathcal{M}}$	61.5%	85.0%	3.55	1.12
+$\tilde{\mathcal{M}}$, +$\mathcal{O}$	68.7%	91.2%	2.81	0.80
+$\tilde{\mathcal{M}}$, +$\mathcal{O}$, +$\mathcal{M}_{prev}$	**71.3%**	**94.2%**	**2.62**	**0.72**

https://doi.org/10.52202/079017-1869

Table 3 presents the evaluation results of our fitting method of Eq. 14 of the main paper with different combinations of guidance on the test set of Top. Utilizing only the sparse UV maps $\tilde{\mathcal{M}}$ as guidance results in the lowest reconstruction accuracy. However, incorporating the guidance of the ISP-recovered panel mask $\mathcal{O}$, which provides garment geometry information, improves the results significantly. Moreover, by incorporating the recovery of the previous frame $\mathcal{M}_{prev}$, we further reduce the reconstruction error by introducing additional regularization on the unobserved surface.

A.5 Robustness



(a) (b)

Figure 17: **Quantitative results** of (a) using different numbers of points as input and (b) under different noise levels with 4000 input points. Blue: the Correspondence Distance D_{cr}. Red: the Chamfer Distance D_{cf}.

To evaluate the influence of point quantity, we analyze the reconstruction errors by varying the number of points used as input on the subset of Folding Pants. The results are reported in Fig. 17 (a). A reduction in points correlates with increased error. However, even with 2000 points, we maintain a relatively low error margin.

To evaluate the influence of input point noise, we add per-point Gaussian noise to the input with varying standard deviation. Fig. 17 (b) shows the results on the subset of Folding Pants. It illustrates that as the noise level rises, so does the reconstruction error; nonetheless, the errors remain relatively low across different noise levels.

A.6 Generative Samples

In Fig. 18, we show the deformed garment mesh generated by using the diffusion model to denoise the randomly sampled noise images. The resulting plausible deformations of the garment surfaces demonstrate the effectiveness of our diffusion model in capturing meaningful deformation priors.

A.7 Evaluation of Panel Mask Fitting

Fig. 19 shows the curve of the mean Intersection over Union (mIoU) between the ground truth panel masks and the masks fitted using Eq. 10 of the main paper. The x-axis represents the volume of the input point cloud, computed as the occupancy of the voxelized 3D space $[-1, 1]^3$ and normalized by the min-max normalization. As shown, a larger point cloud volume corresponds to a better fitting result reflected by a higher mIoU. This validates our choice of using the fitted panel mask of the frame with the maximum volume for the entire sequence.

A.8 Rotational Symmetry

As mentioned in Sec. 4.2 of the main paper, the rotational symmetry of the skirt can result in a relatively large Correspondence Distance (D_{cr}, A_d). Fig. 20 provides an illustrative example for this issue. Specifically, the front and back sides of the ground truth mesh are defined as Fig. 20 (c). However, our model mistakenly identifies the facing-up points of Fig. 20 (b) as the front surface,

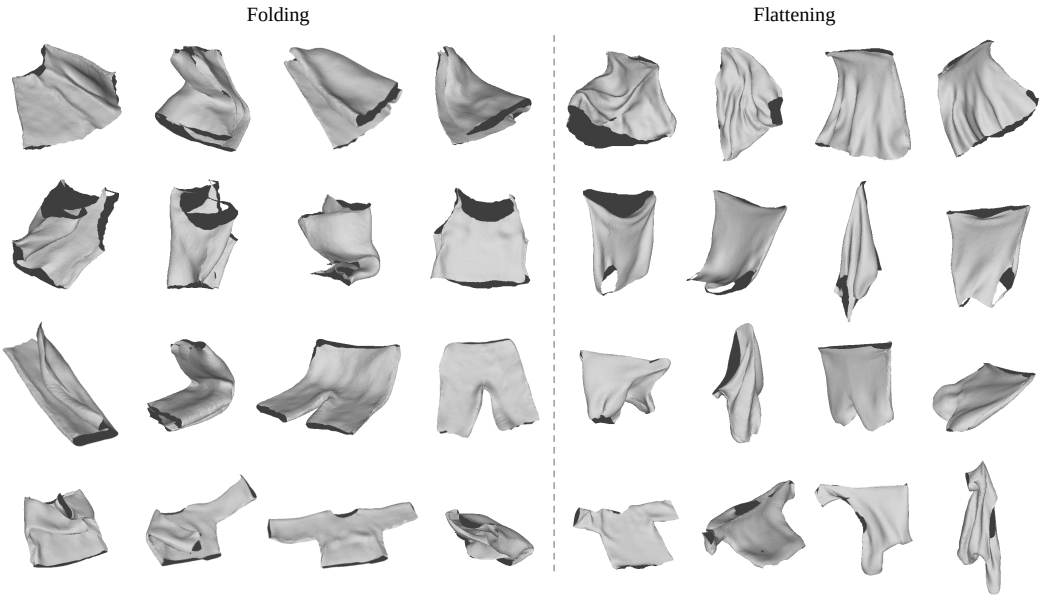


Figure 18: The generative samples for the categories of Skirt, Top, Pants and Shirt (top to bottom).

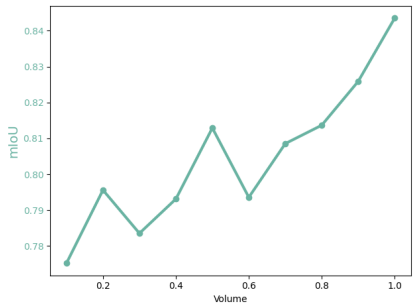


Figure 19: The evaluation of the panel mask fitting results.

yielding a D_{cr} of 19.67 cm. Despite this, as illustrated by Fig. 20 (d), our reconstruction maintains high quality, with a small Chamfer Distance (D_{cf}) of 1.21 cm.

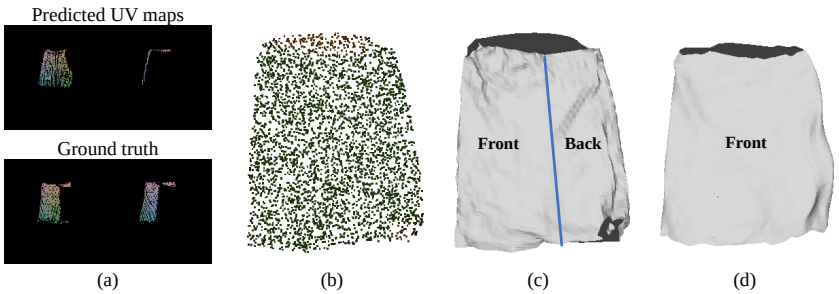


Figure 20: (a) The predicted sparse UV maps and the corresponding ground truth for (b) the input point cloud. (c) The ground truth mesh. (d) Our reconstruction. The front and back sides of the meshes are labeled with 'Front' and 'Back', respectively.

https://doi.org/10.52202/079017-1869

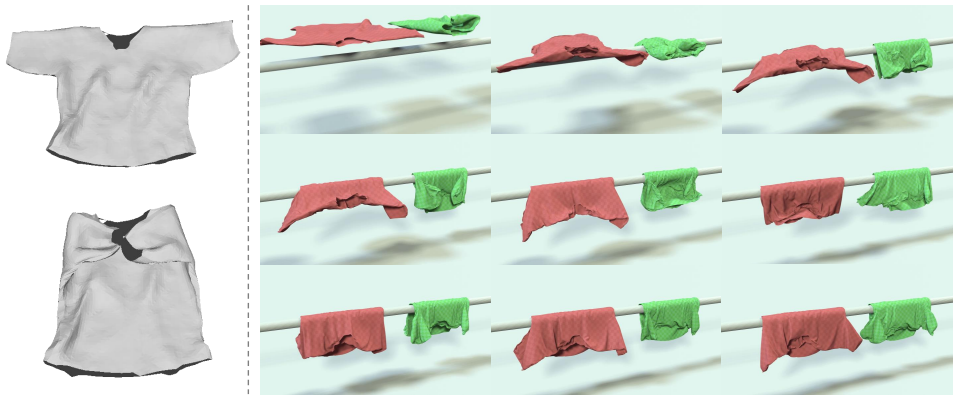


Figure 21: We use the reconstructed shirt meshes (left) to simulate the process of dropping them onto a bar (right).

A.9 Simulation

Our recovered meshes can be used for animation and simulation directly. In Fig. 21, we show the simulated results for the recovered shirts using Blender [60], where we drop them onto a horizontal bar.

B Implementation Details

Following [40, 1], the pattern parameterization network $\mathcal{I}_\Theta$ and the UV parameterization network $\mathcal{A}_\Phi$ of ISP are implemented as MLPs with the latent code $\mathbf{z}$ of size 128. We train $\mathcal{I}_\Theta$ and $\mathcal{A}_\Phi$ jointly for 9000 iterations with a batch size of 50. For the diffusion model, we adopt a U-Net architecture [61]. It takes the concatenated front and back UV maps and panel maps with the dimensions of $128 \times 256 \times 4$ as input. The diffusion model is trained for 100 epochs, with a learning rate of 1e-4, a batch size of 64, and $T = 1000$ steps. The UV mapper $\mathcal{G}$ consists of a sparse 3D CNN implemented as [2], a 6-layer Transformer encoder [62], and a 7-layer MLP for UV coordinate prediction. We choose $K = 128$ and train $\mathcal{G}$ for 100 epochs, using a learning rate of 1e-4 and a batch size of 128. To augment the data, we apply random rotations to both the point cloud and the mesh represented as UV maps. All the models are trained using the Adam optimizer [63] on NVIDIA A100 GPUs.

C Limitations

Due to the thin structure and the close proximity of surfaces during manipulation, self-intersections can occur. An interesting direction for future research is the exploration of learning intersection-free deformation priors with physics-based constraints.

The UV mapper currently relies on single-frame information for predictions. To enhance its performance, we intend to incorporate accumulated point cloud information across frames in future improvements.

Another limitation lies in the dataset used to train our models, which was recorded under controlled settings where participants followed specific manipulation procedures. However, this can be addressed by leveraging the recording platform of [2] to introduce a wider range of potential garment deformations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section C in the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/079017-1869

Justification: Not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Section 4.1 and Section B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release our codes and models upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 4.1 and Section B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Not available.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

https://doi.org/10.52202/079017-1869

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read the ethics review guidelines and ensured that our paper conforms to them.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work is about modeling the garment deformation which has no societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Section 4.1.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

https://doi.org/10.52202/079017-1869

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.