Garment3DGen: 3D Garment Stylization and Texture Generation

Nikolaos Sarafianos, Tuur Stuyck, Xiaoyu Xiang, Yilei Li, Jovan Popovic, Rakesh Ranjan
Meta Reality Labs

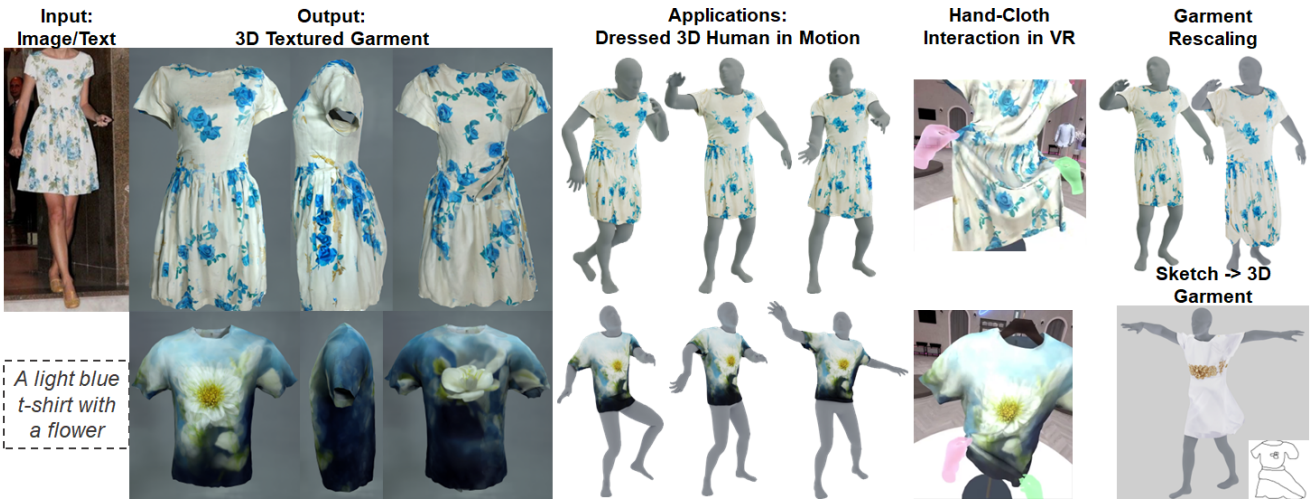

Figure 1. **Garment3DGen**, automatically transforms a base garment mesh to simulation-ready asset directly from images or text in a frictionless manner unlocking applications such as cloth and hand-cloth interaction in a VR.

Abstract

We introduce Garment3DGen a new method to synthesize 3D garment assets from a base mesh given a single input image as guidance. Our proposed approach allows users to generate 3D textured clothes based on both real and synthetic images, such as those generated by text prompts. The generated assets can be directly draped and simulated on human bodies. We leverage the recent progress of image-to-3D diffusion methods to generate 3D garment geometries. However, since these geometries cannot be utilized directly for downstream tasks, we propose to use them as pseudo ground-truth and set up a mesh deformation optimization procedure that deforms a base template mesh to match the generated 3D target. Carefully designed losses allow the base mesh to freely deform towards the desired target, yet preserve mesh quality and topology such that they can be simulated. Finally, we generate high-fidelity texture maps that are globally and locally consistent and faithfully capture the input guidance, allowing us to render the generated 3D assets. With Garment3DGen users can generate the simulation-ready 3D garment of their choice without the need of artist intervention. We present a plethora of quantitative and qualitative comparisons on various assets and demonstrate that Garment3DGen unlocks key applications ranging from sketch-to-simulated garments or interacting with the garments in VR. Details and code are provided in our project page.

1. Introduction

3D asset creation is the process of designing and generating geometries and materials for 3D experiences. It has direct applications across several industries such as gaming, movies, fashion as well as VR applications. Traditionally, simulation-ready garments are hard to obtain and are created through a laborious time-consuming process requiring specialized software [9, 15, 78] relying on experienced artists. Currently, creating virtual clothing for simulation is a challenging task. Garments need to be manually designed and draped onto an underlying body. Additionally, the topology of the garment needs to take simulation requirements into consideration in order to enable stable and visually pleasing results. Low-friction asset creation will be the key enabler to unlock virtual applications at scale. Generative AI will be a cornerstone technology that will allow anyone, from novice users to experts, to create customized avatars and to contribute to building personalized virtual ex-

periences. In addition, it will assist in the design process to facilitate faster exploration and creation of new designs.

To tackle this task, we set out to develop a method that creates 3D garments directly from images. Given a base geometry mesh and a single image, Garment3DGen performs topology-preserving mesh-based deformations to match the image guidance and synthesizes new 3D assets on the fly. Our generated garments comprise of posed geometries that stylistically match the input image, and high-resolution texture maps. The provided image guidance can be either from the real world or synthetically generated [17, 68] which enables us to create both real and fantastical 3D garments. One way to tackle this problem would be to utilize recent image-to-3D techniques [50]. Given a single image as input, such methods synthesize a specific number of views captured from pre-set viewpoints and then employ multi-view reconstruction techniques to obtain the 3D asset. However, the output geometries tend to be coarse and lack fine-level details due to the use of Marching Cubes to extract the output 3D geometry. Another limitation is that the output garments are watertight and have arbitrary scale, making it difficult to drape them on human bodies and simulate them directly. This is because manual intervention would be required to post-process the geometry. For example, creating arm, neck, and waist holes for a t-shirt geometry, as well as re-meshing to produce mesh topologies suitable for downstream applications which are sensitive to poor mesh qualities. Alternatively, one could follow a NeRF-based approach where a handful of views of the base mesh are utilized to train a NeRF which can be stylized in an iterative manner [32]. However, such an approach does not guarantee multi-view consistency of the newly stylized garment because the NeRF training and stylization are happening in an iterative manner. In addition, it is a time-consuming process and the final result is not suitable for simulation. Recent works have focused on 3D Gaussian Splatting [38] to generate 3D assets from image inputs. While such methods are fast and of produce high reconstruction quality, their output splats are hard to be used for any downstream task besides rendering. Another direction of research predicts 2D garment patterns [7, 18, 39] which can be optimized using differentiable simulation [12, 46]. Such approaches generate simulation-ready garments but cannot generalize to fantastical AI-generated garments and are constrained to specific garment styles. For example, when testing SewFormer [49] on upper-body garments, it only works well on input with V-neck shirts, which is very limiting; while for other garments (long-sleeve with round neck) it fails to generate accurate patterns.

To this end, we carefully designed Garment3DGen to tackle each of the aforementioned challenges: i) reconstruction-based approaches output geometries that are watertight, coarse and the garments cannot be draped on human bodies ii) deformation-based approaches are under-constrained when given a single image or text prompt and their outputs do not faithfully match the provided guidance and iii) simulation-based approaches fail to generalize to new garment types. Garment3DGen produces high-quality simulation-ready stylized garment assets along with the associated textures. We approach this task from deformation-based perspective as we believe that it provides better properties and more fine-grained control for the output geometries compared to alternative NeRF-based or reconstruction-based approaches. Mesh-based deformations can preserve the mesh topology which in turn can allow for UV texture transfer, they can preserve the arm/body/head holes of the garment geometry instead of outputting watertight meshes, and can provide output meshes the triangles of which are not distorted and can be draped on human bodies and simulated. Our method takes as inputs a single image and a base template mesh and outputs a deformed mesh that faithfully follows the image guidance while preserving the structure and topology of the base mesh. Our first contribution stems from supervising the mesh deformation process directly in the 3D space which directly allows physics-inspired losses that ensure simulation-readiness instead of solely relying on image-based or embedding-based supervisions [58, 60, 90]. In the absence of 3D ground-truth, we build upon the progress of diffusion-based multi-view consistent image generation to obtain a coarse 3D geometry that can serve as pseudo ground-truth. However, strictly enforcing 3D supervisions using a coarse mesh results in deformed meshes that lack fine-level details and are not simulation-ready. Thus, we utilize a pre-trained MetaCLIP model [90] finetuned to garment data and introduce additional losses in the image as well as the embedding spaces using differentiable rendering. Finally, we propose a carefully designed texture estimation module to predict the texture maps required to create the final 3D garment.

We conducted a plethora of experiments that demonstrate that our method generates 3D garment assets i) directly from images allowing a frictionless experience where users indicate a requested garment by providing a reference image and quickly obtain a high-quality 3D asset without manual intervention and ii) from text describing both real and fantastical garments, and iii) even garment sketches that one can quickly draw. Moreover, we have developed a body-garment co-optimization framework that enables us to scale and fit the garment to a parametric body model. This allows us to animate the body model and perform physics-based cloth simulation, resulting in a more accurate representation of the garment's behavior in various novel scenarios. In summary, our contributions are as follows:

- We propose a new deformation-based approach for 3D geometry and texture generation for garments given a base mesh and a single image guidance as inputs.

- We introduce geometry supervisions directly on the 3D space by generating coarse-guidance meshes from the image inputs and use them as soft constraints during the optimization. In addition, we provide valuable insights on the impact of different losses that ensure that the output geometries are suitable for downstream tasks such as cloth-simulation or hand-garment interaction.
- We introduce a texture enhancement module that generates high-fidelity UV textures from a single image allowing us to render the output geometries.

2. Related Work

Garment Modeling: An important line of work is focused on designing [8, 20, 85], capturing [95], registering [30], reconstructing [16, 36, 45, 61, 66, 92], and representing [53, 80] clothes and their texture [10] from images or videos. BCNet [36] predicts garment vertex displacements whereas Moon et al. [61] map 3D garments to a parametric body using dense keypoints. Both these approaches cannot generalize to loose clothes moving freely. Guo et al. [29] learns a shape diffusion-based prior from captured 4D data in order to enable registration of texture-less cloth, whereas Lin et al. [47] aligns the garment geometry to real world captures using a coarse-to-fine method that leverages intrinsic manifold properties with neural deformation fields. Aiming to model clothes and discover compact ways to represent them, CaPhy [79] recovers a dynamic neural model of clothing similar to SNUG [70] by leveraging 3D supervised training in combination with physics-based losses. Xiang et al. [88] introduced a physically-inspired appearance representation by learning view-dependent and dynamic shadowing effects. Finally, recent methods explored modeling clothes using graph neural networks [27, 31, 62]. An alternative way to represent clothes is via sewing patterns as it ensures an efficient representation of developable [69, 76] and manufacturable garments which can be easily modified. A plethora of works [2, 6, 64] have followed this path for garment reconstruction [49, 82, 83], generation [33, 52, 72] and draping [44].

Garment Deformation and Stylization: The recent progress in language and image-to-3D models has unlocked new ways of representing and reconstructing dressed avatars [37, 84] from text or images. Such methods typically generate multi-view consistent views [50, 51, 54, 65, 87] given a single image or text input or directly optimize a 3D scene [63] using a 3D scene parameterization, similar to Neural Radiance Fields [59]. However, such methods generate coarse, watertight meshes that in the context of garments do not have the required topology and structure to be draped on humans and simulated [77]. Editing and stylizing 3D surfaces has been explored for optimizing directly in 3D [4, 5, 26, 40, 48, 74] and more recently using triplanes [22] and text-to-mesh formulations [13, 60]. For example, Michel et al. [58] performs mesh stylization by predicting color and local geometric details that follow a text prompt. Deformation-based approaches [3, 23, 24, 28, 35, 56, 81, 86, 94, 97] can leverage these foundation models to enforce supervision for text and image-based stylization [19] and manipulation [25] of 3D meshes. Recent work [11, 67, 93, 96] applied text-to-image generation models to create textures based on the mesh and given text/image. Extending such techniques to clothes is a complex task as the supervision signals of a single image or text-prompt are insufficient to ensure that the deformed clothes will be simulation-ready.

3. Methodology

Our approach takes as input a single image I and a base garment template mesh M_{in} and performs a topology-preserving deformation of the input geometry given the image guidance to obtain the target deformed mesh $M_{\text{def}} = D(I, M_{\text{in}})$ where D is a function that optimizes over the input mesh. An overview of our method is depicted in Fig. 2.

3.1. Target Geometry Generation

We propose to leverage the recent progress of single-image-to-3D methods to obtain a coarse geometry of I denoted by $M_{\text{guide}}(I)$ and use it as much stronger supervision both directly in the 3D space as well as in projected 2D space through differentiable rendering. A cross-domain diffusion model [54, 73] is employed which synthesizes RGB and normal images from six views given the input image I captured from the same predefined viewpoint. A multi-view 3D reconstruction algorithm based on the LRM architecture [91] is then utilized that, given the generated views, it outputs a watertight, relatively coarse geometry $M_{\text{guide}}(I)$ of the garment in the input image. Such meshes cannot serve as the final simulation-ready result due to its poor mesh quality which is due to Marching Cubes [55] (or FlexiCubes [71]) or potential inaccuracies of the multi-view generation. Additionally, the fact that it is watertight prevents us from draping the garment on a body (*e.g.* missing armholes). Nonetheless, it provides useful information to serve as a pseudo ground-truth that serves as a reference to deform M_{in} towards. Hence, we update the optimization function as follows $M_{\text{def}} = D(I, M_{\text{in}}, M_{\text{guide}}(I))$. The alternative approach would be to rely on the input image I as the sole supervision for mesh deformation which would result in a severely under-constrained optimization with low-quality output meshes that are uncanny, over-deformed, and they fail to capture the subtle details of the image guidance. For example, starting from a template T-shirt geometry with guidance of the image of an armor, one could extract CLIP embeddings for both I and the renders of M_{def} following a similar approach to TextDeformer [25]. By enforcing supervisions on the embeddings, the goal is obtain an output mesh that would resemble the requested armor. In practice,

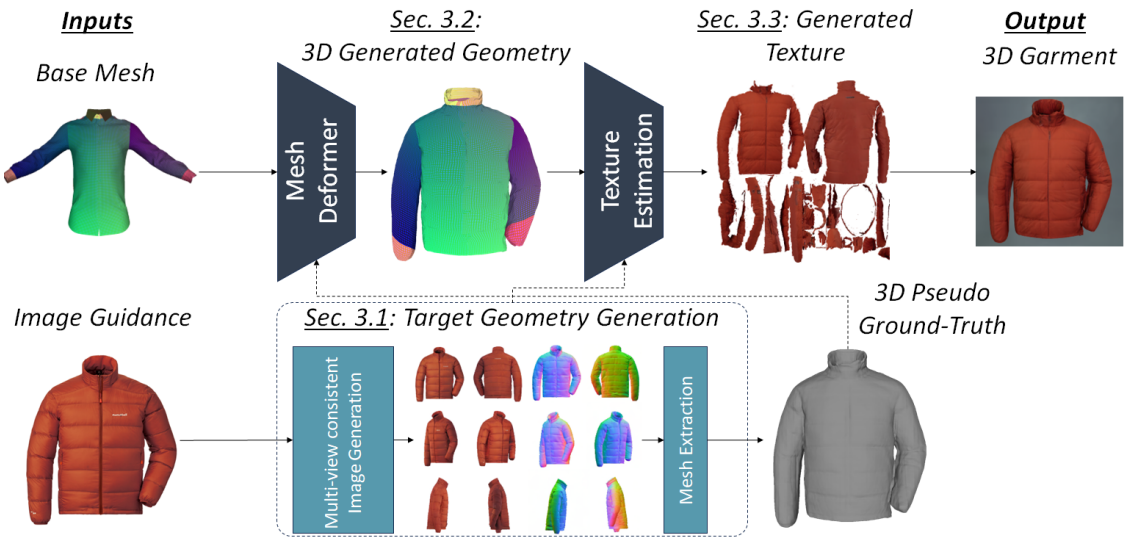

Figure 2. **Overview**: Given an input 3D base mesh and a target garment image we first generate 3D pseudo ground-truth using a diffusion-based method and utilize the output geometry as a soft supervision signal during the deformation process. Our 3D generated geometry preserves the topology and structure of the base mesh as depicted by the colors of the sleeves/collar while accurately reflecting the geometry and details of the input image. Finally, we introduce a texture-estimation module which outputs the corresponding UV texture that along with the geometry comprise our final generated 3D garment.

the supervision from the embedding of a single exemplar image is not strong enough to produce a high-quality output.

3.2. Topology-Preserving Deformations

In order to preserve the structure and topology of the input base mesh while enabling image-base stylizations, we propose an approach which deforms M_{in}. In contrast, generating novel geometry using reconstruction-based methods would be hard to use directly in downstream tasks. Hence, inspired by Neural Jacobian Fields [1], we parameterize M_{in} using a set of per-triangle Jacobians which define a deformation. Following the same formulation, for simplicity we represent per-triangle Jacobians as matrices $J_i \in R^{3x3}$ and solve a Poisson optimization problem to obtain the deformation map Φ^* as the mapping with Jacobian matrices for each triangle that is closest to J_i. More formally, this is represented as:

$$\Phi^* = \min_{\Phi} \sum |t_i| \|\Phi \nabla_i^T - J_i\|^2, \tag{1}$$

where $\nabla(\Phi)$ denotes the Jacobian of Φ at triangle t_i, with $|t_i|$ being the area of the triangle. We optimize the deformation mapping Φ indirectly by optimizing the matrices J_i which define Φ^*. These Jacobians are initialized to identity matrices. With the Jacobian representation at hand, we optimize over the triangles of M_{in} by introducing a several losses each addressing a specific issue.

3D Supervisions: We employ the one-directional Chamfer Distance (CD) loss to evaluate the similarity between sets of points $p_{\text{def}} \in S_{\text{def}}$ and $p_{\text{I}} \in S_{\text{I}}$ sampled randomly in each iteration from M_{def} and $M_{\text{guide}}(I)$. This is defined as:

$$L_{\text{CD}} = \frac{1}{|S_{\text{def}}|} \sum_{p_{\text{def}} \in S_{\text{def}}} \min_{p_{\text{I}} \in S_{\text{I}}} \|p_{\text{def}} - p_{\text{I}}\|_2^2. \tag{2}$$

Regularizations: We introduce several regularizations on the deformed 3D mesh to ensure that it maintains key properties. First, we introduce Laplacian smoothing [21] denoted by L_{Lap} to redistribute vertex positions based on the average positions of neighboring vertices. This smoothing process helps reduce irregularities and improves the overall mesh shape. To produce simulation-ready meshes we penalize very small surface area triangles denoted by L_{triag} as that would result in meshes that are difficult to simulate. We do this by regularizing the edge length and by minimizing the inverse of the squared sum of the triangle areas. Note that there is a trade-off between how much a mesh can freely deform (*e.g.* a shirt becoming a spiky medieval armor) and how much regularization it requires such that the garment can be placed on a parametric body and simulated.

2D Supervisions: We utilize a rasterization-based differentiable renderer [41] denoted by R and pass both the deformed mesh M_{def} in each iteration and target pseudo ground-truth mesh $M_{\text{guide}}(I)$ to obtain K image renders $I_{\text{def}i} = R(M_{\text{def}}, C_i), \quad i = 1 \dots K$ from randomly sampled camera views C_i. With $I_{\text{I}i}$ computed in a similar fashion for $M_{\text{guide}}(I)$ we employ the L1 loss between the deformed and target renders:

$$L_{\text{2D}} = \frac{1}{K} \sum_{i=1}^{K} \|I_{\text{def}i} - I_{\text{I}i}|. \tag{3}$$

This supervision in the 2D space captures well the silhouette of the garment from multiple views as well as its fine-level details thereby enforcing the deformed mesh to not deviate far from the target along each step of the optimization.

Embedding Supervisions: We observe that passing garment images through a pretrained CLIP model results in deformed output garments that are overly distorted, uncanny, and fail to capture the fine-level details provided in the input images. This is due to the weak supervision signal contained in these embeddings. Likewise, this holds true for other mesh classes that are not well represented in the data on which CLIP was trained on, such as human geometries for example. This is because CLIP fails to capture the subtle differences between garments, their properties, and materials. To overcome this limitation, we propose to use garment-specific embeddings obtained from a CLIP model fine-tuned on fashion data named FashionCLIP [14] denoted by FCLIP. The latent space for this model is better tuned for fashion concepts, and as a result, the embeddings provide a stronger guidance for the deformations. We represent this loss as follows:

$$L_{\text{E}} = \frac{1}{K}\sum_{i=1}^{K} \text{Cos}\left(\text{FCLIP}(I_{\text{def}i}), \text{FCLIP}(I_{\text{I}i})\right), \quad (4)$$

where *CosSim* is the cosine similarity. This embedding loss acts as a soft supervision signal between the embeddings of the deformed mesh M_{def} and those of the pseudo-ground-truth $M_{\text{guide}}(I)$. This behavior is desired since we aim to benefit from the embedding representations of the CLIP model and sufficiently deform the base mesh towards the target without capturing all of its shortcomings. For example we need to preserve the arm/head holes of the base geometry while the target mesh is watertight, or capture the fine-level details depicted in the input image that $M_{\text{guide}}(I)$ might have failed to represent well. In summary, the total loss L_{T} is defined as follows where w_* is the corresponding weight for each loss:

$$L_{\text{T}} = w_{\text{CD}}L_{\text{CD}} + w_{\text{Lap}}L_{\text{Lap}} + w_{\text{triag}}L_{\text{triag}} + w_{\text{2D}}L_{\text{2D}} + w_{\text{E}}L_{\text{E}}, \quad (5)$$

where the corresponding weights balance between utilizing the 3D pseudo ground-truth with the CD loss while capturing the finer details through the embedding and 2D losses.

3.3. Texture Estimation

Given the untextured deformed 3D geometry M_{def}, we generate high-fidelity textures that match the input image. We leverage a 2D text-to-image generation model to create high-quality, high-resolution textures with vivid details given a text prompt. When starting from images where text is absent, a text prompt is obtained using an image caption model [43]. Our pipeline builds on the work of Xiang *et al*. [89] and consists of the following steps: generating images of the given fashion asset from multiple views and backprojecting these images onto the mesh surface to create UV texture $T \in \mathbb{R}^{H\times W\times C}$. There are two challenges in adapting a 2D generation model to 3D objects: establishing geometry-texture correspondence and ensuring multi-view consistency.

Shape-Aware Generation: To ensure that the texture faithfully reflects the underlying shapes, we propose using a depth-aware text-to-image generation model. Given a set of camera poses $\mathcal{C} = \{C_i\}_{i=1}^{n}$, we render the depth D_i of each view from M_{def}, and sample the appearance image I_i of the view C_i conditioned on D_i and the text prompt using the text-to-image model.

Multi-view Consistency: Directly conducting view-by-view image synthesis cannot guarantee that the generated views are consistent with each other. To solve this problem, we first synthesize the front and back views simultaneously to implicitly enforce global consistency. Then, we leverage the fact that the geometry of fashion assets is mostly flat, and conduct depth-aware inpainting for the remaining views to ensure the newly generated textures are locally consistent. To handle occluded areas, we design an automatic view selection algorithm to inpaint the textures in a coarse-to-fine manner: from the remaining unpainted area, we select the view with the most unfilled pixels to generate the textures iteratively from large regions to small pieces.

Texture Enhancement: The above texture generation pipeline can also be applied to texture refinement: given a low-quality, low-resolution initial texture T^{LQ}, we leverage the 2D appearance priors to further enhance the details. We adopt SDEdit [57], which perturbs the above sampling procedure with Gaussian noise and progressively denoises by simulating the reverse stochastic differential equations. As a result, the low-quality I_i^{LQ} is projected onto the manifold of realistic images, yielding I_i^{HQ}. By backprojecting these images, we acquire a high-quality texture T^{HQ}.

Our texture estimation module has the following improvements compared with the previous works:

- *Global consistency*: by generating the front and back views together, we enforce better global consistency.
- *Geometry-aware texture inpainting*: Past works tweaked the depth-to-image generation model with masked generation methods (TEXTure [67] uses Blended Diffusion, and Text2Tex [11] uses RePaint). We directly use multi-ControlNet, enabling a better balance between the depth and mask controls, which produces better local content consistency.
- *Backprojection*: we implement a faster backprojection method using PyTorch $grid\_sample$, which has the lowest latency compared to all known methods.
- *Speed*: Our texture generation is significantly faster than: TEXTure: 2.5 mins, Text2Tex:12mins, Ours: 4.56s

We refer the reader to the supplementary for an in-depth discussion of trade-offs.

Figure 3. **Qualitative Comparisons**: We demonstrate several mesh generation methods given an image (top) or text (bottom) and show front and back views of each reconstruction. The 3D Gaussian Splatting [101] method generates distorted frontal colors and dark or blurry back colors while its geometry is not suitable for downstream tasks such as simulation. The second reconstruction approach [54, 100] generates watertight meshes with coarse geometric details and blurry colors. WordRobe[75] which is purely text-based, generates simulation-ready garments but they deviate far from the text prompt (*e.g.*, the puffer jacket and t-shirt are not faithful to the prompt). Our proposed approach outputs 3D geometries that are geometrically correct with fine-level texture details that prior works fail to generate.

4. Experiments

Data: We use the publicly available DiffAvatar [46] dataset which comprises of artist-created garment templates covering several cloth categories (*e.g.* T-shirt, shirt, tank-top, dress, etc). To demonstrate garment generation using image guidance, we collect a variety of real images with garments in different poses, different textures, including garment types that are not covered by our dataset. In addition to real images, we include AI generated images using a text prompt of both realistic and fantastical garments.

Metrics: Quantitatively evaluating our results is a challenging task in the absence of 2D or 3D ground-truth for what are we trying to accomplish. However we can evaluate how consistent the geometries are to the input image and hence we render untextured outputs for all methods from 36 views and compute their perceptual scores using the LPIPS metric [99] as well as their image-based CLIP similarity score using the cosine distance between their embeddings.

Baselines: We evaluate our approach against: i) TextDeformer [25] which deforms input meshes based on text, ii) ImageDeformer - a variation of TextDeformer we developed where the input text is replaced with an image, iii) Wonder3D [54] and iv) Zero123++ [73] both of which generate 3D geometries given a single image using 2D diffusion models, v) ZeroShape [34] which performs zero-shot reconstruction and vi) a triplane-based 3DGS approach [101]. The first two works also take as input a base mesh whereas the latter four reconstruct the final result given a single image as input. Finally, we perform qualitative comparisons with WordRobe [75] which is a recent top-performing method among some concurrent text-based garment generation methods [33, 42, 52] and a pattern-based approach termed SewFormer [49] (see supplementary).

4.1. Image-to-3D Garments

Given an image prompt and a base mesh, Garment3DGen generates textured 3D garments matching the input image guidance. We present a plethora of image-to-3D results in Fig. 1, Fig. 3 and Fig. 5 and showcase that the generated 3D assets can be of various topologies, respect the pose, shape and texture of the input image. Garments can be both real or fantastical. They contain high-quality texture maps which exhibits fine-level details such as cloth wrinkles and folds and preserves the topology and structure of the input (*e.g.*, head, bottom and armholes are respected such that the garments can be draped on a body). Additionally, we quantitatively evaluate Garment3DGen against several baseline methods and report our results in Table 1. Our approach outperforms all methods in terms of both embedding as well as perceptual similarity with the input image guidance while it is the only approach that generates textured geometries that can be used for downstream tasks. Deformation-based approaches (rows 1-2) preserve the topology and the holes of the garment but lack strong supervision signals to generate outputs that match the in-

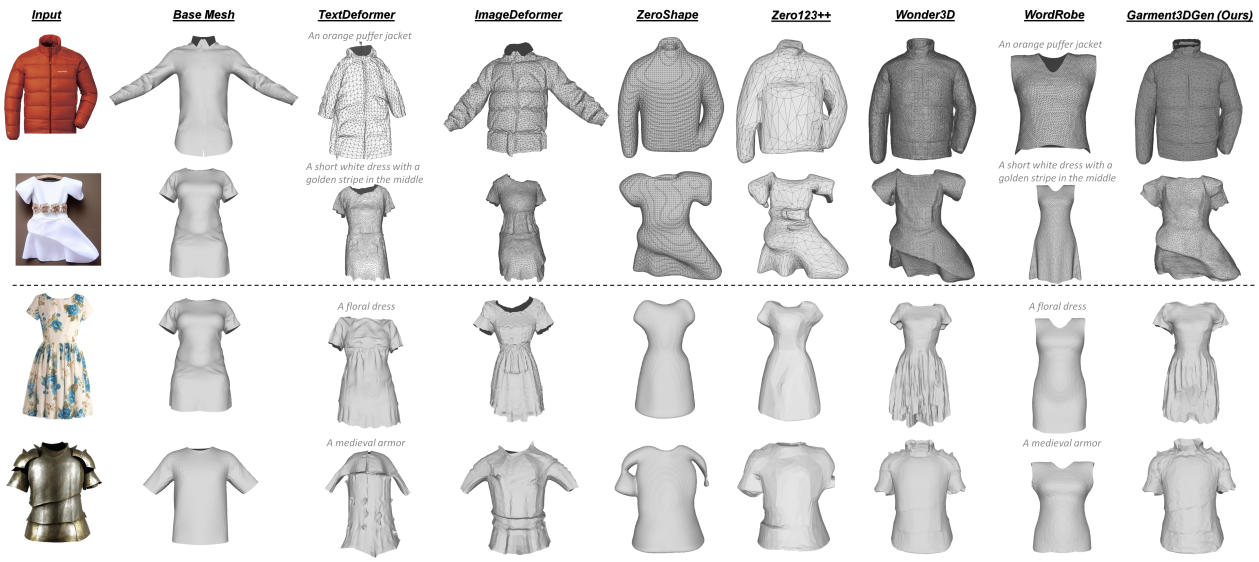

Figure 4. **Mesh Quality (Top) and Geometry Comparisons (Bottom)**: We showcase the wireframes of all approaches. Our method stands out as the only one that produces geometries that adhere to the input image while maintaining good mesh quality and incorporating necessary holes for physics-based simulation tasks. At the bottom we showcase the output geometry of various techniques to highlight that our approach captures fine geometric details without geometric artifacts (Wonder3D). Note that several methods produce smooth meshes where details are incorrectly captured in the texture map instead which results in lower quality visual results.

put image. Reconstruction-based approaches (rows 3-6) are good at preserving and reconstructing what is visible in the input image but generate unusable geometries (or splats) and their color estimates for the non-visible regions are usually blurry or single colored (as shown in Fig. 3).

4.2. Text-to-3D Garments

We easily extend our work to enable textured garment generation using textual prompt inputs. Unlike TextDeformer [25] or WordRobe [75] which utilize text directly, we opt for a text-to-image diffusion model as an initial step. We do this because image-based supervisions provides a stronger guidance for our mesh deformation process which produces better results. This allows the user to easily iterate with several text-prompts to generate the image of their desired garment . This is in stark contrast to the slow iteration time of providing a text-prompt and waiting for the 3D asset creation process to finish to see if the result matches their intent. In Fig. 3 (bottom) we present generated garments using only the provided textual prompt as guidance.

4.3. Ablation Studies

We conducted a variety of ablation studies to assess the impact of the key components of our proposed approach the figures of which are provided in the supplementary material. We start with the off-the-shelf TextDeformer which takes a text prompt and a base-mesh and deforms this to match the target text. Text prompts are not ideal to capture the fine-level details of a garment as there can be many "medieval armors". In addition, a pretrained CLIP model is not capable of capturing the subtle differences between a "jacket" and a "puffer jacket". To overcome this limitation we adapt TextDeformer to take image inputs as guidance (ImageDeformer) and observe that the deformed geometries are improved. Nonetheless, they still fail to capture the details of the image. By swapping the original CLIP model and introducing a model that is fine-tuned on fashion data we observe that details are better preserved across garments. Noting that image-based reconstruction methods can accurately capture geometry but produce coarse and watertight meshes that are unsuitable for subsequent tasks, we utilize these meshes as pseudo ground-truth for our proposed approach. Our Garment3DGen results in garments that faithfully follow the image guidance while containing the wrinkles and fine details. However the quality of the output geometry is not always ideal for physics-based downstream tasks be-

Table 1. **Quantitative Comparisons**: Our approach outperforms deformation-based (rows 1,2) and reconstruction-based (rows 3-6) methods across both metrics while generating textured geometries that can be used for downstream tasks which is not the case for any of the prior methods.

Method	CLIP-Sim↑	LPIPS ↓	Faithful to Image	Colored Output	Head/Arm Holes
TextDeformer [25]	0.51	0.42			✓
ImageDeformer [25]	0.54	0.41			✓
Wonder3D [54]	0.56	0.41	✓	✓	
Zero123++ [73]	0.52	0.42	✓	✓	
ZeroShape [34]	0.48	0.46	✓		
T-3DGS [101]	0.57	0.41	✓	✓	
Garment3DGen	**0.59**	**0.39**	✓	✓	✓

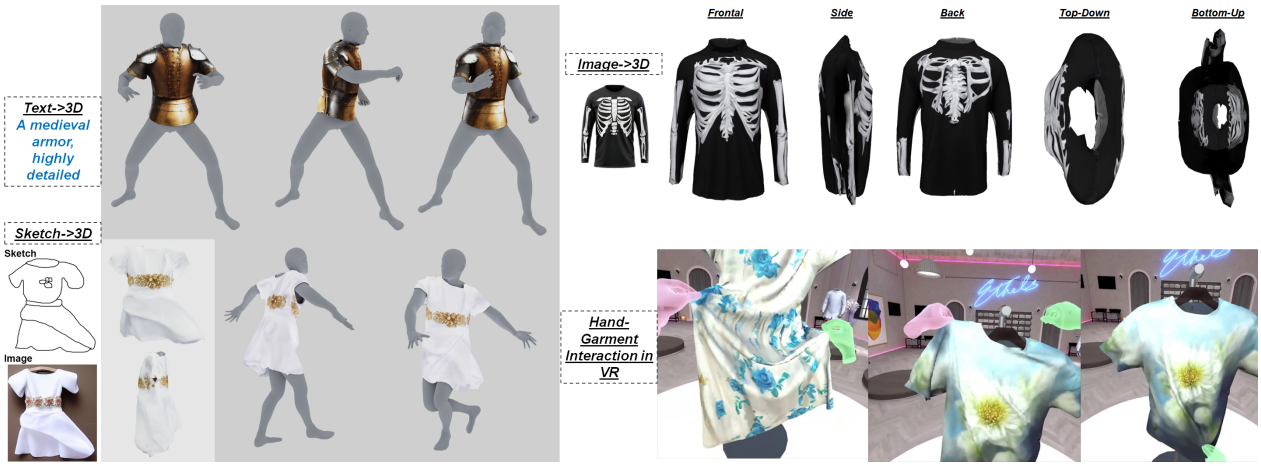

Figure 5. **Applications**: Garment3DGen generates textured 3D garments from images, text prompts, simple sketches, that can be fitted to human bodies and drive them with physics-based cloth simulation or even enable interaction between hands and garments in VR.

cause they produce poorly conditioned triangles which will result in instabilities when simulated in addition to poorly tessellated geometry which will result in unnatural fabric behavior. Because of this, we introduced additional 3D supervisions that preserve a better mesh quality, see Fig. 4. All prior works either do not follow the input image guidance or generate low-quality geometries that cannot be directly used for downstream tasks. Finally we conducted ablation studies to assess the impact of the pre-trained CLIP model on garment data, the impact of regularizations as well as the texture enhancement module and refer the interested reader in the supplementary material for additional information.

4.4. Applications

Physics-based Cloth Simulation We consider the generated 3D shape as rest shape for our simulations and incorporate a zero rest-angle dihedral energy to model the out-of-plane banding of the fabric. Fantastical garments such as armors are not well modeled using a zero rest-angle as they would lose their distinct shape during simulation and wrinkle unnaturally. To obtain visually pleasing results, we take the rest angle to be the 3D generated mesh one which allows it to maintain its shape throughout the simulation. Generated garments can further be manipulated to achieve distinct looks through garment resizing.

Hand-Cloth Interaction in VR Our garments are suitable for real-time simulations with hand-interaction using modern VR headsets. The rightmost part of Fig. 1 and Fig. 5 show a user interacting with the garments through real-time simulation with integrated hand-tracked interactions.

Sketch to Garment: Given a rough dress sketch, we generate a realistic image using ControlNet [98] which then serves as the input to our method to generate the corresponding 3D asset. In Fig. 5 we show simulation results of this automatically generated dress.

4.5. Discussion

Simulation-ready Garments: Our method stands out as the favorable choice for producing garments that are faithful to the prompt with high-quality meshes that can be seamlessly integrated into simulation frameworks as demonstrated by a qualitative and quantitative analysis. Although it might be possible to produce simulation-ready garments using alternative methods, this would require remeshing and potentially several additional manual post-processing steps, such as creating arm holes. Moreover, our use of template meshes ensures a superior UV layout, which is challenging to attain automatically.

Runtime: Garment3DGen takes ~5 minutes on a single H100 which is 3x faster than [25, 42, 58] yet slower than 3DGS-based methods that take a few seconds. While the latter render well, how to accurately re-mesh them and fit them onto bodies is an open research problem.

5. Conclusion

We proposed a new approach to generating high-quality garment assets that can be directly used for downstream applications which require good mesh quality. We introduced a deformation-based approach that takes a base mesh as input along with an image guidance and outputs textured geometry that faithfully matches the input image while preserving the structure and topology of the input mesh. Our key contributions stem from utilizing novel diffusion-based generative models to synthesize 3D pseudo ground-truth that can be used as a soft supervision signal along with additional regularizations, a texture enhancement module that generates high-fidelity texture maps and a body-cloth optimization framework that fits the generated 3D garments to parametric bodies. Our approach clearly outperforms prior work. Finally, we showcased physics-based cloth simulation, hand-garment interaction in a VR environment.

References

[1] Noam Aigerman, Kunal Gupta, Vladimir G. Kim, Siddhartha Chaudhuri, Jun Saito, and Thibault Groueix. Neural jacobian fields: Learning intrinsic mappings of arbitrary meshes. *ACM Trans. Graph.*, 41(4), 2022. 4

[2] Seungbae Bang, Maria Korosteleva, and Sung-Hee Lee. Estimating garment patterns from static scan data. In *Computer Graphics Forum*, pages 273–287. Wiley Online Library, 2021. 3

[3] Ilya Baran, Daniel Vlasic, Eitan Grinspun, and Jovan Popović. Semantic deformation transfer. *ACM Transactions on Graphics (TOG)*, 28(3):1–6, 2009. 3

[4] Amir Barda, Matheus Gadelha, Vladimir G Kim, Noam Aigerman, Amit H Bermano, and Thibault Groueix. Instant3dit: Multiview inpainting for fast editing of 3d objects. *arXiv preprint arXiv:2412.00518*, 2024. 3

[5] Amir Barda, Vladimir Kim, Noam Aigerman, Amit Haim Bermano, and Thibault Groueix. Magicclay: Sculpting meshes with generative neural fields. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–10, 2024. 3

[6] Aric Bartle, Alla Sheffer, Vladimir G Kim, Danny M Kaufman, Nicholas Vining, and Floraine Berthouzoz. Physics-driven pattern adjustment for direct 3D garment editing. *ACM Trans. Graph.*, 35(4):50–1, 2016. 3

[7] Floraine Berthouzoz, Akash Garg, Danny M Kaufman, Eitan Grinspun, and Maneesh Agrawala. Parsing sewing patterns into 3D garments. *Acm Transactions on Graphics (TOG)*, 32(4):1–12, 2013. 2

[8] Remi Brouet, Alla Sheffer, Laurence Boissieux, and Marie-Paule Cani. Design preserving garment transfer. *ACM Transactions on Graphics*, 31(4):Article–No, 2012. 3

[9] Browzwear. v-stitcher. https://browzwear.com/products/v-stitcher, 2024. Accessed on January 2024. 1

[10] Bindita Chaudhuri, Nikolaos Sarafianos, Linda Shapiro, and Tony Tung. Semi-supervised synthesis of high-resolution editable textures for 3D humans. In *CVPR*, 2021. 3

[11] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396*, 2023. 3, 5

[12] Hsiao-yu Chen, Egor Larionov, Ladislav Kavan, Gene Lin, Doug Roble, Olga Sorkine-Hornung, and Tuur Stuyck. Dress anyone: Automatic physically-based garment pattern refitting. *arXiv preprint arXiv:2405.19148*, 2024. 2

[13] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 100–116. Springer, 2019. 3

[14] Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhães, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. Contrastive language and vision learning of general fashion concepts. *Scientific Reports*, 12(1):18958, 2022. 5

[15] CLO3D. Marvelous Designer. https://www.marvelousdesigner.com/, 2024. Accessed on January 2024. 1

[16] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11875–11885, 2021. 3

[17] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 2

[18] Charles de Malefette, Anran Qi, Amal Dev Parakkat, Marie-Paule Cani, and Takeo Igarashi. Perfectdart: Automatic dart design for garment fitting. In *SIGGRAPH Asia 2023 Technical Communications*, pages 1–4, 2023. 2

[19] Dale Decatur, Itai Lang, Kfir Aberman, and Rana Hanocka. 3d paintbrush: Local stylization of 3D shapes with cascaded score distillation. *arXiv preprint arXiv:2311.09571*, 2023. 3

[20] Philippe Decaudin, Dan Julius, Jamie Wither, Laurence Boissieux, Alla Sheffer, and Marie-Paule Cani. Virtual garments: A fully geometric approach for clothing design. In *Computer Graphics Forum*, pages 625–634. Wiley Online Library, 2006. 3

[21] David A Field. Laplacian smoothing and delaunay triangulations. *Communications in applied numerical methods*, 4 (6):709–712, 1988. 4

[22] Anna Frühstück, Nikolaos Sarafianos, Yuanlu Xu, Peter Wonka, and Tony Tung. VIVE3D: Viewpoint-independent video editing using 3D-aware GANs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[23] Daiheng Gao, Xu Chen, Xindi Zhang, Qi Wang, Ke Sun, Bang Zhang, Liefeng Bo, and Qixing Huang. Cloth2tex: A customized cloth texture generation pipeline for 3D virtual try-on. *arXiv preprint arXiv:2308.04288*, 2023. 3

[24] Lin Gao, Jie Yang, Yi-Ling Qiao, Yu-Kun Lai, Paul L Rosin, Weiwei Xu, and Shihong Xia. Automatic unpaired shape deformation transfer. *ACM Transactions on Graphics (ToG)*, 37(6):1–15, 2018. 3

[25] William Gao, Noam Aigerman, Thibault Groueix, Vova Kim, and Rana Hanocka. Textdeformer: Geometry manipulation using text guidance. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3, 6, 7, 8

[26] Will Gao, Dilin Wang, Yuchen Fan, Aljaz Bozic, Tuur Stuyck, Zhengqin Li, Zhao Dong, Rakesh Ranjan, and Nikolaos Sarafianos. 3d mesh editing using masked lrms. *arXiv preprint arXiv:2412.08641*, 2024. 3

[27] Artur Grigorev, Michael J Black, and Otmar Hilliges. Hood: Hierarchical graphs for generalized modelling of clothing dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2023. 3

[28] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. Unsupervised cycle-consistent deformation for shape matching. In *Computer*

Graphics Forum, pages 123–133. Wiley Online Library, 2019. 3

[29] Jingfan Guo, Fabian Prada, Donglai Xiang, Javier Romero, Chenglei Wu, Hyun Soo Park, Takaaki Shiratori, and Shunsuke Saito. Diffusion shape prior for wrinkle-accurate cloth registration. In *3DV*, 2023. 3

[30] Oshri Halimi, Tuur Stuyck, Donglai Xiang, Timur Bagautdinov, He Wen, Ron Kimmel, Takaaki Shiratori, Chenglei Wu, Yaser Sheikh, and Fabian Prada. Pattern-based cloth registration and sparse-view animation. *ACM Transactions on Graphics (TOG)*, 41(6):1–17, 2022. 3

[31] Oshri Halimi, Egor Larionov, Zohar Barzelay, Philipp Herholz, and Tuur Stuyck. Physgraph: Physics-based integration using graph neural networks. *arXiv preprint arXiv:2301.11841*, 2023. 3

[32] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3D scenes with instructions. *arXiv preprint arXiv:2303.12789*, 2023. 2

[33] Kai He, Kaixin Yao, Qixuan Zhang, Jingyi Yu, Lingjie Liu, and Lan Xu. Dresscode: Autoregressively sewing and generating garments from text guidance. *ACM Transactions on Graphics (TOG)*, 43(4):1–13, 2024. 3, 6

[34] Zixuan Huang, Stefan Stojanov, Anh Thai, Varun Jampani, and James M Rehg. Zeroshape: Regression-based zero-shot shape reconstruction. *arXiv preprint arXiv:2312.14198*, 2023. 6, 7

[35] Alec Jacobson, Ilya Baran, Jovan Popovic, and Olga Sorkine. Bounded biharmonic weights for real-time deformation. *ACM Trans. Graph.*, 30(4):78, 2011. 3

[36] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. 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*, pages 18–35. Springer, 2020. 3

[37] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. 3

[38] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2

[39] Maria Korosteleva and Sung-Hee Lee. Neuraltailor: Reconstructing sewing pattern structures from 3D point clouds of garments. *ACM Trans. Graph.*, 41(4), 2022. 2

[40] Dmytro Kotovenko, Olga Grebenkova, Nikolaos Sarafianos, Avinash Paliwal, Pingchuan Ma, Omid Poursaeed, Sreyas Mohan, Yuchen Fan, Yilei Li, Rakesh Ranjan, et al. Wast-3d: Wasserstein-2 distance for scene-to-scene stylization on 3d gaussians. In *ECCV*, 2024. 3

[41] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 4

[42] Boqian Li, Xuan Li, Ying Jiang, Tianyi Xie, Feng Gao, Huamin Wang, Yin Yang, and Chenfanfu Jiang. Garmentdreamer: 3dgs guided garment synthesis with diverse geometry and texture details. *arXiv preprint arXiv:2405.12420*, 2024. 6, 8

[43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 5

[44] Ren Li, Benoît Guillard, and Pascal Fua. Isp: Multi-layered garment draping with implicit sewing patterns. In *NeurIPS*, 2023. 3

[45] Ren Li, Corentin Dumery, Benoit Guillard, and Pascal Fua. Garment Recovery with Shape and Deformation Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[46] Yifei Li, Hsiao-yu Chen, Egor Larionov, Nikolaos Sarafianos, Wojciech Matusik, and Tuur Stuyck. DiffAvatar: Simulation-ready garment optimization with differentiable simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 6

[47] Siyou Lin, Boyao Zhou, Zerong Zheng, Hongwen Zhang, and Yebin Liu. Leveraging intrinsic properties for non-rigid garment alignment. In *ICCV*, 2023. 3

[48] Hsueh-Ti Derek Liu, Michael Tao, and Alec Jacobson. Paparazzi: surface editing by way of multi-view image processing. *ACM Trans. Graph.*, 37(6):221–1, 2018. 3

[49] Lijuan Liu, Xiangyu Xu, Zhijie Lin, Jiabin Liang, and Shuicheng Yan. Towards garment sewing pattern reconstruction from a single image. *ACM Transactions on Graphics (TOG)*, 42(6):1–15, 2023. 2, 3, 6

[50] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3D mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 2, 3

[51] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3

[52] Yufei Liu, Junshu Tang, Chu Zheng, Shijie Zhang, Jinkun Hao, Junwei Zhu, and Dongjin Huang. Clothedreamer: Text-guided garment generation with 3D gaussians. *arXiv preprint arXiv:2406.16815*, 2024. 3, 6

[53] Zhen Liu, Yao Feng, Yuliang Xiu, Weiyang Liu, Liam Paull, Michael J. Black, and Bernhard Schölkopf. Ghost on the shell: An expressive representation of general 3D shapes. *arXiv preprint arXiv:2310.15168*, 2023. 3

[54] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3D using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023. 3, 6, 7

[55] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3D surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, 21(4):163–169, 1987. 3

[56] Arman Maesumi, Paul Guerrero, Noam Aigerman, Vladimir Kim, Matthew Fisher, Siddhartha Chaudhuri, and

Daniel Ritchie. Explorable mesh deformation subspaces from unstructured 3D generative models. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 3

[57] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 5

[58] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022. 2, 3, 8

[59] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3

[60] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*, pages 1–8, 2022. 2, 3

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

[62] Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021. 3

[63] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2D diffusion. *arXiv*, 2022. 3

[64] Anran Qi, Sauradip Nag, Xiatian Zhu, and Ariel Shamir. Personaltailor: Personalizing 2D pattern design from 3D garment point clouds. *arXiv preprint arXiv:2303.09695*, 2023. 3

[65] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 3

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

[67] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3D shapes. *arXiv preprint arXiv:2302.01721*, 2023. 3, 5

[68] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[69] Kenneth Rose, Alla Sheffer, Jamie Wither, Marie-Paule Cani, and Boris Thibert. Developable surfaces from arbitrary sketched boundaries. In *SGP'07-5th Eurographics Symposium on Geometry Processing*, pages 163–172. Eurographics Association, 2007. 3

[70] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Snug: Self-supervised neural dynamic garments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8140–8150, 2022. 3

[71] Tianchang Shen, Jacob Munkberg, Jon Hasselgren, Kangxue Yin, Zian Wang, Wenzheng Chen, Zan Gojcic, Sanja Fidler, Nicholas Sharp, and Jun Gao. Flexible isosurface extraction for gradient-based mesh optimization. *ACM Trans. Graph.*, 42(4):37–1, 2023. 3

[72] Yu Shen, Junbang Liang, and Ming C Lin. Gan-based garment generation using sewing pattern images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 225–247. Springer, 2020. 3

[73] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model, 2023. 3, 6, 7

[74] Olga Sorkine, Daniel Cohen-Or, Yaron Lipman, Marc Alexa, Christian Rössl, and H-P Seidel. Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 175–184, 2004. 3

[75] Astitva Srivastava, Pranav Manu, Amit Raj, Varun Jampani, and Avinash Sharma. Wordrobe: Text-guided generation of textured 3D garments. In *ECCV*, 2024. 6, 7

[76] Oded Stein, Eitan Grinspun, and Keenan Crane. Developability of triangle meshes. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 3

[77] Tuur Stuyck. *Cloth simulation for computer graphics*. Springer Nature, 2022. 3

[78] Style3D. Style3D. https://www.linctex.com/, 2024. Accessed on January 2024. 1

[79] Zhaoqi Su, Liangxiao Hu, Siyou Lin, Hongwen Zhang, Shengping Zhang, Justus Thies, and Yebin Liu. Caphy: Capturing physical properties for animatable human avatars. In *ICCV*, 2023. 3

[80] Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. Deepcloth: Neural garment representation for shape and style editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1581–1593, 2023. 3

[81] Robert W Sumner and Jovan Popović. Deformation transfer for triangle meshes. *ACM Transactions on graphics (TOG)*, 23(3):399–405, 2004. 3

[82] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[83] Angtian Wang, Yuanlu Xu, Nikolaos Sarafianos, Robert Maier, Edmond Boyer, Alan Yuille, and Tony Tung. HISR: Hybrid implicit surface representation for photorealistic 3D human reconstruction. In *AAAI*, 2024. 3

[84] Jionghao Wang, Yuan Liu, Zhiyang Dou, Zhengming Yu, Yongqing Liang, Xin Li, Wenping Wang, Rong Xie, and Li Song. Disentangled clothed avatar generation from text descriptions. *arXiv preprint arXiv:2312.05295*, 2023. 3

[85] Tuanfeng Y. Wang, Duygu Ceylan, Jovan Popović, and Niloy J. Mitra. Learning a shared shape space for multimodal garment design. *ACM Trans. Graph.*, 37(6), 2018. 3

[86] Yu Wang, Alec Jacobson, Jernej Barbič, and Ladislav Kavan. Linear subspace design for real-time shape deformation. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015. 3

[87] Haohan Weng, Tianyu Yang, Jianan Wang, Yu Li, Tong Zhang, CL Chen, and Lei Zhang. Consistent123: Improve consistency for one image to 3D object synthesis. *arXiv preprint arXiv:2310.08092*, 2023. 3

[88] Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, et al. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *ACM Transactions on Graphics (TOG)*, 41 (6):1–15, 2022. 3

[89] Xiaoyu Xiang, Liat Sless Gorelik, Yuchen Fan, Omri Armstrong, Forrest Iandola, Yilei Li, Ita Lifshitz, and Rakesh Ranjan. Make-A-Texture: Fast shape-aware 3D texture generation in 3s. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. 5

[90] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *ICLR*, 2024. 2

[91] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3D mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 3

[92] Shan Yang, Zherong Pan, Tanya Amert, Ke Wang, Licheng Yu, Tamara Berg, and Ming C. Lin. Physics-inspired garment recovery from a single-view image. *ACM Trans. Graph.*, 37(5), 2018. 3

[93] Yu-Ying Yeh, Jia-Bin Huang, Changil Kim, Lei Xiao, Thu Nguyen-Phuoc, Numair Khan, Cheng Zhang, Manmohan Chandraker, Carl S Marshall, Zhao Dong, et al. Texturedreamer: Image-guided texture synthesis through geometry-aware diffusion. *arXiv preprint arXiv:2401.09416*, 2024. 3

[94] Wang Yifan, Noam Aigerman, Vladimir G Kim, Siddhartha Chaudhuri, and Olga Sorkine-Hornung. Neural cages for detail-preserving 3D deformations. In *CVPR*, 2020. 3

[95] Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, and Yebin Liu. Simulcap: Single-view human performance capture with cloth simulation. In *CVPR*, 2019. 3

[96] Xianfang Zeng. Paint3d: Paint anything 3D with lighting-less texture diffusion models. *arXiv preprint arXiv:2312.13913*, 2023. 3

[97] Hao Zhang, Alla Sheffer, Daniel Cohen-Or, Quan Zhou, Oliver Van Kaick, and Andrea Tagliasacchi. Deformation-driven shape correspondence. In *Computer Graphics Forum*, pages 1431–1439. Wiley Online Library, 2008. 3

[98] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 8

[99] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[100] Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, and Jingyi Yu. Human performance modeling and rendering via neural animated mesh. *ACM Trans. Graph.*, 41(6), 2022. 6

[101] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3D reconstruction with transformers. *arXiv preprint arXiv:2312.09147*, 2023. 6, 7