GarmentCrafter: Progressive Novel View Synthesis for Single-View 3D Garment Reconstruction and Editing

Yuanhao Wang[1] Cheng Zhang[2] Gonçalo Frazão[1] Jinlong Yang[3]
Alexandru-Eugen Ichim[3] Thabo Beeler[3] Fernando De la Torre[1]

[1] Carnegie Mellon University [2] Texas A&M University [3] Google AR

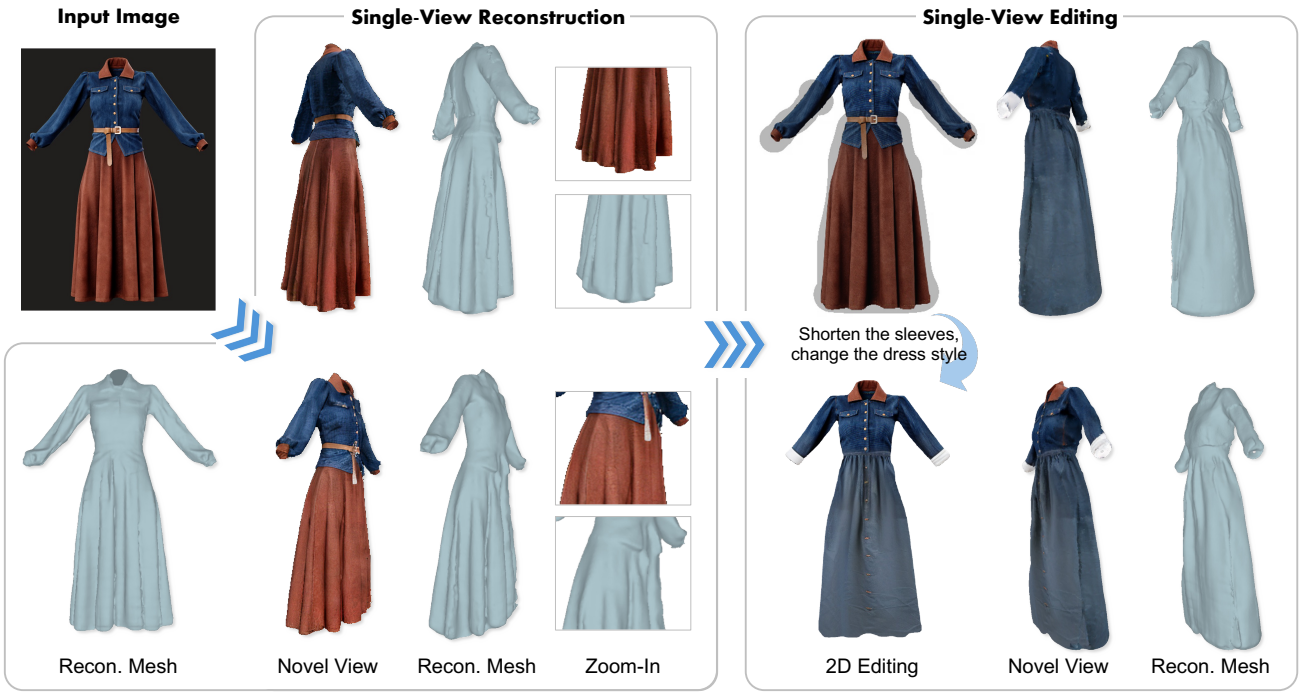

Figure 1. From a real-world clothing image, GarmentCrafter synthesizes high-quality novel views, enabling the reconstruction of garment meshes with accurate geometry and rich detail. Additionally, users can easily apply 2D edits (e.g., modifying parts or surface details) using off-the-shelf tools on a single image, and GarmentCrafter seamlessly applies these edits across the 3D model with multi-view consistency.

Abstract

We introduce GarmentCrafter, a new approach to enable non-professional users to create and modify 3D garments from a single-view image. While recent advances in image generation have facilitated 2D garment design, creating and editing 3D garments remains challenging for non-professional users. Existing methods for single-view 3D reconstruction often rely on pre-trained generative models to hallucinate novel views conditioning on the reference image and camera pose, yet they lack cross-view consistency, failing to capture the internal relationships across different views. In this paper, we tackle this challenge through progressive depth prediction and image warping to approximate novel views. Subsequently, we train a multi-view diffusion model to complete occluded and unknown clothing regions, informed by the evolving camera pose. By jointly inferring RGB and depth, GarmentCrafter enforces inter-view coherence and reconstructs precise geometries and fine details. Extensive experiments demonstrate that our method achieves superior visual fidelity and inter-view coherence compared to state-of-the-art single-view 3D garment reconstruction methods. Our code will be publicly available.

1. Introduction

Professional fashion designers use sophisticated software to create and edit garments in 3D, crafting highly detailed virtual apparels [6, 13, 59, 62]. However, as digital garments become integral to virtual environments and personalized digital experiences [8, 19, 25, 52, 73], there is a growing demand for intuitive tools that allow non-professional users to design and interact with 3D garments. For broader accessibility, such tools should allow users to work with 3D garments with minimal input, ideally from just a single image. This raises a key question: *How can we create and edit 3D garments with simple manipulations in an image?*

Recent advancements in image generation models [49, 51, 53, 66] and image editing techniques [5, 46, 48, 67, 84, 87] have enabled high-quality garment design in 2D. Yet, achieving the same level of control and realism for 3D garments remains challenging for common users. Currently, state-of-the-art methods on single-view 3D garments rely either on 1) deforming, matching, and registration with the human body prior [41] and/or predefined garment templates [3, 14, 18, 35, 37, 43, 55], or 2) novel view synthesis techniques [39, 70] that use pre-trained 2D diffusion models conditioned on a reference image and target pose. However, they often fall short in capturing accurate, realistic geometry and appearance.

Two characteristics of garments pose challenges. First, garments exhibit diverse shapes, complex geometries, and rich textures, making template-based methods limited in their ability to generalize across clothing styles. Most existing methods prioritize either geometry [14, 42] or texture [50, 79], rarely balancing both [18, 43, 55]. Second, the fine details in garments demand stronger multi-view consistency. Existing novel view synthesis methods [40, 74], conditioned on a reference image and target pose, often neglect critical semantic connections across different views.

How can we ensure that a pixel in one view corresponds to a point visible in another, with consistent appearance? In this paper, we propose a different approach, *progressive novel view synthesis*, to enhance cross-view coherence. Our method begins by estimating the depth of the input image and warping projected points to approximate unseen views. We then apply a multi-view diffusion model to complete missing and occluded regions based on the evolving camera pose. Furthermore, we incorporate a monocular depth estimation model to generate depth maps that remain consistent with the warped depths. Unlike existing novel view synthesis, our key insight is to use the depth-based warped image as an additional condition to guide cross-view alignment. By progressively synthesizing views and depths along a predefined camera trajectory, our method gradually refines the geometry and texture of the garment across viewpoints.

We name our method *GarmentCrafter*, a novel solution for 3D garment creation and editing while users just need to operate on a single-view image, as shown in Figure 1. Specifically, GarmentCrafter not only generates high-quality 3D garments but also extends garment editing from 2D to 3D. Thanks to our progressive novel view synthesis, users can make local edits (e.g., editing surface details) or perform part-based manipulations (e.g., modifying garment parts) directly on a single-view image, with precise effects reflected in 3D space — capabilities that are absent in the existing methods [55]. Trained on large-scale 3D garment datasets [4, 16, 88], GarmentCrafter demonstrates superior performance on held-out 3D garment data as well as in-the-wild clothing images. Extensive experiments show that our method outperforms state-of-the-art 2D-to-3D garment reconstruction approaches in terms of geometric accuracy, visual fidelity, and cross-view consistency.

Remark. Professional digital fashion designers typically construct 3D garments from reference images through a progressive workflow that involves initial reconstruction followed by iterative refinement. Our method is designed to support this established practice by offering two modules: a reconstruction module that generates a base 3D garment from a single reference image, and an editing module that allows for detailed 3D adjustments. We explicitly adopt garments in a canonical T-pose or rest pose, which is consistent with industry conventions and facilitates both geometric manipulation and downstream processing. While this constraint may appear limiting, it aligns with common authoring pipelines and enables a more controllable and reproducible design process for non-expert users.

2. Related Work

Single-View 3D Garment Reconstruction and Editing. Reconstructing 3D garments from a single image has been widely explored, with existing methods approaching the task from several perspectives. One line of work relies on parametric body templates, such as SMPL [3, 14, 27, 45], or employs 2D shape priors and keypoint-based techniques [83] to optimize garment structure. Another category of work uses explicit or implicit 3D parametric garment models [3, 15, 18, 35, 42, 43, 55, 86] to capture garment shape and support pose-guided deformations. Additionally, some methods incorporate garment sewing patterns [2, 9, 11, 26, 37, 76, 88], offering flexibility by reconstructing garments from 2D panels. However, these works often struggle to capture diverse garment styles and fine surface details (e.g., wrinkles), and lack support for intuitive garment manipulation, such as modifying surface details or garment parts. In contrast, GarmentCrafter prioritizes novel view synthesis for detailed geometry and texture reconstruction, without relying on garment templates or human body priors, allowing it to handle a wide range of garment styles. Furthermore, single-view edits can also be

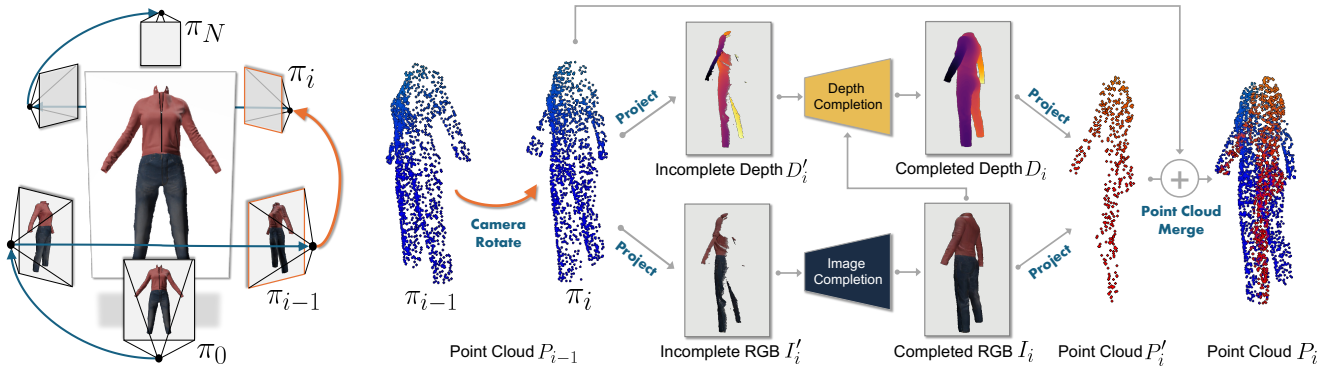

Figure 2. **An illustration of progressive novel view synthesis in GarmentCrafter. Left:** Given a garment image, our method performs depth-aware novel view synthesis along a predefined zigzag camera trajectory. **Right:** For each camera rotation from π_{i-1} to π_i, we project the current point cloud P_{i-1} into the image space based on camera pose π_i, resulting in incomplete RGB and depth images. Our diffusion model completes the RGB image using the warped view, input image, and camera pose as conditions, while a depth completion network refines the depth map based on the completed RGB, warped depth, and camera pose. The re-projected point cloud P_i' is then merged with P_{i-1} to produce an updated point cloud P_i. This iterative process continues until a full 3D representation of the garment is achieved.

seamlessly extended to the 3D model. Note that, our focus in this paper is on garments in a rest pose — well suited to the fashion industry, where ease of adjustment is essential.

Novel View Synthesis from Sparse Images. Our method is inspired by novel view synthesis. Popular approaches such as Neural Radiance Fields (NeRFs) [44] and 3D Gaussian Splatting (3D-GS) [30] rely on numerous posed inputs, limiting their use in single-view scenarios. Recently, distillation from pre-trained 2D generative models has emerged as a promising solution for hallucinating novel views from limited input, with applications in human digitization [1, 20, 21, 32, 54, 71, 72, 82] and object-centric reconstruction [24, 24, 38–40, 47, 57, 60, 70, 85]. However, these methods often lack cross-view consistency and high-quality details, crucial for garment-focused tasks. Unlike models that sample views independently, our method takes semantic cues (i.e., wrapped images) from other views as an additional condition for view synthesis. This might be reminiscent of scene-level approaches, such as Perpetual View Synthesis [7, 12, 28, 36, 63, 78], which condition on warped images for neighbor view image completion. However, we note that scene-centric methods often lack the precision needed for object-centric cases (e.g., garment manipulation) and overlook loop closure for garment shape completion. Our work represents a novel attempt of progressive view synthesis with a predefined camera trajectory for garment reconstruction and editing.

Image-to-3D Reconstruction. Our approach builds on recent advancements in image-to-3D reconstruction, where most methods distill pre-trained generative models via per-scene optimization [10, 33, 47, 58, 65] or multi-view diffusion techniques [24, 38–40, 56, 64, 85]. With the availability of large-scale 3D datasets [16, 17], generalizable Large Reconstruction Models (LRMs) [22, 34, 61, 74, 75] are being trained for feed-forward image-to-3D generation. Unlike Zero-1-to-3 and its variants [39], our method leverages diffusion models to progressively condition on projected images with carefully designed camera trajectory and error reduction methods to enhance cross-view consistency. Additionally, we curated a 3D garment dataset, incorporating assets from existing 3D collections [4, 16, 88], allowing our model to synthesize highly detailed, multi-view images and corresponding depth maps. This process yields multi-view image and depth maps, enabling high-quality mesh reconstruction through standard point cloud-to-mesh methods [29]. While we demonstrate point aggregation and mesh reconstruction in our work, our primary focus is on advancing the multi-view and depth synthesis stages rather than optimizing the point-to-mesh conversion process itself.

3. Approach

We first present problem statement in Section 3.1, followed by our proposed progressive novel view synthesis in Section 3.2. We introduce garment-centric applications enabled by our method in Section 3.3. We describe the details of data curation and model training methods in Section 3.4.

3.1. Problem Definition

Given a single-view garment image I_0, our goal is to generate consistent novel views with detailed RGB textures and accurate depths, which support both single-view 3D reconstruction and editing. Specifically, we first estimate a depth map D_0 based on the input I_0. Then, we project every pixel in the foreground of the garment to the world space, creating a colored point cloud P_0. Our goal is to complete this point cloud by sequentially incorporating information from synthesized novel views. To achieve this, we propose an progressive 3D completion process with a predefined cam-

era trajectory $\boldsymbol{\pi} = \{\pi_1, \pi_2, ..., \pi_N\}$ that forms a closed loop around the garment object. Figure 2 illustrates the overall framework. Next, we elaborate the details of an arbitrary step in the following sections.

3.2. Progressive Novel View Synthesis

Overview. At the step i of the progressive novel view synthesis (see Figure 2), we first project the existing point cloud P_{i-1} to the image plane of camera $\pi_i \in \boldsymbol{\pi}$, producing an incomplete image I'_i and an incomplete depth map D'_i. We then apply an image completion model to inpaint the missing areas in I'_i, resulting in I_i. Next, we use an monocular depth estimation model to estimate the corresponding depth map D_i consistent with the known depths in D'_i. Finally, we integrate I_i and D_i with the existing point cloud to obtain a merged P_i. By following a predefined camera trajectory, our method can generate view-dependent images and corresponding depths that enable high-quality garment reconstruction and edit with improved cross-view consistency.

3.2.1 Conditional Image Generation.

At step i, the goal is to synthesize $I_i \in \mathbb{R}^{H\times W\times 3}$, the image of the garment object from the viewpoint of camera π_i, given the input image I_0, the projected image I'_i, and the relative camera rotation $R_i \in \mathbb{R}^{3\times 3}$ and translation $T_i \in \mathbb{R}^3$ from π_0 to π_i. We aim to train a model f_{img} such that:

$$I_i = f_{\text{img}}(I_0, I'_i, R_i, T_i), \tag{1}$$

where I_i is the synthesized complete image that retains the appearance of I'_i in the known regions, and synthesizes plausible appearance in the unknown regions that remain perceptually consistent with I'_i and the original input I_0.

To learn f_{img}, we fine-tune a denoising diffusion model, leveraging its strong generalization capabilities in image generation. Specifically, we adopt a latent diffusion architecture based on Stable Diffusion [51] with an image encoder $\mathcal{E}$, a denoising network ϵ_θ, and a decoder $\mathcal{D}$. At denoising step $s \in S$, let z_s denote the noisy latent of the target image $x = I_i$, and let $\boldsymbol{c} = c(I_0, I'_i, R_i, T_i)$ be the embedding of the anchor view image, target view projected image, and relative camera extrinsics. We optimize the following latent diffusion objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathcal{E}(I_0),\mathcal{E}(I'_i),\epsilon\sim\mathcal{N}(0,\mathbf{I}),s}\left[\|\epsilon - \epsilon_\theta(z_s, s, \boldsymbol{c})\|^2\right]. \tag{2}$$

Unlike existing multi-view diffusion models (e.g., [39, 56]), which synthesize novels views from an arbitrary input viewpoint, we unify our garment-centric task by fixing the input image to a near-frontal view of the garment. This allows R_i and T_i to be interpreted as the absolute camera transformation from the frontal view. Furthermore, in addition to conditioning on the anchor view image, we incorporate the warped image (i.e., I'_i in Figure 2 and Equation 1) at the target view as an additional condition input, which provides a strong prior that enhances cross-view consistency in garment reconstruction, as demonstrated in Section 4.4.

Conditional Depth Generation. Given complete RGB image I_i, we learn a depth model f_{depth} to estimate the depth map $D_i \in \mathbb{R}^{H\times W\times 1}$ conditioned on the warped incomplete depth map D'_i:

$$D_i = f_{\text{depth}}(I_i, D'_i) \tag{3}$$

Similar to the conditional image generation, we enforce depth preservation in known regions by framing the task as metric depth estimation. To ensure consistency, we align the depth values of D_i and D'_i during training. The model is optimized using an $\mathcal{L}_1$ loss:

$$\mathcal{L}_1 = ||(D_i - \hat{D}_i) \cdot m||, \tag{4}$$

where $\hat{D}_i$ is the ground-truth depth, and m is the foreground mask. To train f_{depth}, we fine-tune the pretrained human foundation model, Sapiens [31], leveraging its strong priors for human-related tasks. To condition the model on D'_i, we concatenate D'_i with I_i as input and add an extra channel to the first projection layer of Sapiens model. The weights of the added channel are initialized to zero.

Point Cloud Merging and Projection. To integrate novel view observations (i.e., I_i and D_i) into the existing point cloud P_{i-1}, we first identify the inpainted regions from the image model. Pixels in these regions are projected into world space and merged with P_{i-1} to form P_i, with expanded borders to include overlapping regions. To minimize stitching artifacts, we align the depth map of the inpainted regions with the warped depth map of P_{i-1}. When projecting a partial point cloud to a novel view, only surfaces facing the camera should be rendered. To enforce this, we track the orientation of each point. For a point x added at step i, its orientation vector v is derived from the normal direction of the corresponding pixel in D_i. During projection, a point is ignored if dot $(v, v_0) < 0$, where v_0 is the viewing direction.

We illustrate the whole process of P-NVS and show the intermediate results in the supplementary. After completing all steps along the camera trajectory, we optionally sample a few random views for additional inpainting to cover any region occluded in previous views.

3.3. Garment Digitization and Editing

Garment Digitization. Our method enables garment digitization from a single image by progressively synthesizing novel views, generating multi-view consistent RGBD images and a colored point cloud. This output serves as an intermediate representation that can be converted to other 3D representations. In this work, we employ Screened Poisson surface reconstruction [29] to convert the point cloud into

a textured mesh. Note that each point of the point cloud contains both the RGB color information and the surface normals. The Screened Poisson method interpolates these attributes to map textures onto the mesh vertices and produces a watertight mesh. To preserve the non-watertight garment topology, we apply a trimming operation to remove unwanted mesh surfaces introduced by Poisson Reconstruction. Please see supplementary for a comparison with and without surface trimming. To further reduce artifacts, we remove floating faces unconnected to the main mesh and apply Laplacian smoothing to refine the mesh surface.

Interactive Editing. Redesigning a 3D garment model typically requires significant expertise, making it impractical for most users. GarmentCrafter provides an intuitive alternative, allowing users to edit a rendered image of the garment from a selected view, which is then lifted into 3D. In this work, we focus on two types of edits: (1) *Part-based Editing*: Modifies the geometry or texture of specific garment parts, such as sleeves or pant legs. Users can add, remove, or resize components. (2) *Local surface editing*: Adjusts the geometry and texture of localized regions, such as adding a pocket or modifying the neckline design.

The garment part editing is achieved with the following strategy. Given a 3D garment object G, the user selects an anchor view π and edits the rendered image I to obtain I_{edit}. This editing step can be done using any image editing tool, such as Photoshop or AI-based methods. We first identify the edited region in I_{edit} and remove the corresponding garment parts from G, leaving a partial 3D garment G'. We reformulate the editing task as single-view 3D garment part reconstruction, conditioned on G'. We follow the process described in Section 3.2 with two modifications: (1) At each step along the camera trajectory, the conditional image and depth are generated by combining the projected point cloud with observations from the partial garment G'. (2) After computing image and depth maps, only pixels within the edited region are projected and merged with the existing point cloud. The final output is a colored point cloud of the edited parts, which is then merged with G'. For local surface editing, instead of removing and reconstructing an entire garment part, we apply the same process to a localized surface region.

3.4. Data Preparation and Training

We construct the training dataset by replicating the inference procedure. For each 3D garment, we sample 6 uniform views at 20° elevation (following the full camera trajectory) and 4 additional random views between 60° and -30° for inpainting.

Training Data for Reconstruction. We follow the zigzag camera trajectory (Figure 2) and at each step i, we form a training pair for the image generation model f_{img}: $\{(I'_i, I_0, R_i, T_i), I_i\}$, where I'_i is the projected image, I_0 is the anchor view, and (R_i, T_i) are the relative camera transformations. Similarly, the depth generation model f_{depth} is trained with $\{(D'_i, I_i), D_i\}$, where D'_i is the projected depth, and D_i is the ground-truth depth. We merge the point cloud with I_i and D_i before proceeding. Finally, we repeat the process for four random views to simulate inpainting.

Training Data for Editing. We generate training data for 3D editing by randomly removing parts of a 3D garment. At each step, we create a partial image I''_i and depth map D''_i by merging I'_i and D'_i with known observations. The training pairs become $\{(I''_i, I_0, R_i, T_i), I_i\}$ for f_{img} and $\{(D''_i, I_i), D_i\}$ for f_{depth}.

Joint Training. To learn a unified model for both reconstruction and editing, we combine their training data. We randomly apply small rotations to the 3D object when generating the training data, enabling the model to handle in-the-wild inputs that may not be well-posed. Please refer to the supplementary materials for details.

4. Experiments

We present experimental results of our method on single-view garment reconstruction and editing. Please see supplementary for additional details, analyses, and results.

4.1. Datasets, Metrics, and Baselines

Datasets. We validate GarmentCrafter using 3D garment assets from a number of sources. (1) Curated dataset: We collect $\sim$700 3D garments with diverse shape and texture from online sources. (2) Objaverse 1.0 (Garment) [16]: the original v1.0 dataset contains more than 800K 3D objects, where most of the existing method trained on [39, 74, 77]. We manually curated a subset only contain $\sim$900 high-quality garment assets. (3) BEDLAM [4]: 114 garments, each has many textures, $\sim$1600 assets in total. (4) Cloth4D [88]: $\sim$1100 artists made garments.

Quantitative Metrics. (1) Texture and appearance quality: we evaluate the novel view synthesis using commonly used LPIPS [80], PSNR [23], SSIM [68]. (2) Geometry quality: we measure the performance using geometric errors with Chamfer distance (bi-directional point-to-mesh) between ground-truth and reconstructed meshes.

Baselines. We compare GarmentCrafter with state-of-the-art models for image-to-3D object and image-to-garment reconstruction. (1) Hunyuan3D-2.0 [81]: a powerful large model for high-quality image-to-3D object reconstruction. (2) InstantMesh [74]: object reconstruction by generating novel views using Zero-1-to-3++ [56]. (3) CRM [69]: generate six orthographic views for 3D object reconstruction. (4) Garment3DGen [55]: a state-of-the-art garment-specific model based on template optimization, with templates initialized by InstantMesh [74]. As the texture code is not released, we compare only mesh geometry.

Figure 3. **Qualitative comparison on single-view 3D garment reconstruction.** Our method demonstrates better performance in handling complex texture patterns and geometric structures compared to Hunyuan3D-2.0 [81], InstantMesh [74], and CRM [69].

For fair comparisons, we fine-tuned InstantMesh* on our garment dataset. Hunyuan3D-2.0 and CRM require significant computing for full fine-tuning, making it impractical given our resource constraints.

4.2. Results on Single-View Reconstruction

We evaluate GarmentCrafter on single-view reconstruction using a held-out test dataset of 150 garment assets. For each test case, we sample 12 views with alternating elevations of 0° and 20° and azimuth angles evenly spaced over 360°. To assess image quality, we convert the generated point clouds to meshes using a classical surface reconstruction method and render multi-view images. For geometry evaluation, we compute the Chamfer distance directly between the generated point cloud and the ground-truth mesh.

Qualitative Results. Figure 3 shows qualitative comparisons, where GarmentCrafter demonstrates superior texture and geometry generation compared to all other baselines.

Table 1. **Quantitative comparison of texture and geometry quality.** Garment3DGen provides no texture reconstruction code.

	Appearance			Geometry
	LPIPS↓	PSNR↑	SSIM↑	Chamfer↓
Hunyuan3D-2.0	<u>0.1743</u>	18.79	<u>0.8158</u>	<u>0.0088</u>
InstantMesh*	0.1848	<u>19.14</u>	0.7944	0.0139
CRM	0.2213	17.51	0.8131	0.0127
Garment3DGen	–	–	–	0.0123
GarmentCrafter	**0.1190**	**22.36**	**0.8317**	**0.0044**

Our method, benefiting from consistent multi-view generation, produces sharp textures, intricate geometric details, and tight alignment with the input image. Notably, even against the highly capable Hunyuan3D-2.0, our method maintains a clear edge across these dimensions. Figure 4 shows additional qualitative results of GarmentCrafter.

Quantitative Results on Geometry Quality. We present

Figure 4. More qualitative results of GarmentCrafter on single-view reconstruction. Please see supplementary for more results.

Table 2. Ablation study on Progressive Novel View Synthesis (P-NVS) and analysis on multi-view consistency. We show results with and without P-NVS. CVCS: Cross-View Consistency Score.

P-NVS	LPIPS ↓	PSNR ↑	SSIM ↑	CVCS↑
✗	0.1195	21.512	0.8369	0.9030
✓	**0.1052**	**22.776**	**0.8557**	**0.9512**

quantitative geometry evaluation results in Table 1. GarmentCrafter outperforms baseline methods in terms of Chamfer distance, highlighting its enhanced ability to capture detailed surface geometries in 3D garments.

Quantitative Results on Texture Quality. We conduct a quantitative analysis of texture quality on our held-out test dataset and show results in Table 1. Across all image metrics, GarmentCrafter surpasses all baseline methods, demonstrating its effectiveness in producing high-fidelity textures and preserving fine-grained details.

4.3. Results on Single-View Editing

Figure 5 shows qualitative results on single-view editing, including various types of edits such as resizing, element swapping, and surface edits. GarmentCrafter applies these edits in 3D while preserving cross-view consistency.

4.4. Analyses and Ablation Studies

Importance of Progressive Novel View Synthesis. A key insight of our method is to progressively synthesize novel view by conditioning the generation on the projected images. We conduct an ablation study on the effect of projected image conditioning. For each test case, we select an anchor view π_1, and a second camera view, π_2, at a 60° azimuthal angle relative to π_1. We compare the performance of our image model with or without projected image conditioning at synthesizing view π_2 in Table 2. We observe a drop in performance measured in image similarity metrics when removing the projected condition.

Analysis on multi-view consistency. Common image metrics (e.g., LPIPS, PSNR, and SSIM) measure similarity but do not directly reflect cross-view consistency. We propose a new metric, the Cross-View Consistency Score (CVCS), to gain deeper insights into the consistency of our results.

$$\text{CVCS} = 1 - \frac{\Sigma |I - I'| \cdot m'}{\Sigma m'} \tag{5}$$

where I is the synthesized image at camera view π, I' is a partial image projected from an observed view π_0 with known depth, and m' is a binary mask indicating the projection regions. This assumes π and π_0 are relatively close.

We use the CVCS metric to ablate the impact of P-NVS. As shown in Table 2, GarmentCrafter achieves superior cross-view consistency with P-NVS. We further validate this with a visual example in the supplementary.

Effect of Trajectory on Loop Closure. For better loop closure, we use a "zigzag" camera trajectory where we rotate the camera to left and right alternatively and converge at the center back of the garment (see Figure 2). This design aims to better capture overlapping views, thereby improving reconstruction accuracy. We validate this design choice by comparing the quality of the 3D meshes generated using zigzag and sequential trajectories. We report quantitative results in Table 3. We find that our chosen trajectory achieves better performance across both image and geometry metrics. We additionally show a qualitative comparison in Figure 6. When using a circular trajectory, achieving loop closure from the side view is challenging; the generated geometry (left sleeve) often conflicts with prior predictions, leading to model failure.

5. Conclusion

We present GarmentCrafter, a new approach to reconstruct and edit 3D garments from a single input image. Our method synthesizes novel view images progressively to ensure cross-view consistency, thereby achieving high quality

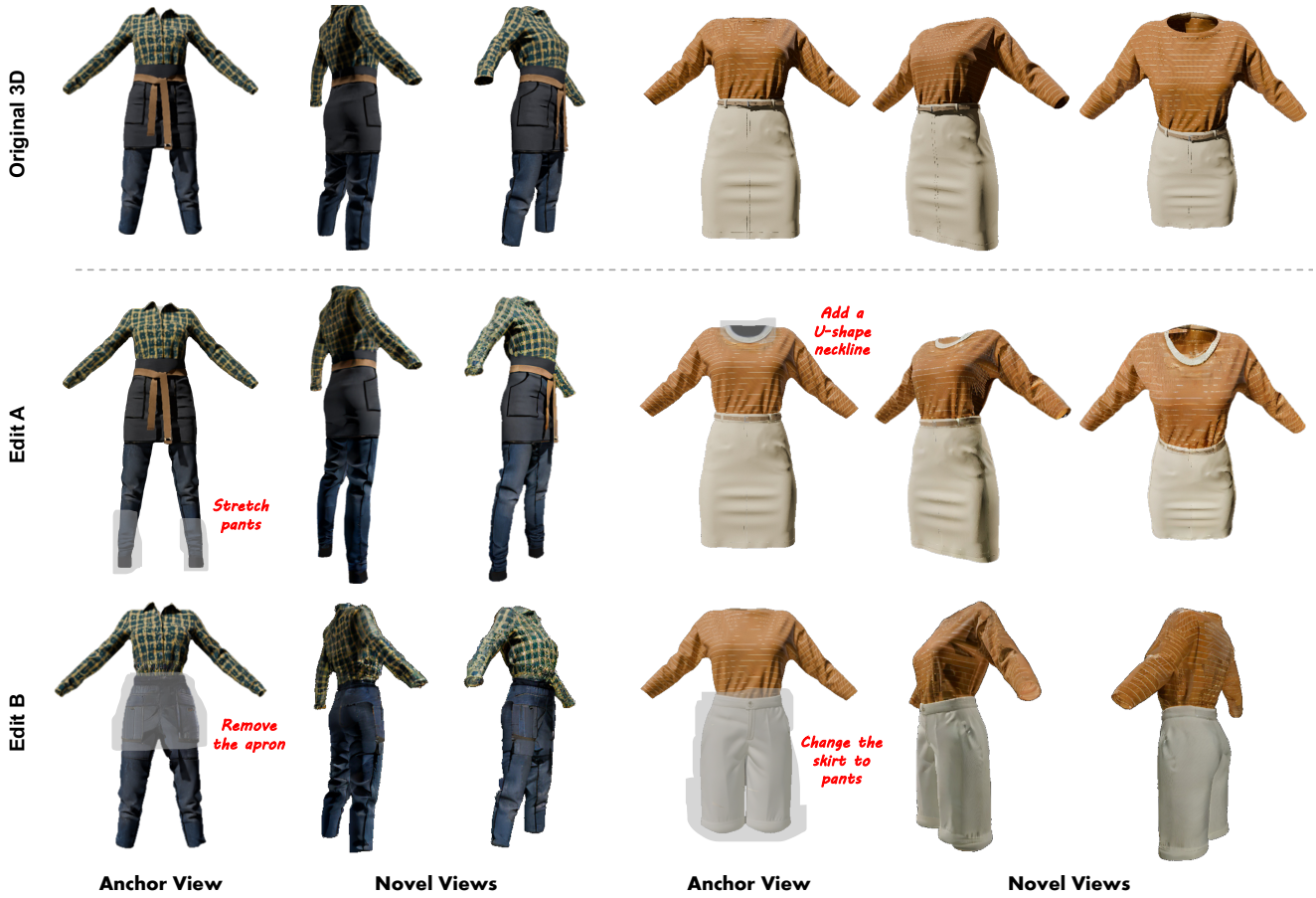

Figure 5. **Qualitative results on single-view 3D garment editing.** GarmentCrafter enables single-view edit such as modify the geometry and surface details of the garment, with the changes accurately reflected across the 3D model. Please see supplementary for more results.

Table 3. **Ablation study on camera trajectory selection.** We study two types camera trajectory for progressive novel view synthesis. **Circular**: the camera moves around the object in regular steps, either clockwise or counterclockwise. **Zigzag**: the camera alternates directions with each step, as shown in Figure 2. Results indicate that our proposed zigzag achieves better appearance and geometry quality compared to using circular trajectory. We show an actual example in Figure 6 for qualitative analyses.

Trajectory	LPIPS ↓	PSNR ↑	SSIM ↑	Chamfer ↓
Circular	0.1503	20.79	0.8130	0.0054
Zigzag (Ours)	**0.1454**	**21.22**	**0.8173**	**0.0044**

Figure 6. **Camera trajectory selection for loop closure.** Zigzag achieves better loop closure, while the circular trajectory struggles with side-view closure, leading to geometric conflicts and failure. We argue that there are numerous ways to select camera trajectories, our proposed approach offers an intuitive solution tailored for single-view reconstruction and editing.

geometry and texture results. We have conducted extensive experiments to demonstrate the superior performance of GarmentCrafter with other baseline methods. Please see supplementary materials for additional implementation and training details, more qualitative results on garment reconstruction and editing, as well as an ablation study on the rotation angles in the camera trajectory.

We focus on garments in a rest pose rather than arbitrary poses. This scope is a deliberate choice, as GarmentCrafter is designed as a tool to facilitate digital garment design and editing, and rest poses provide a consistent and intuitive baseline well suited to this purpose. Future work could extend the training dataset to include 3D garments in varied poses, allowing the method to generalize to arbitrary garment inputs. Additionally, our model reconstructs only the external surface of the garments, without accounting for inner layers or internal structures. This will be addressed in future work.

References

[1] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1506–1515, 2022. 3

[2] 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

[3] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5420–5430, 2019. 2

[4] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 2, 3, 5

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2

[6] Browzwear. Browzwear. https://browzwear.com/, 2025. 2

[7] Shengqu Cai, Eric Ryan Chan, Songyou Peng, Mohamad Shahbazi, Anton Obukhov, Luc Van Gool, and Gordon Wetzstein. Diffdreamer: Towards consistent unsupervised single-view scene extrapolation with conditional diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2139–2150, 2023. 3

[8] Andrés Casado-Elvira, Marc Comino Trinidad, and Dan Casas. Pergamo: Personalized 3d garments from monocular video. In *Computer Graphics Forum*, pages 293–304. Wiley Online Library, 2022. 2

[9] Cheng-Hsiu Chen, Jheng-Wei Su, Min-Chun Hu, Chih-Yuan Yao, and Hung-Kuo Chu. Panelformer: Sewing pattern reconstruction from 2d garment images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 454–463, 2024. 2

[10] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22246–22256, 2023. 3

[11] Xipeng Chen, Guangrun Wang, Dizhong Zhu, Xiaodan Liang, Philip Torr, and Liang Lin. Structure-preserving 3d garment modeling with neural sewing machines. *Advances in Neural Information Processing Systems*, 35:15147–15159, 2022. 2

[12] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 3

[13] CLO3D. CLO3D. https://www.clo3d.com/en/, 2025. 2

[14] 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. 2

[15] R Daněřek, Endri Dibra, Cengiz Öztireli, Remo Ziegler, and Markus Gross. Deepgarment: 3d garment shape estimation from a single image. In *Computer Graphics Forum*, pages 269–280. Wiley Online Library, 2017. 2

[16] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 3, 5

[17] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[18] 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. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 2

[19] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12858–12868, 2023. 2

[20] Vishnu Mani Hema, Shubhra Aich, Christian Haene, Jean-Charles Bazin, and Fernando De la Torre. Famous: High-fidelity monocular 3d human digitization using view synthesis. *arXiv preprint arXiv:2410.09690*, 2024. 3

[21] I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 538–549, 2024. 3

[22] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 3

[23] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 5

[24] Hanzhe Hu, Zhizhuo Zhou, Varun Jampani, and Shubham Tulsiani. Mvd-fusion: Single-view 3d via depth-consistent multi-view generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9698–9707, 2024. 3

[25] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 634–644, 2024. 2

[26] Moon-Hwan Jeong, Dong-Hoon Han, and Hyeong-Seok Ko. Garment capture from a photograph. *Computer Animation and Virtual Worlds*, 26(3-4):291–300, 2015. 2

[27] 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. 2

[28] Biliana Kaneva, Josef Sivic, Antonio Torralba, Shai Avidan, and William T Freeman. Infinite images: Creating and exploring a large photorealistic virtual space. *Proceedings of the IEEE*, 98(8):1391–1407, 2010. 3

[29] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006. 3, 4

[30] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4):1–14, 2023. 3

[31] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2024. 4

[32] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 505–515, 2024. 3

[33] Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, et al. Generalizable human gaussians for sparse view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 3

[34] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 3

[35] Ren Li, Corentin Dumery, Benoît Guillard, and Pascal Fua. Garment recovery with shape and deformation priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1586–1595, 2024. 2

[36] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14458–14467, 2021. 3

[37] 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

[38] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[39] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2, 3, 4, 5

[40] 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. 2, 3

[41] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2

[42] Zhongjin Luo, Haolin Liu, Chenghong Li, Wanghao Du, Zirong Jin, Wanhu Sun, Yinyu Nie, Weikai Chen, and Xiaoguang Han. Garverselod: High-fidelity 3d garment reconstruction from a single in-the-wild image using a dataset with levels of details. *arXiv preprint arXiv:2411.03047*, 2024. 2

[43] Sahib Majithia, Sandeep N Parameswaran, Sadbhavana Babar, Vikram Garg, Astitva Srivastava, and Avinash Sharma. Robust 3d garment digitization from monocular 2d images for 3d virtual try-on systems. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3428–3438, 2022. 2

[44] 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 *European Conference on Computer Vision (ECCV)*, 2020. 3

[45] 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. 2

[46] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[47] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[48] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. Swapnet: Image based garment transfer. In *Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XII 15*, pages 679–695. Springer, 2018. 2

[49] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2

[50] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d

shapes. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[51] 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, 4

[52] Boxiang Rong, Artur Grigorev, Wenbo Wang, Michael J Black, Bernhard Thomaszewski, Christina Tsalicoglou, and Otmar Hilliges. Gaussian garments: Reconstructing simulation-ready clothing with photorealistic appearance from multi-view video. *arXiv preprint arXiv:2409.08189*, 2024. 2

[53] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[54] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 3

[55] Nikolaos Sarafianos, Tuur Stuyck, Xiaoyu Xiang, Yilei Li, Jovan Popovic, and Rakesh Ranjan. Garment3dgen: 3d garment stylization and texture generation. In *3DV*, 2025. 2, 5

[56] 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. *arXiv preprint arXiv:2310.15110*, 2023. 3, 4, 5

[57] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 3

[58] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 3

[59] Style3D. Style3D. https://www.linctex.com/, 2025. 2

[60] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 3

[61] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2025. 3

[62] TUKA3D. TUKA3D. https://tukatech.com/tuka3d/, 2025. 2

[63] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. In *European Conference on Computer Vision*, pages 197–214. Springer, 2025. 3

[64] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2025. 3

[65] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 3

[66] Junyan Wang, Zhenhong Sun, Zhiyu Tan, Xuanbai Chen, Weihua Chen, Hao Li, Cheng Zhang, and Yang Song. Towards effective usage of human-centric priors in diffusion models for text-based human image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8446–8455, 2024. 2

[67] Tongxin Wang and Mang Ye. Texfit: Text-driven fashion image editing with diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10198–10206, 2024. 2

[68] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[69] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *European Conference on Computer Vision*, pages 57–74. Springer, 2025. 5, 6

[70] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 2, 3

[71] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 3

[72] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 512–523, 2023. 3

[73] Yuliang Xiu, Yufei Ye, Zhen Liu, Dimitrios Tzionas, and Michael J Black. Puzzleavatar: Assembling 3d avatars from personal albums. *arXiv preprint arXiv:2405.14869*, 2024. 2

[74] 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. 2, 3, 5, 6

[75] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024. 3

[76] Shan Yang, Tanya Ambert, Zherong Pan, Ke Wang, Licheng Yu, Tamara Berg, and Ming C Lin. Detailed garment recovery from a single-view image. *arXiv preprint arXiv:1608.01250*, 2016. 2

[77] Xianghui Yang, Huiwen Shi, Bowen Zhang, Fan Yang, Jiacheng Wang, Hongxu Zhao, Xinhai Liu, Xinzhou Wang, Qingxiang Lin, Jiaao Yu, et al. Hunyuan3d-1.0: A unified framework for text-to-3d and image-to-3d generation. *arXiv preprint arXiv:2411.02293*, 2024. 5

[78] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6658–6667, 2024. 3

[79] Cheng Zhang, Yuanhao Wang, Francisco Vicente Carrasco, Chenglei Wu, Jinlong Yang, Thabo Beeler, and Fernando De la Torre. FabricDiffusion: High-fidelity texture transfer for 3d garments generation from in-the-wild images. In *ACM SIGGRAPH Asia*, 2024. 2

[80] 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. 5

[81] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025. 5, 6

[82] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 3170–3184, 2021. 3

[83] Bin Zhou, Xiaowu Chen, Qiang Fu, Kan Guo, and Ping Tan. Garment modeling from a single image. In *Computer graphics forum*, pages 85–91. Wiley Online Library, 2013. 2

[84] Shizhe Zhou, Hongbo Fu, Ligang Liu, Daniel Cohen-Or, and Xiaoguang Han. Parametric reshaping of human bodies in images. *ACM transactions on graphics (TOG)*, 29(4):1–10, 2010. 2

[85] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12588–12597, 2023. 3

[86] Heming Zhu, Lingteng Qiu, Yuda Qiu, and Xiaoguang Han. Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3845–3854, 2022. 2

[87] Luyang Zhu, Yingwei Li, Nan Liu, Hao Peng, Dawei Yang, and Ira Kemelmacher-Shlizerman. M&m vto: Multi-garment virtual try-on and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1346–1356, 2024. 2

[88] Xingxing Zou, Xintong Han, and Waikeung Wong. Cloth4d: A dataset for clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12847–12857, 2023. 2, 3, 5