

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

Yuanhao Wang¹ Cheng Zhang² Gonçalo Frazão¹ Jinlong Yang³ Alexandru-Eugen Ichim³
Thabo Beeler³ Fernando De la Torre¹

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

humansensinglab.github.io/garment-crafter



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 that enables 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 synthesize novel views conditioning on the reference image and camera pose, yet they lack cross-view consistency, failing to capture the internal relationships across differ-

ent 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.

1. Introduction

Professional fashion designers use sophisticated software to create and edit garments in 3D, crafting highly detailed virtual apparels [1–4]. However, as digital garments become integral to virtual environments and personalized digital experiences [11, 21, 27, 54, 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 [51, 53, 55, 66] and image editing techniques [9, 48, 50, 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 [43] and/or predefined garment templates [7, 16, 20, 37, 39, 45, 57], or 2) novel view synthesis techniques [41, 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 [16, 44] or texture [52, 80], rarely balancing both [20, 45, 57]. Second, the fine details in garments demand stronger multi-view consistency. Existing novel view synthesis methods [42, 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 pre-defined 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 [57]. Trained on large-scale 3D garment datasets [8, 18, 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.

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 [7, 16, 29, 47], 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 [7, 17, 20, 37, 44, 45, 57, 86] to capture garment shape and support pose-guided deformations. Additionally, some methods incorporate garment sewing patterns [6, 12, 14, 28, 39, 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 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) [46] and 3D Gaussian Splatting (3D-GS) [32] 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 [5, 22, 23, 34, 56, 71, 72, 82] and object-centric reconstruction [26, 26, 40–42, 49, 59, 61, 70, 85]. However, these methods often lack cross-view consistency and

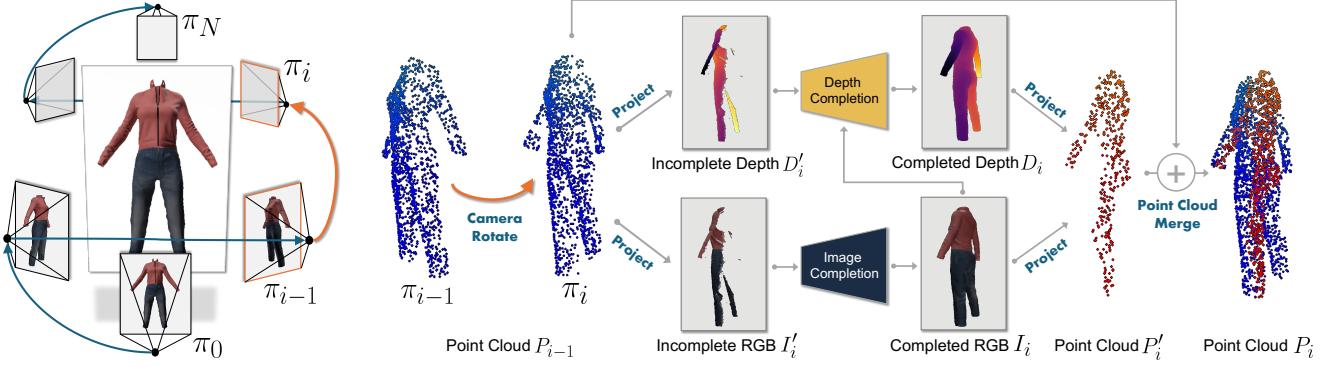


Figure 2. **An illustration of progressive novel view synthesis in GarmentCrafter.** **Left:** Given a garment image, our method performs depth-aware novel view synthesis along a predefined zigzag camera trajectory. **Right:** For each camera rotation from π_{i-1} to π_i , we project the current point 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.

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 [10, 15, 30, 38, 63, 79], 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 [13, 35, 49, 60, 65] or multi-view diffusion techniques [26, 40–42, 58, 64, 85]. With the availability of large-scale 3D datasets [18, 19], Large Reconstruction Models (LRMs) [24, 36, 62, 74, 75] are being trained for feed-forward image-to-3D generation. Unlike Zero-1-to-3 and its variants [41], our method leverages diffusion models to progressively condition on warped 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 [8, 18, 88], allowing our model to synthesize highly detailed, multi-view images and corresponding depth maps. This process yields multi-view images alongside accurate depth maps, enabling high-quality mesh reconstruction through standard point cloud-to-mesh methods [31]. 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 camera trajectory $\pi = \{\pi_1, \pi_2, \dots, \pi_N\}$ 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 $\pi_i \in \pi$, 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 a 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.

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:

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

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 [53] 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 $c = c(I_0, I'_i, R_i, T_i)$ be the embedding of the anchor view image, target view projected image, and relative camera extrinsics. We optimize the following latent diffusion objective:

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

Unlike existing multi-view diffusion models (e.g., [41, 58]), which synthesize novel 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. After obtained 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 as follows:

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

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:

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

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 [33], 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 $\text{dot}(v, v_0) < 0$, where v_0 is the viewing direction. After completing all steps along the camera trajectory, we optionally sample a few random views for additional inpainting to recover any occluded regions. Please see supplementary for additional details.

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 for various 3D reconstruction. In this work, we employ Screened Poisson surface reconstruction [31] to convert the point cloud into a textured mesh. Specifically, we project multi-view RGBD images to form a colored point cloud, where each point encodes geometry and color. The Screened Poisson method then interpolates these attributes, mapping textures onto mesh vertices.

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} .

We first identify the edited region in I_{edit} and remove the corresponding garment parts from G , leaving a partial garment G' that remains unchanged. This reformulates the task as single-view 3D garment part reconstruction, conditioned on G' . We then 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 simulating inference. For each 3D garment, we sample 6 uniform views at 0° 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. For 3D editing, we generate training data by randomly removing parts of a 3D garment to create a partial known model. 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 ~ 700 3D garments with diverse shape and texture from Artstation¹. (2) Objaverse 1.0 (Garment) [18]: the original v1.0 dataset contains more than 800K 3D objects, where most of the existing method trained on [41, 74, 77]. We manually curated a subset only contain ~ 900 high-quality garment assets. (3) BEDLAM [8]: 114 garments, each has many textures, ~ 1600 garments in total. (4) Cloth4D [88]: ~ 1100 artists made garments.

Quantitative Metrics. (1) Texture and appearance quality: we evaluate the novel view synthesis using commonly used LPIPS [81], PSNR [25], 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) InstantMesh [74]: object reconstruction by generating novel views using Zero-1-to-3++ [58]. (2) CRM [69]: generate six orthographic views for 3D object reconstruction. (3) Hunyuan3D-1.0 [77]: a newly released model for high-quality image-to-3D object reconstruction. (4) Garment3DGen [57]: 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.

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. Our method, benefiting from consistent multi-view generation, produces sharp textures and intricate geometric details, whereas other baselines often result in blurry textures and overly smoothed geometries. Figure 4 shows additional qualitative results of GarmentCrafter.

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 quality metrics, GarmentCrafter consistently surpasses baseline methods, demonstrating its effectiveness in producing high-fidelity textures and preserving fine-grained details.

¹<https://www.artstation.com/>



Figure 3. **Qualitative comparison on single-view 3D garment reconstruction with state-of-the-art methods.** Our method demonstrates better performance in handling complex texture patterns and geometric structures compared to InstantMesh [74], Hunyuan3D-1.0 [78], and Convolutional Reconstruction Model (CRM) [69].



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

Table 1. **Quantitative comparison of texture and geometry quality.** InstantMesh*: with fine-tuned Zero-1-to-3++ on our garment data for a fair comparison. CRM and Hunyuan3D-1.0 require significant computing for full fine-tuning, making it impractical. Garment3DGen does not provide texture reconstruction code.

	Appearance			Geometry
	LPIPS↓	PSNR↑	SSIM↑	Chamfer↓
InstantMesh* [74]	0.1848	19.14	0.7944	0.0139
CRM [69]	0.2213	17.51	0.8131	0.0127
Hunyuan3D-1.0 [77]	0.2216	17.77	0.7794	0.0121
Garment3DGen [57]	–	–	–	0.0123
GarmentCrafter	0.1190	22.36	0.8317	0.0044

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 Results on Geometry Quality. We present 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 garment shapes.

4.3. Results on Single-View Editing

We present qualitative results on single-view editing in Figure 6, showcasing various types of edits, including resizing, element swapping, and surface editing. GarmentCrafter successfully applies 3D edits that are consistent with the 2D edits, 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. To address this, we propose a new metric, the Cross-View Consistency

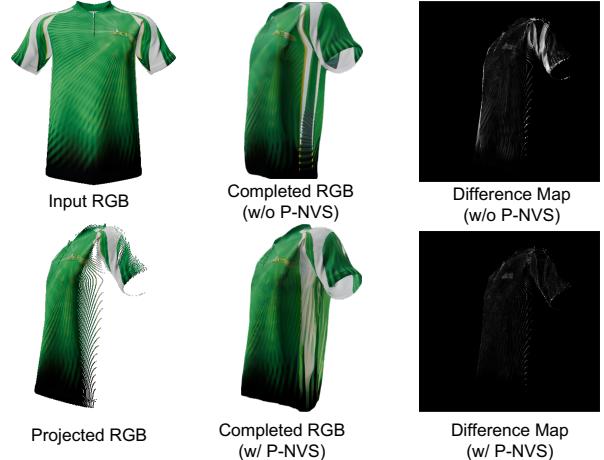


Figure 5. **Analysis of projected image conditioning.** Left: we show original input and projected RGB images. Middle: completed RGB images with and without Progressive Novel View Synthesis (P-NVS). Right: difference between completed and projected images, showing our novel view aligns more closely with the ground-truth projected RGB. Zoom-in for details.

Score (CVCS), to gain deeper insights into the consistency performance of our model.

$$\text{CVCS} = 1 - \frac{\sum |I - I'| \cdot m'}{\sum m'} \quad (5)$$

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 claim with a visual example in Figure 5. While both model synthesizes plausible novel views, GarmentCrafter with P-NVS aligns more closely with the input observation.

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 7. 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.

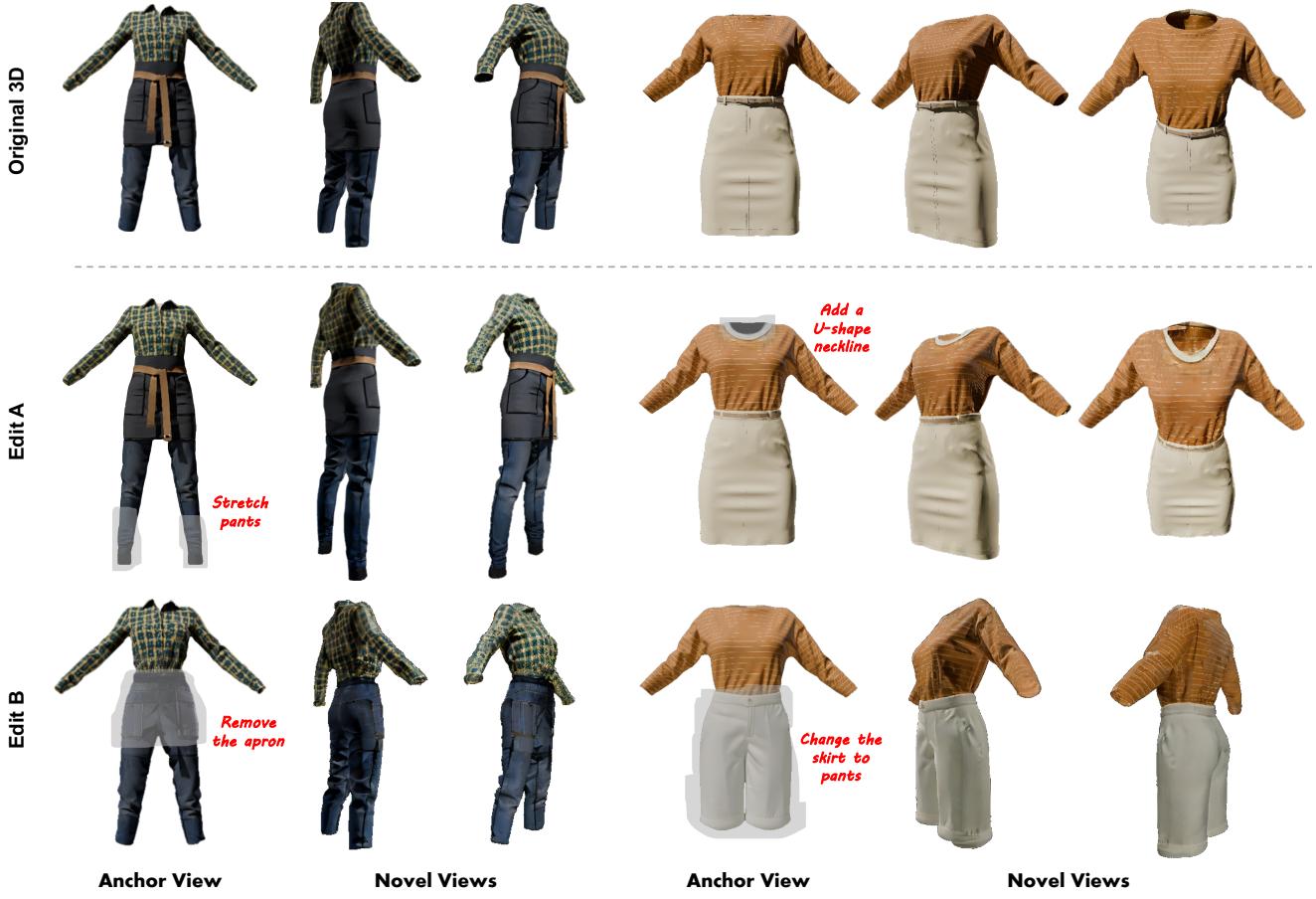


Figure 6. **Qualitative results on single-view 3D garment editing.** GarmentCrafter enables single-view edit such as modify the geometry and surface details of the garment, with the changes accurately reflected across the 3D model. Please see supplementary for more results.

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 7 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

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 geometry and texture results. We have conducted extensive experiments to demonstrate the superior performance of GarmentCrafter with other baseline methods. Please see



Figure 7. **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 model failure. We argue that there are numerous ways to select camera trajectories, our proposed approach just offers an intuitive solution tailored for single-view garment reconstruction and editing.

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.

Limitation and future works. We focus on garments in a rest pose and cannot handle arbitrary poses. In addition, our model reconstructs only the external surface, not inner layers or structures. These will be addressed in future work.

References

- [1] CLO3D. <https://www.clo3d.com/en/>. 2
- [2] Style3D. <https://www.linctex.com/>.
- [3] TUKA3D. <https://tukatech.com/tuka3d/>.
- [4] Browzwear. <https://browzwear.com/>. 2
- [5] Thieno 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. 2, 13
- [6] 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
- [7] 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
- [8] 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
- [9] 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
- [10] 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
- [11] 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
- [12] 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
- [13] 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
- [14] 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
- [15] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoungh Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 3
- [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. 2
- [17] R Daněk, Endri Dibra, Cengiz Öztureli, 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
- [18] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weih, 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
- [19] 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
- [20] 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
- [21] 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
- [22] 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. 2
- [23] 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. 2
- [24] 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
- [25] 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
- [26] 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. 2, 3, 14
- [27] 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
- [28] 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
- [29] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnets: 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 XXI 16*, pages 18–35. Springer, 2020. 2
- [30] 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
- [31] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006. 3, 4
- [32] 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. 2
- [33] 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
- [34] 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. 2
- [35] 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
- [36] 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
- [37] 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
- [38] 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
- [39] 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
- [40] 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. 2, 3
- [41] 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
- [42] 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
- [43] 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
- [44] 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
- [45] Sahib Majithia, Sandeep N Parameswaran, Sadhbhavana 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
- [46] 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. 2
- [47] 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
- [48] 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
- [49] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 3
- [50] 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
- [51] 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
- [52] 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
- [53] 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
- [54] Boxiang Rong, Artur Grigorev, Wenbo Wang, Michael J Black, Bernhard Thomaszewski, Christina Tsaligoglou, and Otmar Hilliges. Gaussian garments: Reconstructing simulation-ready clothing with photorealistic appearance from multi-view video. *arXiv preprint arXiv:2409.08189*, 2024. 2
- [55] 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
- [56] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Moriguchi, 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. 2, 13
- [57] 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, 7, 13, 15, 16
- [58] 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, 14
- [59] 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. 2
- [60] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 3
- [61] 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. 2
- [62] 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
- [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, 7
- [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
- [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. 2
- [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. 2, 13
- [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, 7
- [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 Albert, 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, 7
- [78] Xianghui Yang, Huiwen Shi, Bowen Zhang, Fan Yang, Jiacheng Wang, Hongxu Zhao, Xinhai Liu, Xinzhou Wang, Qingxiang Lin, Jiaao Yu, et al. Tencent hunyuan3d-1.0: A unified framework for text-to-3d and image-to-3d generation. *arXiv e-prints*, pages arXiv–2411, 2024. 6
- [79] 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
- [80] 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
- [81] 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
- [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. 2, 13
- [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. 2, 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

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

Supplementary Material

A. Contribution, Novelty, and Limitation

We reiterate our contribution, novelty, and limitations.

Contribution and Novelty. Our main contribution lies in a new direction: enabling non-professional users to create and edit 3D garment with single-view input. While existing works have made strides in reconstructing clothed humans [5, 56, 72, 82] or garment [57] from a single image, they mainly rely on optimizing pre-defined garment or human templates. In contrast, we target a more flexible, template-free garment reconstruction framework. Specifically, we propose to progressively synthesize depth-accurate novel view images with enhanced cross-view consistency. Moreover, our method enables single-view 3D editing, including part-based or local surface edits — capabilities that are absent in the aforementioned methods.

Scope and Limitations. As discussed in Section 5 of the main paper, our method has certain limitations. We mainly focus on garment in a rest pose. As will be shown in Section D.4, our method may struggle to accurately capture the geometry of garments in non-rest poses. With that said, this scope is a deliberate choice, as rest poses provide a consistent and intuitive baseline that aligns well with the needs of garment editing applications.

B. Ethics and Social Impacts

We focus on advancing garment digitization. We do not foresee any ethical concerns or negative societal impacts arising from our work. Our training and evaluation processes do not involve any sensitive data, human identities, or personal information. All experiments and datasets used in this study are compliant with ethical research practices. By advancing template-free garment reconstruction for non-professional users, our method avoids potential biases associated with specific body or garment templates, promoting inclusiveness in digital garment reconstruction.

C. Additional Implementation Details

In this section, we provide additional implementation details of our method omitted in the main text.

C.1. Conditional Image Generation

Our image generation model is finetuned from the Stable Zero-1-to-3 checkpoint². To account for the additional projected image as input, we add 4 additional channels to the

input convolution layer of the denosing UNet and initialize the weights to be zeros. The training resolution is 512×512 . We train the mode on 4 NVIDIA A6000 GPUs with a total batch size of 256 for 20k iterations for 2 days.

C.2. Conditional Depth Generation

Our conditional image generation model is finetuned from the Sapiens-0.3B depth checkpoint³. To add the projected partial depth map as the additional condition, we add 1 extra channels to the input projection layer of the vision transformer backbone and initialize its weights to be zeros. The training resolution is 512×512 . We train the model on 4 A6000 GPUs with a total batch size of 24 for 3 days.

C.3. Computational Efficiency

The inference time and memory consumption of our method are approximately 1 minute and 10 GB, respectively, on a single A6000 GPU. These values are comparable to those of most baseline methods, which have inference times ranging from 10 seconds to 1 minute.

C.4. Measures to Reduce Error Accumulation

Since our method synthesizes novel views in sequential steps, it is susceptible to error accumulation. To address this, we incorporate a series of techniques aimed at mitigating such errors and improving overall robustness.

Point Cloud Outlier Removal. Depth predictions near the edges of discontinuities (with large jumps in depth values) are occasionally inaccurate, resulting in some floating points in the point cloud. To address this, we apply a classical outlier removal method at each step to eliminate these floating points, ensuring a cleaner and accurate point cloud.

Open Hole Detection. We observe that depth predictions are less reliable in open-hole regions of a garment surface, such as holes in collars and sleeves. Additionally, the surface orientation derived from the estimated depth map in these areas can be reversed. These errors can propagate and lead to artifacts in subsequent steps. To address this issue, we develop a simple algorithm to detect open holes and exclude these regions during point cloud completion, improving the robustness of the pipeline.

The detection algorithm is based on the observation that the interior regions of open holes typically exhibit greater depth values compared to the boundary pixels. As shown in Figure S8, after synthesizing the completed image and

²<https://huggingface.co/stabilityai/stable-zero123>

³<https://huggingface.co/facebook/sapiens-depth-0.3b>

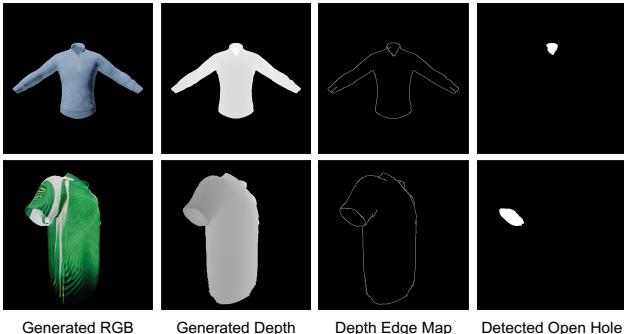


Figure S8. Open hole detection in garments. We note that interior regions of open holes in a garment exhibit greater depth values compared to the boundary pixels. Leveraging this observation, we propose a simple yet effective algorithm to detect open holes and exclude these regions during point cloud completion, improving the robustness of the pipeline.

depth maps from a novel viewpoint, we first detect edges in the depth map and identify connected regions enclosed by these edges using classical methods. A connected region R is classified as an open hole if more than a threshold ϵ of its boundary pixels have depth values smaller than the average depth of the region. For all our experiments, we found that ϵ can be robustly set to 0.85.

Clipping Distant Depth Values. Our observations indicate that synthesized images and depth maps are more robust in regions closer to the camera compared to those farther away. At steps 3 and 4 (corresponding to azimuth angles of 120° and -120°), the entire back side of the garment is synthesized from a side view. For these steps, we only use pixels with smaller depth values for point cloud completion, disregarding pixels with larger depth values.

C.5. Point-to-Mesh Reconstruction

We use Screened Poisson surface reconstruction to convert point clouds to meshes. Note that the point orientations are estimated using depth maps at each step, as described in Section 3.3 of the main paper. While Screened Poisson reconstruction generates a watertight mesh, we aim to preserve the non-watertight topology of garments (e.g., maintaining holes in collars, sleeves, etc.). To achieve this, we perform an additional trimming operation⁴ to remove unwanted mesh faces introduced during the Poisson reconstruction that fill open holes by leveraging the point cloud density. To further reduce artifacts, we remove floating faces unconnected to the main mesh and apply Laplacian smoothing to refine the mesh surface.

C.6. Scope of Single-View 3D Editing

As introduced in Section 3.3 of the main paper, GarmentCrafter enables single-view editing through a simple

workflow: identify the edited 3D region, remove the original mesh in the identified area, and reconstruct the edited components. We support two types of editing operations, differentiated by their assumptions about the edited regions.

The first category is local surface editing. Given a camera viewpoint and a mask, this approach assumes that only the visible surface intersected by the camera rays corresponding to the masked pixels will be edited. Occluded surfaces are ignored, even if their mesh vertices project within the mask. To facilitate reconstruction, we remove the mesh vertices of the selected surface. Additionally, internal vertices near the external surface are also removed to account for surface thickness.

The second category, part-based editing, involves modifying a 3D garment part, including not only the “front” surface but also the “back” and “internal” surfaces within a masked region. For ease of implementation, we always use the frontal view as the editing perspective and remove all mesh vertices whose 2D projections fall within the mask.

Our editing pipeline is designed under the assumption that both the geometry and the texture will be edited. Therefore, it is not optimized for cases where (1) surface texture is modified while preserving the geometry, or (2) the geometry or pose is altered while preserving the texture.

C.7. 2D Editing Assumptions

In theory, our method is agnostic to the tools used for 2D editing. The edits can be created using deep learning-based image editing models or traditional tools like Photoshop. However, our approach requires the edits to be confined to regions specified by masks in the 2D input. Therefore, global edits such as style transfer that alters the entire image, are not recommended.

D. Additional Results and Analyses

D.1. Intermediate Results of Progressive NVS

In Figure 2 of the main paper, we showed results at one specific camera rotation step during the progressive novel view synthesis. Here, we illustrate the whole process and show the intermediate results in Figure S9.

D.2. Additional Baseline Comparisons

D.2.1. Comparison with SoTA NVS methods

We present additional quantitative comparisons for novel view synthesis against state-of-the-art methods (Zero-1-to-3++ [58] & MVD-Fusion [26], fine-tuned with same data). For each object in the held-out test set of 150 garment assets, we sample six camera viewpoints with an elevation of 20 degrees and evenly spaced azimuth angles covering 360 degrees. Each method takes a frontal image as input and generates six corresponding novel views, which we evaluate against ground truth images using image similarity metrics

⁴<https://github.com/mkazhdan/PoissonRecon>



Figure S9. **Intermediate results of progressive novel view synthesis along a full camera trajectory.** From an input RGB image (top-left), GarmentCrafter progressively synthesizes novel view RGB and depth maps following a zigzag camera trajectory.

(LPIPS, PSNR, and SSIM). We also report our proposed CVCS score. Table S4 shows that our method achieves superior performance across all metrics.

D.2.2. Qualitative Comparison with Garment3DGen

As the texture reconstruction code of Garment3DGen [57] is not released, we provide qualitative comparison with Garment3DGen on the reconstructed mesh geometry in Figure S10. Our method reconstructs 3D garments with much richer geometric details and much less inference time (1 min vs. 3 hours).

D.3. Additional Analyses and Applications

D.3.1. Degree of Zigzag Camera Trajectory

We have studied all major design choices in our pipeline in the main paper, including the effect of progressive novel

Table S4. **Quantitative comparison for novel view synthesis.** Our method outperforms all state-of-the-art novel view synthesis methods cross both image similarity and consistency metrics.

	LPIPS ↓	PSNR ↑	SSIM ↑	CVCS↑
Zero123++	0.1611	18.023	0.7979	0.8957
MVD-Fusion	0.1528	18.529	0.8026	0.9090
Ours	0.1052	22.776	0.8557	0.9512

view synthesis and camera trajectory. Here, we analyze the impact of the degree of Zigzag Camera Trajectory and show the results in Table S5. In our experiments, we use a 60° trajectory as it provides a good balance between view coverage and efficiency. While the choice of degree slightly affects the ability to synthesize side-view garments (i.e., 90°), our analysis indicates that the overall performance is not highly

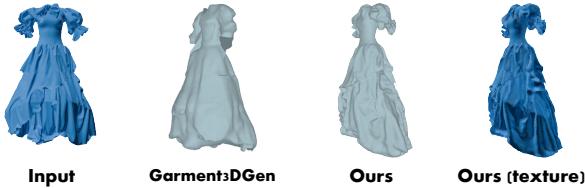


Figure S10. **Qualitative comparison with Garment3DGen [57].** Our GarmentCrafter reconstructs garment meshes with richer details with much lower computational costs.

Table S5. Analysis of the degree of zigzag camera trajectory. In our experiments, we use a 60° trajectory as it provides a good balance between view coverage and efficiency. While the choice of degree slightly affects the ability to synthesize side-view garments (i.e., 90°), our analysis indicates that the overall performance is not highly sensitive to this parameter.

Degree	Appearance			Geometry
	SSIM↑	LPIPS↓	PSNR ↑	Chamfer↓
30°	0.8044	0.1675	20.62	0.0051
60°	0.8066	0.1638	<u>20.62</u>	0.0050
90°	0.8003	0.1709	20.19	0.0070
120°	<u>0.8053</u>	<u>0.1654</u>	20.51	<u>0.0050</u>



Figure S11. **Failure case.** GarmentCrafter may fail to reconstruct the garment with arbitrary poses.

sensitive to this parameter. We do not notice any other significant hyperparameters in our framework.

D.3.2. Digitizing AI-generated Apparel

We explore the potential of combining GarmentCrafter with AI-generated garment image and show examples in Figure S14. Using a text-to-image generative model, we produce synthetic garment images and apply GarmentCrafter to digitize them. The results demonstrate the broad applicability of our method in handling diverse inputs, including AI-generated designs.

D.4. Failure Cases

The focus of our work is on reconstructing and editing garments in their rest pose. Consequently, our method struggles with input images in arbitrary poses as such instances lie outside of the training data distribution. As illustrated in Figure S11, an input garment image in a non-resting pose results in the failure of our model to synthesize coherent novel view images, leading to nonsensical reconstructions.

D.5. More Qualitative Results

Reconstruction. Please see more results in Figure S12.

Editing. We provide more qualitative results in Figure S13.



Input RGB

Recon. Mesh

Novel Views

Figure S12. More qualitative result on single-view 3D garment reconstruction.



Figure S13. More results on single-view 3D garment editing. The top row illustrates how GarmentCrafter effectively handles surface edits, even for regions with complex textures. The middle row demonstrates the capability of GarmentCrafter to support full garment changes and swaps, showcasing the potential in virtual try-on scenarios. The bottom row presents an example of removing an entire garment part.



Figure S14. Compatibility with generative apparel. By reconstructing both geometry and texture from synthetic garment images, GarmentCrafter demonstrates its adaptability to AI-generated designs. The results showcase the ability of GarmentCrafter to handle diverse and complex inputs, expanding its potential applications to generative fashion and virtual apparel workflows.