GarVerseLOD: High-Fidelity 3D Garment Reconstruction from a Single In-the-Wild Image using a Dataset with Levels of Details

ZHONGJIN LUO, SSE, CUHKSZ, China
HAOLIN LIU, FNii, CUHKSZ, China and SSE, CUHKSZ, China
CHENGHONG LI, FNii, CUHKSZ, China and SSE, CUHKSZ, China
WANGHAO DU, SSE, CUHKSZ, China
ZIRONG JIN, SSE, CUHKSZ, China
WANHU SUN, SSE, CUHKSZ, China
YINYU NIE, Huawei Noah's Ark Lab, UK
WEIKAI CHEN, DCC Algorithm Research Center, Tencent Games, USA
XIAOGUANG HAN*, SSE, CUHKSZ, China and FNii, CUHKSZ, China

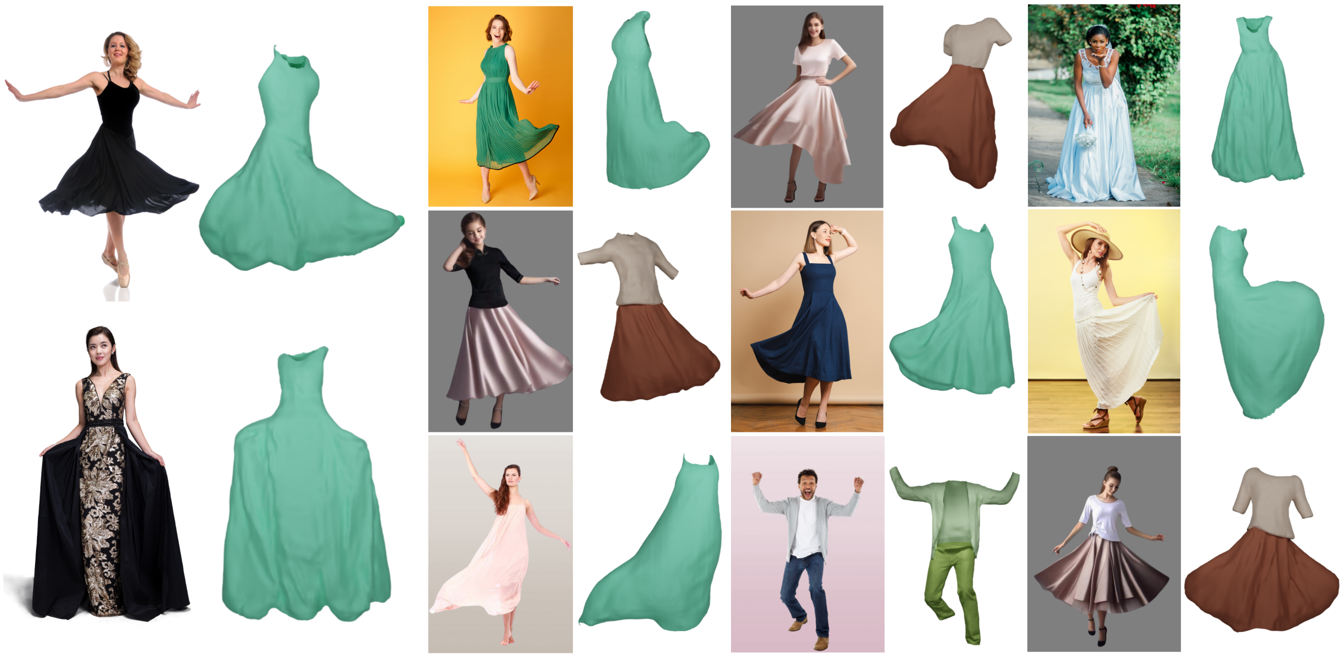

Fig. 1. We propose a hierarchical framework to recover different levels of garment details by leveraging the garment shape and deformation priors from the GarVerseLOD dataset. Given a single clothed human image, our approach is capable of generating high-fidelity 3D standalone garment meshes that exhibit realistic deformation and are well-aligned with the input image. Original images courtesy of licensed photos and Stable Diffusion [Rombach et al. 2022]. The images with a gray background are synthesized, while the rest are licensed photos.

Neural implicit functions have brought impressive advances to the state-of-the-art of clothed human digitization from multiple or even single images. However, despite the progress, current arts still have difficulty generalizing to unseen images with complex cloth deformation and body poses. In this work, we present GarVerseLOD, a new dataset and framework that paves the way to achieving unprecedented robustness in high-fidelity 3D garment reconstruction from a single unconstrained image. Inspired by the recent success of large generative models, we believe that one key to addressing the generalization challenge lies in the quantity and quality of 3D garment data. Towards this end, GarVerseLOD collects 6,000 high-quality

*Corresponding author is Xiaoguang Han (hanxiaoguang@cuhk.edu.cn).

Authors' addresses: Zhongjin Luo, SSE, CUHKSZ, Shenzhen, China; Haolin Liu, FNii, CUHKSZ, China and SSE, CUHKSZ, China; Chenghong Li, FNii, CUHKSZ, China and SSE, CUHKSZ, China; Wanghao Du, SSE, CUHKSZ, China; Zirong Jin, SSE, CUHKSZ, China; Wanhu Sun, SSE, CUHKSZ, China; Yinyu Nie, Huawei Noah's Ark Lab, UK; Weikai Chen, DCC Algorithm Research Center, Tencent Games, USA; Xiaoguang Han, SSE, CUHKSZ, Shenzhen, China and FNii, CUHKSZ, Shenzhen, China.

ACM 0730-0301/2024/12-ART
https://doi.org/10.1145/3687921

cloth models with fine-grained geometry details manually created by professional artists. In addition to the scale of training data, we observe that having disentangled granularities of geometry can play an important role in boosting the generalization capability and inference accuracy of the learned model. We hence craft GarVerseLOD as a hierarchical dataset with *levels of details (LOD)*, spanning from detail-free stylized shape to pose-blended garment with pixel-aligned details. This allows us to make this highly under-constrained problem tractable by factorizing the inference into easier tasks, each narrowed down with smaller searching space. To ensure GarVerseLOD can generalize well to in-the-wild images, we propose a novel labeling paradigm based on conditional diffusion models to generate extensive paired images for each garment model with high photorealism. We evaluate our method on a massive amount of in-the-wild images. Experimental results demonstrate that GarVerseLOD can generate standalone garment pieces with significantly better quality than prior approaches while being robust against a large variation of pose, illumination, occlusion, and deformation. Code and dataset are available at garverselod.github.io.

CCS Concepts: • **Computing methodologies → Shape inference**; **Reconstruction**.

Additional Key Words and Phrases: Image-based Modeling, 3D Garment Reconstruction, 3D Garment Dataset

ACM Reference Format:
Zhongjin Luo, Haolin Liu, Chenghong Li, Wanghao Du, Zirong Jin, Wanhu Sun, Yinyu Nie, Weikai Chen, and Xiaoguang Han. 2024. GarVerseLOD: High-Fidelity 3D Garment Reconstruction from a Single In-the-Wild Image using a Dataset with Levels of Details. *ACM Trans. Graph.* 43, 6 (December 2024), 21 pages. https://doi.org/10.1145/3687921

1 INTRODUCTION

High-quality 3D garment models are critical assets for a large variety of applications, ranging from entertainment to professional concerns, such as visual effects, physical simulation, and VR/AR telepresence. In the production-level pipeline, independent garment pieces are more desirable than a single clothed human model, as the former allows layered compositions with an internal body mesh to ensure the realism of physical motion and the flexibility of garment transfer. However, unlike clothed human reconstruction that can directly utilize the latest advances of neural implicit representation [Saito et al. 2019, 2020; Xiu et al. 2022], standalone garment modeling mostly relies on deforming parametric templates with open boundaries due to its strict requirement of correct topology.

Nonetheless, reconstructing high-fidelity 3D garment from a single image remains a nuisance to current vision algorithms. While the high diversity of garment styles and the scarcity of the inputs render the problem highly ill-posed, the complex deformations resulted from the cloth dynamics make the inference even more challenging. There are two mainstream approaches for estimating the deformations of standalone garments from posed humans. Linear blend skinning (LBS)-based methods [Corona et al. 2021; Jiang et al. 2020] focus on predicting the deformations caused by human poses, where the learned skinning weights of the garment mesh are either bound to the skeleton or the surface vertices of a parametric model of unclothed humans (e.g., SMPL [Loper et al. 2015]). While this line of approaches can effectively represent posed-induced deformations, they struggle to model other intricate deformations caused by the environments or physical dynamics. Feature-line-based methods [Zhu et al. 2020, 2022] reconstruct garment meshes from SMPL surfaces and further fit them with garments' manifold boundaries, making it versatile to model any type of deformations. However, the problem of boundary estimation from single images itself is challenging, due to the severe occlusions and 2D-to-3D ambiguities.

Apart from the technical challenges, the other obstacle to learning-based garment reconstruction is the limited quantity and quality of 3D dataset. Due to the lack of local geometry details in existing garment datasets, current LBS-based methods are incapable of learning fine-grained geometries (e.g., wrinkles), resulting in coarse 3D garment quality. ReEF [Zhu et al. 2022] annotates the feature lines for only 400 garment models in the RenderPeople dataset [RenderPeople 2018]. The limited data scale hampers the prior approaches from generalizing to unseen images and often leads to poor reconstruction quality of feature lines (i.e., garment boundaries).

In this work, we strive to address the above issues for standalone 3D garment reconstruction from the perspectives of both data and algorithm. We thereby introduce GarVerseLOD, a dedicated dataset and framework that achieves unprecedented robustness in reconstructing high-fidelity 3D garments from a single in-the-wild image (Fig. 1). To promote the quantity and quality of 3D garment data, GarVerseLOD collects 6,000 high-quality hand-crafted garment meshes with fine-grained details created by professional artists. It covers 5 most commonly seen categories – each category shares the same mesh topology, facilitating cross-instance interpolation and construction of blendshape models. While garment shapes differ globally in terms of style and topology, the local deformations are determined by a wide range of factors, including body poses, garment-environment interactions, self-collisions, *etc.* We, therefore, propose to craft GarVerseLOD as a hierarchical dataset with *levels of details (LOD)* to accommodate this key observation.

In particular, as shown in Fig. 2, GarVerseLOD contains three basic levels of databases: 1) *Garment Style Database* with T-posed and detail-free coarse garment; 2) *Local Detail Database* enclosing pairs of T-posed models with and without fine-level local geometric details; and 3) *Garment Deformation Database* consisting of pairs of T-posed garment and its deformed counterpart (i.e., with global deformations). As the mesh topologies are identical within each category, we can easily extract the local details and global deformations from paired models in the corresponding database and combine all levels of geometries to obtain the *Fine Garment Dataset*. The disentangled granularities of geometry allows us to make this highly underconstrained problem tractable by factorizing the inference into smaller tasks, each can be tackled with narrowed solution space. Furthermore, we introduce a novel data labeling paradigm to generate extensive paired images for each garment model. Specifically, we leverage the latest advances in conditional diffusion model to transfer the textureless renderings to photorealistic images with diverse appearances. This further elevates the generalization capability of GarVerseLOD in handling unconstrained images.

Algorithm-wise, we propose to connect the good ends of both LBS and feature-line based approaches. We first build a parametric model of the T-posed coarse shapes in the garment style database. After estimating the blendshape coefficients of the coarse garment, we progressively refine the result by adding pose-induced global deformations and fine-scale local deformations. Thanks to the LOD structure of GarVerseLOD, these three steps can be performed in a disentangled manner with eased complexity. While we employ linear blend skinning to estimate deformations caused by body poses, an implicit garment representation is learned to capture pixel-aligned fine surface from estimated 2D normal maps. We then fit the posed coarse garment with fine surfaces by aligning their open boundaries for the purpose of transferring the local details to the globally deformed mesh with correct topology. To combat with the occlusions, we present a novel geometry-aware boundary prediction strategy that equips the 2D features with 3D information from the estimated fine surface for better localization of 3D boundaries. Our experimental results show that GarVerseLOD can effectively reconstruct garments with diversified shapes and intricate deformations, demonstrating significantly better generalization ability over the prior arts. We summarize our contributions as follows:

- We present the GarVerseLOD dataset, a large collection of high-fidelity 3D *hand-crafted* garments. It encloses 6,000 professionally hand-crafted garments, covers 5 categories, and, for the first time, contains 3 disentangled levels of details to ease the learning task.
- We propose a novel data simulation pipeline to generate extensive paired images for supporting single-view reconstruction.
- We devise a specially-tailored coarse-to-fine approach to fully utilize the LOD structure of the GarVerseLOD dataset. Experimental results show that our method excels in reconstructing high-quality garments from single images.

2 RELATED WORK

3D Human Reconstruction. 3D reconstruction has seen significant advancements recently [Habermann et al. 2019, 2020; Jiang et al. 2022; Li et al. 2021; Loper et al. 2015; Luo et al. 2023, 2021; Poole et al. 2022; Saito et al. 2019; Xu et al. 2018; Yan et al. 2024]. Some single-view human reconstruction methods [Anguelov et al. 2005; Hasler et al. 2009; Lassner et al. 2017; Pavlakos et al. 2019; Xu et al. 2019] restrict the solution space to a parametric human model and simplify the problem, which can only reconstruct nude human 3D models without garments. Inspired by SMPL [Loper et al. 2015], some methods [Alldieck et al. 2019; Tan et al. 2020; Xiang et al. 2020; Yang et al. 2018; Zheng et al. 2019] approximate human body geometry by deforming the SMPL. These methods can reconstruct realistic results from an unconstrained image but fail to handle loose garments. Contrary to the SMPL-based approaches, other methods enable clothed human body reconstruction with arbitrary topology. Siclope [Natsume et al. 2019] reconstructs clothed 3D human models using multi-view silhouettes predicted from a frontal image. DeepHuman [Zheng et al. 2019] generates progressively refined voxels, which are embossed with details from a surface normal. While both methods can produce clothed human shapes with arbitrary topology, the details are relatively coarse. Recent works [Li et al. 2020; Saito et al. 2019, 2020; Xiu et al. 2023, 2022] address this issue with pixel-aligned implicit functions and achieve high reconstruction fidelity. However, all the above methods fail to provide the garment mesh separated from the human body.

3D Garment Reconstruction. Compared to clothed 3D human bodies, reconstructing high-fidelity 3D independent garments from a single image is challenging. Many prior arts rely on learning-based strategies to fit 3D garment deformations from a collection of 2D image-3D garment pairs for generalization. Two mainstream approaches are often used in estimating standalone garments from single images. Linear blend skinning (LBS)-based methods (e.g., BCNet [Jiang et al. 2020], ClothWild [Moon et al. 2022]) focus on predicting deformations caused by human poses, where explicit or implicit garment parametric models are used. These methods can address posed-guided deformations but fail to reproduce large cloth deformations (e.g., those caused by complex environmental factors) and fine-grained surface details (e.g., wrinkles). Recent works, such as ISP [Li et al. 2024b] and Neural-ABC [Chen et al. 2024], have proposed more advanced implicit parametric models, but their reconstruction methods still rely solely on their parametric models. Similar to BCNet [Jiang et al. 2020] and ClothWild [Moon et al. 2022], the limited representational capacity of the parametric models prevents them from accurately recovering complex garment deformations from images. Feature-line-based methods [Zhu et al. 2020, 2022] reconstruct garment meshes from SMPL and further fit them with the estimated pixel-aligned clothed human and the predicted garment's manifold boundaries, making them flexible to model cloth deformations and geometric details. However, the problem of boundary estimation from single images itself is challenging, due to the severe occlusions and 2D-to-3D ambiguities. Garment Recovery [Li et al. 2024a] relies on a normal estimator trained on human data with limited clothing diversity and a deformation prior trained with limited deformation variations, preventing it from accurately reconstructing high-fidelity surface details and complex deformations that reflect the inputs. All existing methods cannot faithfully recover intricate clothing deformations and fine-grained geometric details from single-view images.

3D Garment Datasets. 3D garment datasets are an important foundation for learning-based tasks, but current available datasets are very limited in quality and scale. Existing datasets can be divided into two major categories: scanning-based datasets [Bhatnagar et al. 2019; Lin et al. 2023; Pons-Moll et al. 2017; Tiwari et al. 2020; Wang et al. 2024; Zhang et al. 2017; Zhu et al. 2020] and simulation-based datasets [Bertiche et al. 2020; Black et al. 2023; Gundogdu et al. 2019; Jiang et al. 2020; Patel et al. 2020; Zou et al. 2023]. Scanning-based datasets allow for realistic garment appearance and shape. However, separating garment models from 3D scans is laborious and often results in surface damage due to occlusion. These datasets usually are restricted by data scale and suffer from the inability to separate garments from the mannequin [Zhang et al. 2017], as well as insufficient clothing diversity. Simulation-based datasets synthesize 3D garments by simulating motion using physics-based engines. However, these synthetic datasets are unsatisfactory in terms of cloth style, body pose and garment deformation variations, as well as the quality of paired images. It is difficult to generalize the trained models to in-the-wild images. We introduce a large-scale 3D garment dataset characterized by intricate deformations and fine-grained surface details. Additionally, we present a novel data simulation strategy to collect extensive image-3D garment pairs by leveraging the generative capabilities of conditional stable diffusion models [Mou et al. 2023; Rombach et al. 2022; Zhang et al. 2023a].

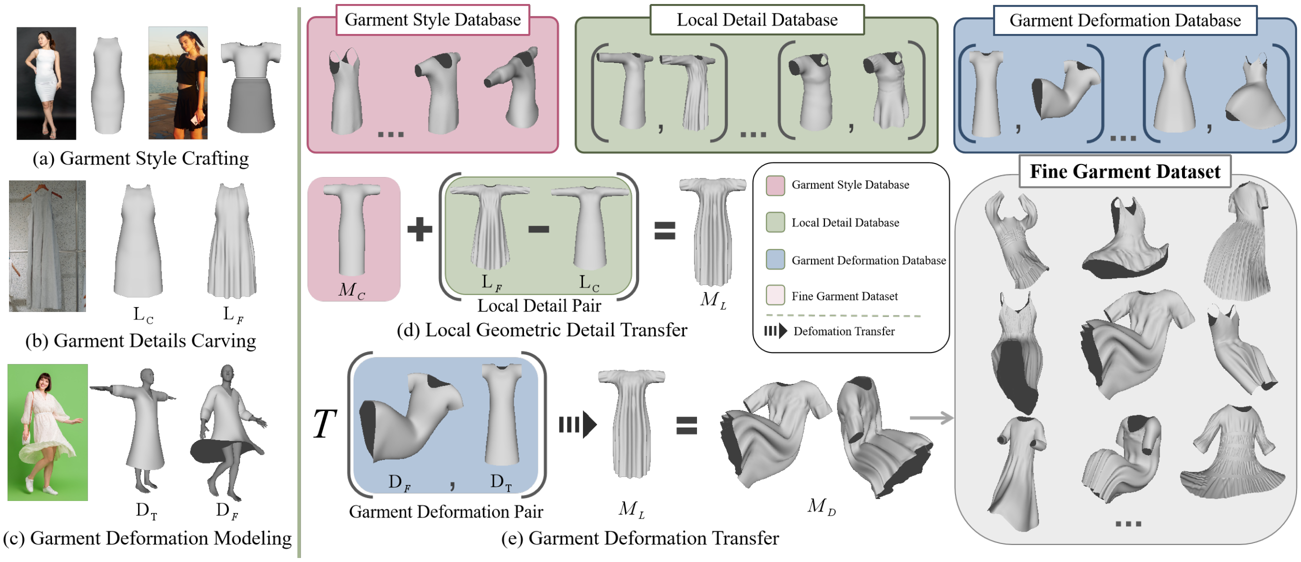

Fig. 2. The pipeline of our novel strategy for constructing a progressive garment dataset with levels of details. (a) Each case shows the reference image and the artist-crafted T-pose coarse garment in **Garment Style Database**. (b) A example of the reference image and the artist-crafted detail-pair in **Local Detail Database**. (c) A example of the reference image and the artist-crafted deformation-pair in **Garment Deformation Database**. (d) To obtain an T-pose garment with geometric details, we first sample a shape M_C from the Garment Style Database and a "Local Detail Pair" (L_C, L_F) from the Local Detail Database. Then we transfer the geometric details depicted by (L_C, L_F) to M_C to obtain M_L. (e) The deformation depicted by a sampled "Garment Deformation Pair" (D_T, D_F) is transferred to M_L to obtain the fine garment M_D, which contains fine-grained geometric details and complex deformations (**Fine Garment Dataset**). Original images courtesy of licensed photos.

3 DATASET

Reconstructing accurate and standalone garments from single images remains a significant challenge due to the absence of a well-established dataset, especially for scenarios involving complex deformations like human-garment or garment-environment interactions. Existing datasets suffer from limited data scale [Bhatnagar et al. 2019; Jiang et al. 2020], or a lack of paired 2D image-3D garment data [Bertiche et al. 2020], or solely monotonous rendered images paired with clothing models [Zou et al. 2023]. In this work, we fill this gap by introducing GarVerseLOD, a progressive dataset with levels of details (LOD). Additionally, we present a novel data generation pipeline to construct a large-scale dataset with realistic images paired with 3D models in GarVerseLOD.

Overview. GarVerseLOD has four features: 1) **Broad diversity**. GarVerseLOD contains 5 common garment categories, i.e. dress, skirt, coat, top, and pant. Each 3D model comprises fine-grained geometric details and intricate clothing physical deformations. 2) **Levels of details**. We collect three basic databases with different *levels of details* (LOD) to obtain high-quality 3D clothes. Based on these databases, we create a large number of posed 3D garments with complex deformations and fine-grained geometric details. 3) **Topological consistency**. Each 3D garment in GarVerseLOD is created by carefully deforming a pre-defined template mesh. All 3D garments within different categories share a unified topology, paving the way to learn a parametric model. 4) **Extensive paired data**. To create high-quality image-3D garment paired data, we employ ControlNet [Zhang et al. 2023a] and T2I-Adapter [Mou et al. 2023] as the data simulator and transform monotonous rendered images into photorealistic images with diverse appearances. Some 2D image-3D garment pairs are shown in Fig. 3. Please refer to the supplementary materials for more details.

3.1 LOD Garment Crafting

As shown in Fig. 2, we first construct three basic databases with different levels of detail: 1) **Garment Style Database**. In Fig. 2(a), we collected a set of reference images of clothed humans with diverse garment styles from the Internet and hired eight artists to craft a T-posed coarse garment for each reference image (without surface geometric details like wrinkles, only depicting the overall cloth shape); 2) **Local Detail Database**. In Fig. 2(b), we collected a set of reference images of clothes with diverse surface details (e.g., wrinkles). The eight artists were asked to carve two T-posed garments for each image: one without surface details (L_C) and one with fine surface details (L_F). These garment pairs (L_C, L_F) describe garment local geometric details; 3) **Garment Deformation Database**. In Fig. 2(c), we collected a set of reference images of clothed humans with diverse poses and garment deformations. For each image, we use PyMAF [Zhang et al. 2021] to estimate the SMPL shape β and pose θ from the images. The artists were asked to construct two over-smoothed garments (i.e., garments without local geometric details): a T-posed garment (i.e, D_T on top of the estimated T-pose body) and a garment with global deformation aligned with the image (i.e., D_F, on top of the posed body). These two garments (D_T, D_F) form a pair that depicts garment deformations. Note that the

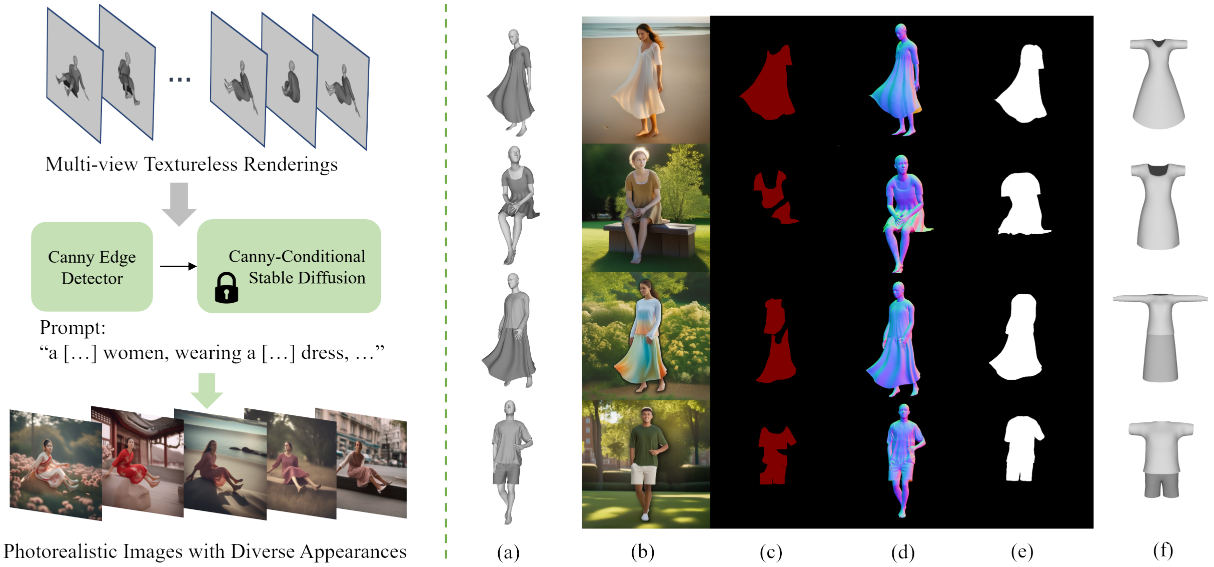
Multi-view Textureless Renderings
Canny Edge Detector
Canny-Conditional Stable Diffusion
Prompt:
“a [...] women, wearing a [...] dress, ...”
Photorealistic Images with Diverse Appearances
(a)
(b)
(c)
(d)
(e)
(f)

Fig. 3. Left: Our novel strategy for generating extensive photorealistic paired images. We acquire rendered images of 3D garments with random camera views. These rendered images are processed through Canny-Conditional Stable Diffusion [Mou et al. 2023; Rombach et al. 2022; Zhang et al. 2023a] to produce photorealistic images. Right: (a) The garment sampled from Fine Garment Dataset; (b) The synthesized image; (c) The pixel-aligned mask; (d) The normal map rendered using (a); (e) The garment mask rendered by (a); (f) The counterpart T-pose coarse garment of (a). In Sec. 4, (b, f) is used to train the coarse garment estimator, while (b,c,d) is adopted to train the normal estimator. (d, e, a) is utilized to train the fine garment estimator and the geometry-aware boundary predictor. Synthesized images courtesy of Stable Diffusion.

estimated SMPL parameters are also stored to assist deformation transfer in the following fine garment synthesis.

Fine Garment Dataset. All models in the above three databases are created by deforming predefined templates (i.e., dress, skirt, coat, top, and pant). Thus, all 3D garments within different categories are homeomorphic in topology. The feature of **topological-consistency** not only paves the way to learning a parametric model (Sec. 4.1), but also enables the incorporation of our three basic databases to obtain the fine garment dataset. As shown in Fig. 2, we first sample a coarse garment shape M_C by interpolating between garments in the Garment Style Database. Then we sample a “Local Detail Pair” (L_C, L_F) and apply their vertex offsets to M_C by,

$$M_L = M_C + L_F - L_C, \tag{1}$$

where M_L denotes the T-posed garment with local details transferred from L_F. Subsequently, we sample a “Garment Deformation Pair” (D_T, D_F) from the Garment Deformation Database and transfer the deformation to M_L to obtain the fine garment M_D by,

$$M_D = LBS(M_L + T), \tag{2}$$

$$T = LBS^{-1}(D_F) - D_T, \tag{3}$$

where we apply the inverse LBS of SMPL to garment D_F to obtain the deformation offsets T in the rest-pose space. Then T is applied to M_L in the rest-pose space. The forward LBS is used to deform M_L to pose space to obtain the fine garment M_D, which contains both fine-grained surface details and complex garment deformations.

3.2 Photorealistic Paired Image Generation

We present a data simulation pipeline to synthesize images paired with our 3D garments. Specifically, we utilize ControlNet and T2I-Adapter to transfer monotonous rendered images to photorealistic images with diverse appearances. As illustrated in Fig. 3, we first obtain 3D garment renderings with random camera views. Then, the rendered images are fed to Canny-Conditional Stable Diffusion (i.e., ControlNet and T2I-Adapter) to obtain realistic RGB images. We calculate the L_1 loss on canny edges between renderings and the generated images, and manually pick generated images that closely approximate the 3D garment shape. Finally, all generated images are manually inspected to ensure high consistency between the image and the corresponding 3D garment.

Local Alignment. Although ControlNet and T2I-Adapter perform well in generating images with correct global shapes, it is still challenging to produce images with pixel-aligned details, such as wrinkles (see Fig. 3(a, b, d)). To support fine-grained inference, as shown in Fig. 3(c), a pixel-level alignment mask is labeled manually to mark out the alignment region between the synthesized image (b) and the rendered normal map (d). This leads to a collection of high-quality paired data that can be utilized for 3D garment reconstruction.

4 METHOD

As shown in Fig. 4, given an RGB image, our approach initially estimates the coarse explicit garment shape M_P (Sec. 4.1). Geometric details are recovered from the implicit function with the assistance of the normal map to obtain a fine garment mesh with a closed boundary M_I (Sec. 4.2). Next, our method combines the 2D image

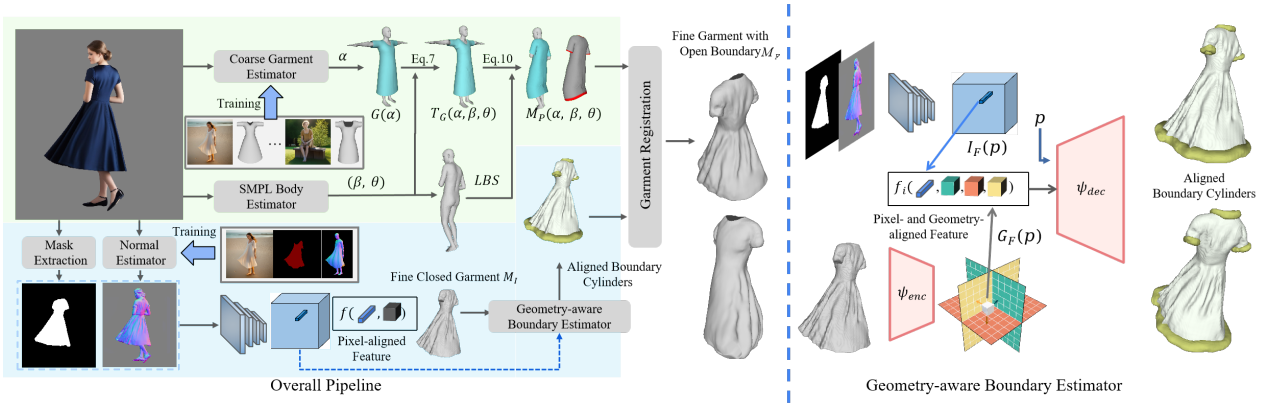
Coarse Garment Estimator
α
Eq.7
Eq.10
Training
$G(\alpha)$
$T_G(\alpha, \beta, \theta)$
$M_P(\alpha, \beta, \theta)$
SMPL Body Estimator
(β, θ)
LBS
Garment Registration
Fine Garment with Open Boundary M_F
Mask Extraction
Normal Estimator
Training
Fine Closed Garment M_I
Aligned Boundary Cylinders
Geometry-aware Boundary Estimator
$f($, $)$
Pixel-aligned Feature
Overall Pipeline
p
$I_F(p)$
$f_i($, , , $)$
ψ_{dec}
Pixel- and Geometry-aligned Feature
$G_F(p)$
ψ_{enc}
Aligned Boundary Cylinders
Geometry-aware Boundary Estimator

Fig. 4. The pipeline of our proposed method. Given an RGB image, our method first estimates the T-pose garment shape $G(\alpha)$ (Eq. 4) and computes its pose-related deformation $M_P(\alpha, \beta, \theta)$ with the help of the predicted SMPL body (Eq. 7, Eq. 10). Then a pixel-aligned network is used to reconstruct implicit fine garment M_I and the geometry-aware boundary estimator is adopted to predict the garment boundary. Finally, we register $M_P(\cdot)$ to M_I to obtain the final mesh M_F, which has fine topology and open-boundaries. Images courtesy of Stable Diffusion.

and the 3D fine garment to predict the garment boundary (Sec. 4.3). Finally, we fit the coarse shape with the fine garment mesh by aligning the 3D boundaries to generate the target garment mesh M_F with an open boundary (Sec. 4.4).

4.1 Coarse Explicit Garment Estimation

4.1.1 Unposed Coarse Garment Inference.

Garment Blendshape Construction. Various current works [Jiang et al. 2020; Luo et al. 2023; Patel et al. 2020] demonstrate that linear statistical models are able to represent the basic geometries of diverse shapes. Inspired by [Jiang et al. 2020; Loper et al. 2015], we utilize PCA to parameterize our unposed (i.e. T-posed) coarse garments by,

$$G(\alpha) = \mathbf{T}_g + B_g(\alpha), \tag{4}$$

where $G(\alpha)$ denotes the statistical garment model, which is worn on top of SMPL's mean shape. $\mathbf{T}_g \in \mathbb{R}^{N_G \times 3}$ is a garment template with N_G vertices. We define an independent T-posed garment template $\mathbf{T}_g$ for each garment category. $B_g(\alpha) \in \mathbb{R}^{N_G \times 3}$ models the *Garment Shape Blend Shapes* (GSBS) in T-posed space, while $\alpha \in \mathbb{R}^{32}$ is the PCA coefficients that control the GSBS.

Coarse Garment Estimator. Given an input image, we firstly utilize a lightweight image classifier [Jiang et al. 2020; Zhu et al. 2020, 2022] to categorize it into one of five common types and select the corresponding statistical garment model. With the selected statistical model, we use a CNN encoder to map the image to the parametric space and obtain T-posed coarse garment $G(\alpha)$ through Eq. 4.

4.1.2 Posed Coarse Garment Estimation.

To model garment's coarse deformation, we extend SMPL's skinning procedure to the pose garment. SMPL incorporates *Body Shape Blend Shapes* (BSBS) and *Body Pose Blend Shapes* (BPBS) to define a T-posed body. Given the shape parameters (β) and pose parameters (θ), SMPL can generate the T-posed body mesh by,

$$T_B(\beta, \theta) = \mathbf{T}_b + B_s(\beta) + B_p(\theta), \tag{5}$$

where $\mathbf{T}_b \in \mathbb{R}^{N_B \times 3}$ is a body template mesh with N_B vertices. $B_s(\beta) \in \mathbb{R}^{N_B \times 3}$ denotes the shape-related displacements, while $B_p(\theta) \in \mathbb{R}^{N_B \times 3}$ models the pose-dependent correctiveness. Then SMPL uses LBS to pose a rigged template. The mapping can be summarized as the following equation:

$$M_B(\beta, \theta) = W\left(T_B(\beta, \theta), J(\beta), \theta, \mathcal{W}\right), \tag{6}$$

where $W(\cdot)$ is a skinning function with skinning weights $\mathcal{W} \in \mathbb{R}^{N_B \times 24}$, joint locations $J(\beta) \in \mathbb{R}^{24 \times 3}$, and pose parameters θ that rig a T-posed body mesh $T_B(\beta, \theta)$.

SMPL Body Estimator. To pose our garment, we use PyMAF [Zhang et al. 2021] to estimate SMPL parameters from the image, and apply BSBS and BPBS to the T-posed garment by,

$$T_G(\alpha, \beta, \theta) = G(\alpha) + \widetilde{B}_s(G(\alpha), \beta) + \widetilde{B}_p(G(\alpha), \theta), \tag{7}$$

$$\widetilde{B}_s(G(\alpha), \beta) = w(G(\alpha))B_s(\beta), \tag{8}$$

$$\widetilde{B}_p(G(\alpha), \theta) = w(G(\alpha))B_p(\theta), \tag{9}$$

where $G(\alpha)$ is the T-pose garment obtained by Eq. 4. $\widetilde{B}_s(\cdot) \in \mathbb{R}^{N_G \times 3}$ and $\widetilde{B}_p(\cdot) \in \mathbb{R}^{N_G \times 3}$ are the corresponding garment displacements influenced by the BSBS and BPBS of the body, respectively. $w(\cdot) \in \mathbb{R}^{N_G \times N_B}$ is a weighted matrix that can be computed by searching the K-Nearest Neighbors (KNN) body vertices for each garment vertex [Jiang et al. 2020; Peng et al. 2021].

Posed Coarse Garment Modeling. Then we transfer SMPL's LBS to $T_G(\cdot)$ to obtain the posed garment $M_P(\cdot)$ by

$$M_P(\alpha, \beta, \theta) = W\left(T_G(\alpha, \beta, \theta), J(\beta), \theta, \widetilde{\mathcal{W}}\right), \tag{10}$$

$$\widetilde{\mathcal{W}} = w(G(\alpha))\mathcal{W}, \tag{11}$$

where $\widetilde{\mathcal{W}} \in \mathbb{R}^{N_G \times 24}$ represents the garment skinning weights extended from SMPL with the same weighted matrix $w(\cdot)$ in Eq. 9.

4.2 Fine Implicit Garment Reconstruction

To generate the fine implicit garment field, we first obtain the garment mask and the normal map of the input image (*Please refer to the supplementary materials for details about mask extraction and our normal estimator*). Then we apply an Hourglass filter [Saito et al. 2019] to extract the image feature from the input normal map. The 3D point p is projected to 2D image coordinate by camera projection $\pi(\cdot)$ to exact pixel-aligned local image feature $I_F(p) = F(\pi(p))$. Then we define an implicit function f for arbitrary point p in 3D space as,

$$f(F(\pi(p)), z(p)) = s : s \in (0, 1), \tag{12}$$

where s denotes the occupancy of p, and $z(p)$ is the depth in the camera coordinate space. $f(\cdot)$ is designed as MLPs to decode the occupancy status of p. L_1 loss is chosen to measure the error between the predicted occupancy and the ground truth during training. To compute the occupancy, we use MeshLab's close hole operation [Cignoni et al. 2008] to create a closed topology for each garment.

4.3 Geometry-aware Boundary Prediction

Garment boundaries are thin 3D curves that are challenging to capture with implicit functions. Inspired by ReEF [Zhu et al. 2022], we adopt a cylinder structure to represent the garment boundary. To obtain garment boundaries, a straightforward approach is to use 2D image cues to regress the boundary cylinders. However, relying solely on 2D pixel-aligned features suffers from depth ambiguity, leading to inconsistent 3D results. To address the ambiguity, we integrate 2D clues and 3D geometry-aligned features to enhance global boundary shape alignment (see Fig. 4). We utilize the previous pixel-aligned image feature $F(\pi(p))$ as 2D clues. To produce geometry-aware features from the fine garment M_I, we employ a triplane encoder ψ_{enc} to obtain 3D features aligned with three axis-aligned orthogonal planes. Specifically, point clouds sampled from M_I are projected onto the triplane, and a 3D-aware UNet ψ_{enc} is used to obtain high-level triplane feature maps [Liu et al. 2024]. Then we query any 3D position $p \in \mathbb{R}^3$ by projecting it onto each feature plane, retrieving three corresponding feature vectors $G_F(p) = (F_{xy}, F_{xz}, F_{yz})$ via bilinear interpolation. A small MLP-based decoder ψ_{dec} is used to interpret the aggregated concatenated 2D pixel-aligned and 3D triplane features as 3D boundary fields. For an arbitrary 3D point p, its occupancy value o_i to the i-th boundary is computed as,

$$f_i(F(\pi(p)), F_{xy}, F_{xz}, F_{yz}) = o_i : o_i \in (0, 1), \tag{13}$$

where we model each garment boundary as an implicit cylinder to compute the ground-truth occupancy. L_1 loss is used to measure the error between the predicted occupancy and the ground truth.

4.4 3D Garment Shape Registration

To obtain the target garment mesh M_F, we first establish the boundary correspondence between the boundaries of the coarse garment and the predicted boundary cylinders for registration, as the boundaries possess prominent geometrical features of the garment shape. We fit the coarse mesh boundary strip to the predicted 3D boundary cylinders by minimizing the objective function:

$$L_{boundary} = \lambda_c L_c + \lambda_{lap} L_{lap} + \lambda_{edge} L_{edge} + \lambda_{normal} L_{normal}, \tag{14}$$

where L_c is the Chamfer loss [Ravi et al. 2020] that restricts the positions of boundary mesh vertices; L_{lap}, L_{edge} and L_{normal} are Laplacian Smooth, Edge Length and Normal Consistency regularizers [Ravi et al. 2020], respectively. After fitting the coarse mesh boundary strip, the deformed boundary strip is generated to guide the registration of the target garment mesh. Then, we utilize non-rigid ICP [Amberg et al. 2007] to register the coarse garment template to the target garment mesh under the constraints of:

$$L_{nicp} = \lambda_d L_d + \lambda_b L_b + \lambda_s L_s + \lambda_{reg} L_{reg}, \tag{15}$$

$$L_{reg} = \lambda_{lap} L_{lap} + \lambda_{edge} L_{edge} + \lambda_{normal} L_{normal}, \tag{16}$$

where L_d penalizes the distance between the deformed garment template and the ground truth. L_b is the landmark cost between the coarse mesh boundary strips and the deformed boundary strips, while the stiffness term L_s penalizes differences between the transformation matrices assigned to neighboring vertices. Different from the original non-rigid ICP, we incorporate the mesh regularization term L_{reg} [Ravi et al. 2020] to stabilize the registration process.

Table 1. **Quantitive comparison between our method with others**.

Method	BCNet	ClothWild	Deep Fashion3D	ReEF	Ours
Chamfer Distance ↓	18.742	16.136	17.159	11.357	**7.825**
Normal Consistency ↑	0.781	0.812	0.793	0.838	**0.913**

Table 2. **Quantitative comparison between our method and alternative strategies for predicting garment boundary**.

Method	Chamfer Distance ↓	Normal Consistency ↑	IoU ↑
ReEF	16.428	0.809	55.425
Ours	**10.571**	**0.862**	**69.775**

5 EXPERIMENTS

In our experiments, we trained our method and all compared methods using our synthetic dataset (as shown in Fig. 2 and Fig. 3), allocating 80% for training and reserving the remaining 20% for testing. To evaluate reconstruction quality, we employ commonly used metrics [Mescheder et al. 2019] for quantitative comparisons, including Chamfer Distance, Normal Consistency, and Intersection over Union (IoU). Quantitative comparisons were conducted on the test-set of our synthetic data, while qualitative comparisons were performed on in-the-wild images. Fig. 1 and Fig. 5 present our representative results. Please refer to our supplementary materials for more results and implementation details.

Comparison Study. We compare our method with the state-of-the-art single-view garment reconstruction methods, i.e., BCNet, ClothWild, DeepFashion3D and ReEF, both quantitatively and qualitatively. Tab. 1 shows the quantitative comparisons. Our proposed method achieves the best scores against the baselines. Fig. 6 provides qualitative comparisons given the input of in-the-wild images. LBS-based methods (e.g., BCNet, ClothWild) only capture coarse deformations caused by pose and often neglect surface details. Although BCNet utilizes a displacement network to model garment deformations and geometric details, it is challenging for a simple

Table 3. **Quantitative comparison between our method and alternative strategies**.

Method	*Ablation Study on*					Ours
	Data	*Coarse Garment Estimation*	*Implicit Representation*			
	ReEF's dataset	Crop from SMPL	UDF w/o Registering	UDF w/ Registering	Occupancy w/o Registering	
Chamfer Distance ↓	16.363	14.635	9.616	9.375	8.658	**7.825**
Normal Consistency ↑	0.805	0.823	0.841	0.848	0.851	**0.913**

MLP-based network to regress a larger number of vertex offsets. Using only global features makes it inefficient for DeepFashion3D to obtain accurate boundaries, resulting in poor reconstruction quality. Although ReEF performs well on simple poses and simple clothing deformations by leveraging pixel-aligned features, it presents artifacts on garments with complex human poses and garment deformations. Our method demonstrates proficiency in capturing both large garment deformations and geometric details.

Ablation Study on Boundary Prediction. Tab. 2 shows the quantitative comparisons between ReEF and our method on boundary field prediction. Our proposed method achieves better scores. Fig. 7 provides qualitative comparisons on garment boundary reconstruction. As noted, relying solely on 2D pixel-aligned features makes ReEF fail to predict accurate boundaries with complex poses and deformations, resulting in discontinuous boundaries. Our geometry-aware boundary prediction excels in reconstructing complex garment boundaries that are well-aligned with the garment shape.

Ablation Study on Data. We verify the significance of our data by training our method on both: 1) ReEF's dataset; and 2) our GarVerseLOD. As shown in Fig. 8 and Tab. 3, the model trained with our data achieves the best results, indicating that our data enhances the network's generalization in reconstructing in-the-wild images.

Ablation Study on Coarse Garment Estimation. To demonstrate the significance of using our dataset in building the garment parametric model, we conduct an ablation study on different methods to obtain coarse garments. Apart from using our parametric model and estimator, there is an alternative strategy: cropping a part of the mesh from a posed SMPL body, as used in DeepFashion3D and ReEF. Fig. 9 and Tab. 3 present the comparisons between our method and the ablated strategies. Our method is superior in estimating a more reasonable coarse garment. The registered results show that a good coarse initialization significantly stabilizes the registration process.

Ablation Study on Implicit Representation. Apart from registering coarse garments to fine garments, another strategy for obtaining open-boundary meshes is to use UDF (Unsigned Distance Field). However, UDF encounters two problems (Fig. 10, Tab. 3): 1) Although some methods [Guillard et al. 2022] can extract open-boundary meshes from UDF, the quality is poor and may result in unexpected open regions and incomplete meshes. Garment registration is still required to achieve fine topology. 2) The regression problem with UDF is more challenging to converge than classification, resulting in inferior surface details compared to the occupancy field.

6 CONCLUSION

Capturing diversified garment shapes and intricate garment deformations robustly from single RGB images remains difficult due to garment complexity and data scarcity. Our work presents a large-scale 3D garment dataset GarVerseLOD, which is extensively annotated at different levels of detail, ranging from coarse stylized garments to deformed models with intricate deformations and fine-grained geometric details. Based on the well-established dataset, we propose a framework for high-quality 3D garment reconstruction from single-view images. The core of our approach is a hierarchical design to recover different levels of garment details, i.e., from pose-independent stylized coarse garments to pose-blended and open-boundary garments with pixel-aligned details. Experiments indicate that our framework is capable of reconstructing garments with various shapes and fine-grained deformations, showcasing its superior generalization ability against state-of-the-art methods.

Limitation. Although our work provides faithful reconstructed results on a wide range of in-the-wild images, it may fail when reconstructing garments with complex topology: 1) As shown in Fig. 11(a), although our method is able to reconstruct faithful clothing details, it fails to represent the multi-layer structures present in dresses or skirts. This problem is largely due to the reliance on the single-layer occupancy field and the single-layer garment parametric model, which are unable to capture multi-layered structures. One possible solution is to design a new representation that effectively supports the reconstruction of garments with multi-layer structures; 2) As shown in Fig. 11(b), our method struggles to accurately reconstruct dresses or skirts with slits. This issue primarily stems from the limited representation of such features in our current dataset. The lack of sufficient examples of slits in the training data restricts the model's ability to generalize and accurately reconstruct these specific structures. A potential strategy is to expand the dataset by incorporating a broader range of clothing styless, thereby enhancing the model's capability to handle these intricate features.

ACKNOWLEDGMENTS

The work was supported in part by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, by Guangdong Provincial Outstanding Youth Fund (No. 2023B1515020055), by Shenzhen Science and Technology Program No. JCYJ20220530143604010, and No. NSFC61931024. It is also partly supported by the National Key R&D Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104, by Key Area R&D Program of Guangdong Province (Grant No. 2018B030338001), by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), and by Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055). Stable Diffusion [Rombach et al. 2022] was utilized to generate the synthetic images used in this work.

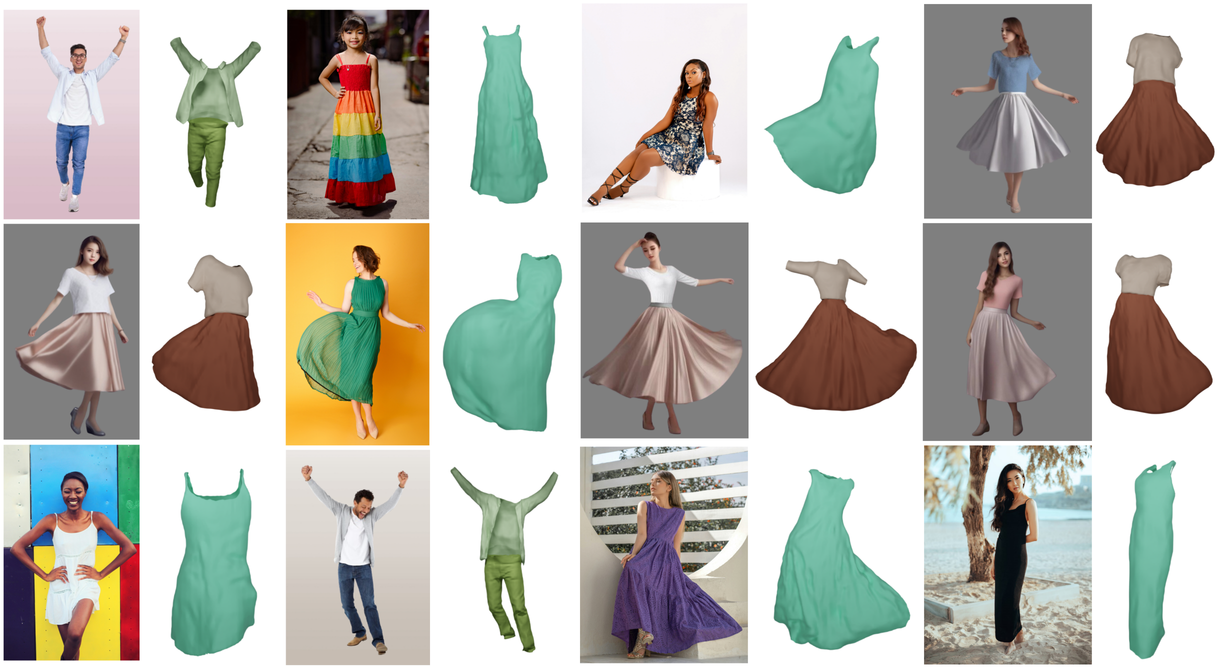

Fig. 5. Result gallery of our method. Each image is followed by the reconstructed garment mesh. As illustrated, our method can effectively reconstruct garments with intricate deformations and fine-grained surface details. To support the modeling of folded structures, such as collars, we assembled a repository of diverse real-world collars that were crafted based on our topologically-consistent garments. A lightweight classification network was trained to select the collar that best matches the given image in terms of appearance [Zhu et al. 2022]. Original images courtesy of licensed photos and Stable Diffusion. The images with a gray background are synthesized, while the rest are licensed photos.

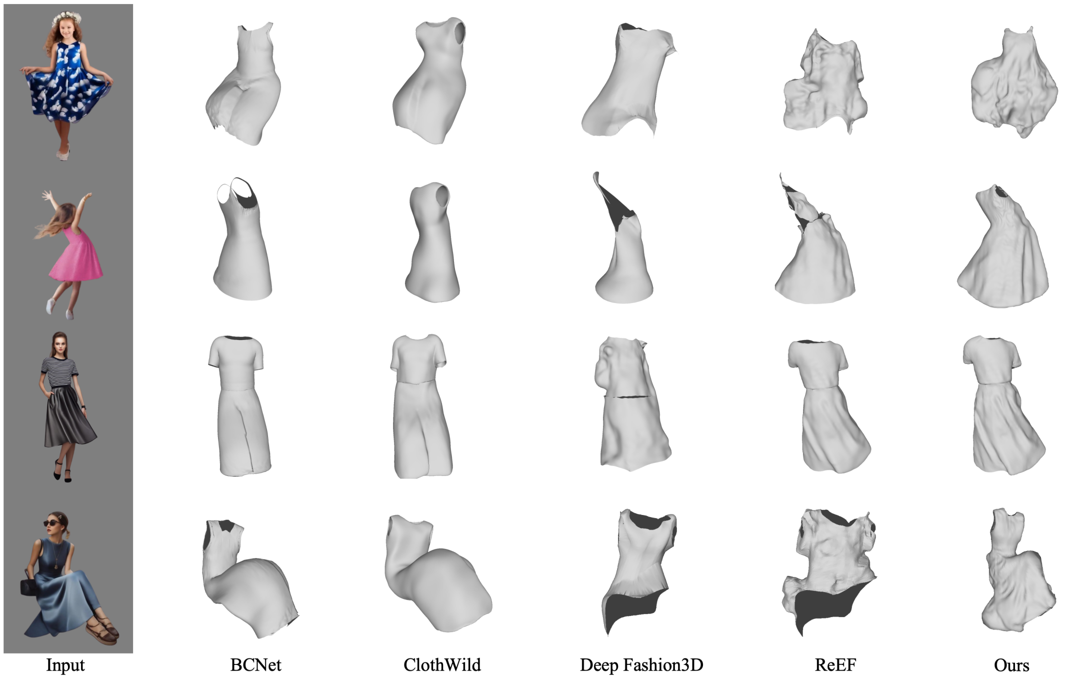

Fig. 6. Qualitative comparison between ours and the state of the arts. For each row, the input image is followed by the results generated by BCNet [Jiang et al. 2020], ClothWild [Moon et al. 2022], Deep Fashion3D [Zhu et al. 2020], ReEF [Zhu et al. 2022] and our method. Input images courtesy of Stable Diffusion.

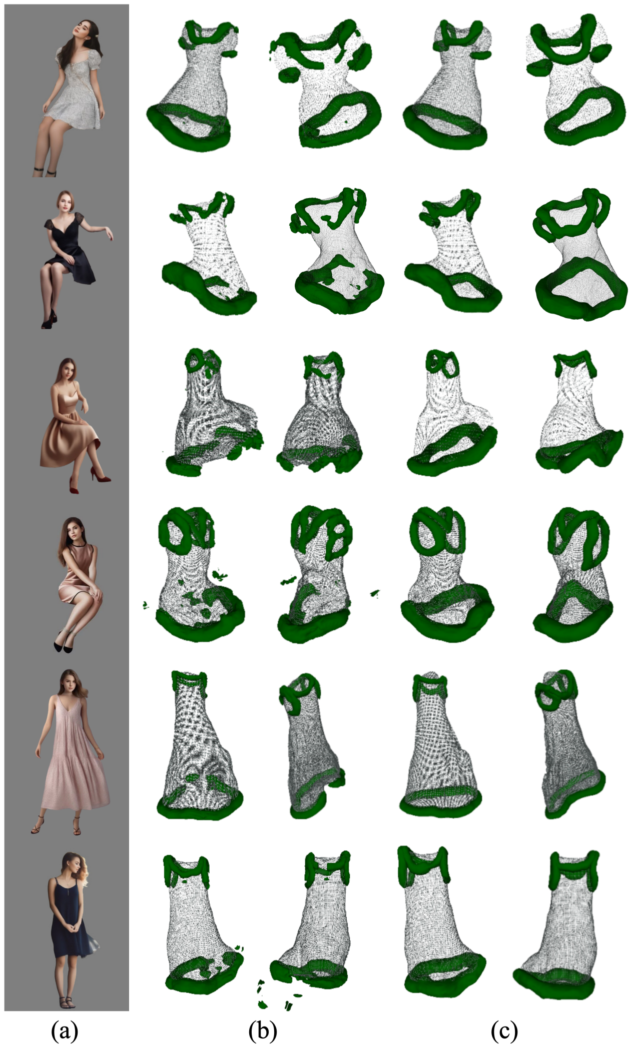

Fig. 7. Qualitative comparison between our method and the alternative strategy for predicting garment boundary from in-the-wild images. The input image (a) is followed by the boundaries generated by (b) ReEF's strategy and (c) our geometry-aware estimator. ReEF fails to accurately predict boundaries with complex poses and deformations, leading to discontinuous boundaries. Our geometry-aware boundary prediction outperforms ReEF in reconstructing complex garment boundaries that are well-aligned with the garment shape. Input images courtesy of Stable Diffusion.

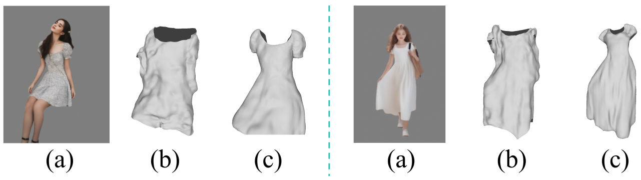

Fig. 8. Qualitative comparison on different data. The input image (a) is followed by the results generated by networks trained with (b) ReEF's data and (c) our GarVerseLOD. Input images courtesy of Stable Diffusion.

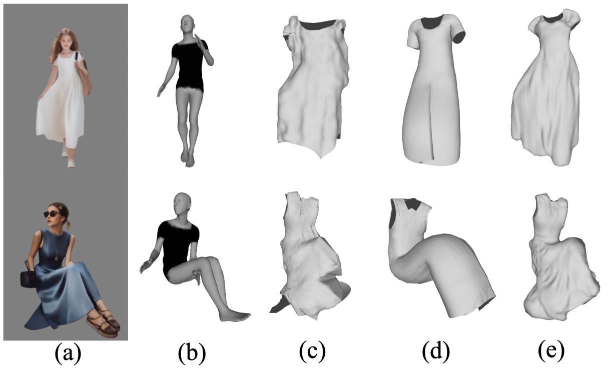

Fig. 9. Qualitative comparison between our method and the alternative strategy for obtaining coarse garment template. (a) the input image; (b) the template (black part) cropped from SMPL; (c) the registration result using (b); (d) the coarse garment estimated by our coarse garment estimator; and (e) the registration result using (d). Input images courtesy of Stable Diffusion.

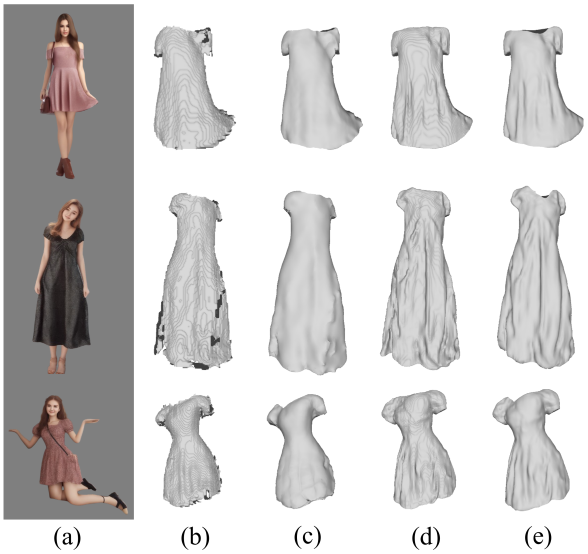

Fig. 10. Qualitative comparison on different representation. The input image (a) is followed by the result generated by (b) UDF, (c) registering to (b), (d) occupancy field and (e) registering to (d). Input images courtesy of Stable Diffusion.

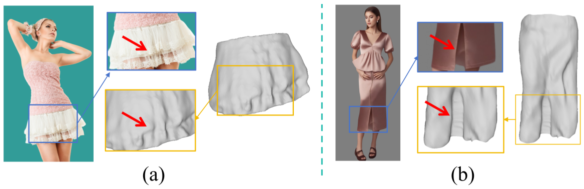

Fig. 11. Failure cases. Our framework may struggle to reconstruct garments with complex topology, such as those multi-layered structures (a) or featuring slits (b). Images courtesy of licensed photos and Stable Diffusion.

REFERENCES

Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. 2019. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2293–2303.

Brian Amberg, Sami Romdhani, and Thomas Vetter. 2007. Optimal step nonrigid ICP algorithms for surface registration. In *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 1–8.

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

Hugo Bertiche, Meysam Madadi, and Sergio Escalera. 2020. CLOTH3D: clothed 3d humans. In *European Conference on Computer Vision*. Springer, 344–359.

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

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

Honghu Chen, Yuxin Yao, and Juyong Zhang. 2024. Neural-ABC: Neural Parametric Models for Articulated Body With Clothes. *IEEE Transactions on Visualization and Computer Graphics* (2024).

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

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

Benoit Guillard, Federico Stella, and Pascal Fua. 2022. MeshUDF: Fast and Differentiable Meshing of Unsigned Distance Field Networks. In *European Conference on Computer Vision*.

Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. 2019. Garnet: A two-stream network for fast and accurate 3d cloth draping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8739–8748.

Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. 2019. Livecap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics (TOG)* 38, 2 (2019), 1–17.

Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. 2020. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5052–5063.

Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and H-P Seidel. 2009. A statistical model of human pose and body shape. In *Computer graphics forum*, Vol. 28. Wiley Online Library, 337–346.

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

Yue Jiang, Marc Habermann, Vladislav Golyanik, and Christian Theobalt. 2022. Hifecap: Monocular high-fidelity and expressive capture of human performances. *arXiv preprint arXiv:2210.05665* (2022).

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 694–711.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment Anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), 3992–4003. https://api.semanticscholar.org/CorpusID:257952310

Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. 2017. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6050–6059.

Ren Li, Corentin Dumery, Benoît Guillard, and Pascal Fua. 2024a. Garment Recovery with Shape and Deformation Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1586–1595.

Ren Li, Benoît Guillard, and Pascal Fua. 2024b. Isp: Multi-layered garment draping with implicit sewing patterns. *Advances in Neural Information Processing Systems* 36 (2024).

Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. 2020. Monocular real-time volumetric performance capture. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*. Springer, 49–67.

Yue Li, Marc Habermann, Bernhard Thomaszewski, Stelian Coros, Thabo Beeler, and Christian Theobalt. 2021. Deep physics-aware inference of cloth deformation for monocular human performance capture. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 373–384.

Siyou Lin, Boyao Zhou, Zerong Zheng, Hongwen Zhang, and Yebin Liu. 2023. Leveraging intrinsic properties for non-rigid garment alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14485–14496.

Haolin Liu, Chongjie Ye, Yinyu Nie, Yingfan He, and Xiaoguang Han. 2024. LASA: Instance Reconstruction from Real Scans using A Large-scale Aligned Shape Annotation Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20454–20464.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

Zhongjin Luo, Shengcai Cai, Jinguo Dong, Ruibo Ming, Liangdong Qiu, Xiaohang Zhan, and Xiaoguang Han. 2023. RaBit: Parametric Modeling of 3D Biped Cartoon Characters with a Topological-consistent Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12825–12835.

Zhongjin Luo, Jie Zhou, Heming Zhu, Dong Du, Xiaoguang Han, and Hongbo Fu. 2021. Simpmodeling: Sketching implicit field to guide mesh modeling for 3d animalmorphic head design. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 854–863.

Sébastien Marcel and Yann Rodriguez. 2010. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*. 1485–1488.

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

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

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453* (2023).

Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. 2019. Siclope: Silhouette-based clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4480–4490.

Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. 2020. TailorNet: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10975–10985.

Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2021. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14314–14323.

Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. 2017. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–15.

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

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. 2020. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501* (2020).

RenderPeople. 2018. In *https://renderpeople.com/3d-people*.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 234–241.

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

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 84–93.

Feitong Tan, Hao Zhu, Zhaopeng Cui, Siyu Zhu, Marc Pollefeys, and Ping Tan. 2020. Self-supervised human depth estimation from monocular videos. In *Proceedings of* *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 650–659.

Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. 2020. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 1–18.

Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 2024. 4D-DRESS: A 4D Dataset of Real-World Human Clothing With Semantic Annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 550–560.

Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. 2020. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *2020 International Conference on 3D Vision (3DV)*. IEEE, 322–332.

Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. 2023. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 512–523.

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

Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. 2018. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (ToG)* 37, 2 (2018), 1–15.

Yuanlu Xu, Song-Chun Zhu, and Tony Tung. 2019. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7760–7770.

Zizheng Yan, Jiapeng Zhou, Fanpeng Meng, Yushuang Wu, Lingteng Qiu, Zisheng Ye, Shuguang Cui, Guanying Chen, and Xiaoguang Han. 2024. DreamDissector: Learning Disentangled Text-to-3D Generation from 2D Diffusion Priors. *arXiv preprint arXiv:2407.16260* (2024).

Shan Yang, Zherong Pan, Tanya Amert, Ke Wang, Licheng Yu, Tamara Berg, and Ming C Lin. 2018. Physics-inspired garment recovery from a single-view image. *ACM Transactions on Graphics (TOG)* 37, 5 (2018), 1–14.

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. 2021. Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*.

Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. 2017. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4191–4200.

Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. 2023b. PyMAF-X: Towards Well-aligned Full-body Model Regression from Monocular Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. 2021. PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop. In *Proceedings of the IEEE International Conference on Computer Vision*.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023a. Adding Conditional Control to Text-to-Image Diffusion Models.

Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. 2019. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7739–7749.

Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. 2020. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 512–530.

Heming Zhu, Lingteng Qiu, Yuda Qiu, and Xiaoguang Han. 2022. 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*. 3845–3854.

Xingxing Zou, Xintong Han, and Waikeung Wong. 2023. CLOTH4D: A Dataset for Clothed Human Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12847–12857.

A MORE RESULTS AND IMPLEMENTATION

More Results. We report more results on challenging loose cloth reconstruction in Fig. 14, Fig. 15, Fig. 16 and Fig. 17.

Implementation details. In our implementation, all networks were implemented using PyTorch and trained on an Ubuntu server equipped with four A100 GPUs. Quantitative evaluations and qualitative assessments were also performed on this server. For the coarse garment estimator, the parameter size of our statistical garment model $G(\alpha)$ is 32 (i.e., the length of α is 32). We use ResNet-50 blocks [Luo et al. 2023; Marcel and Rodriguez 2010] to map the input image ($512 \times 512 \times 3$) to α and obtain the T-posed coarse garment through Eq. 4 in the Main Paper. All the data in the Garment Style Database were used to learn $G(\alpha)$. To obtain a powerful estimator, we collected 10,000 image-3D T-pose garment paired data (as shown in Fig. 3(b) and Fig. 3(f) of the Main Paper) for training. The coarse garment estimator is trained for 1000 epochs using the Adam optimizer with a batch size of 128 and a learning rate of 3×10^{-4}. For the SMPL body estimator, we employ the well-established body estimator PyMAF [Zhang et al. 2023b, 2021] to predict body shape and pose.

Inspired by the normal-aided approaches [Saito et al. 2020; Xiu et al. 2022], we employ the normal map and garment segmentation mask as inputs to accurately carve the garment surface details. In the training stage, the ground-truth normal maps and garment masks of our synthetic data are directly rendered from the 3D garment models of GarVerseLOD. Given the ground-truth normal map ($512 \times 512 \times 3$) and the corresponding collected images ($512 \times 512 \times 3$), we train a normal estimator using a U-Net [Ronneberger et al. 2015] network with the following loss:

$$L_N = L_{pixel}[M] + \lambda_{VGG} L_{VGG}[M], \qquad (17)$$

where L_{pixel} is a L_1 Loss between the ground-truth and predicted normal maps and L_{VGG} is a perceptual loss [Johnson et al. 2016] weighted by λ_{VGG}. M is the pixel-aligned mask as shown in Fig. 3(c) of the Main Paper. We collect 5,000 paired data (as illustrated in Fig. 3(b,c,d) of the Main Paper to train the normal estimator. To improve the performance of the normal estimator, we also incorporate data from THUman2 [Yu et al. 2021]. The normal estimator is trained for 80 epochs using the Adam optimizer with a batch size of 32 and a learning rate of 3×10^{-4}. In the testing stage, the garment masks of in-the-wild images are generated by leveraging the segmentation of SAM [Kirillov et al. 2023], and the normal maps are predicted by our trained estimator.

For the fine garment estimator and the geometry-aware boundary predictor, the input size of the normal map and the garment mask is $512 \times 512 \times 3$. Inspired by [Saito et al. 2019], we utilize an Hourglass filter to extract image features and employ an MLP network to decode the features of each sampled point into an occupancy value. The fine garment estimator undergoes 100 epochs of training with the RMSprop optimizer, utilizing a batch size of 16 and a start-up learning rate of 1×10^{-3}. The learning rate is reduced by a factor of 10 after epochs 30, 60, and 90. The geometry-aware triplane features are set to a resolution of 256×256. The boundary estimator is trained over 100 epochs using the RMSprop optimizer, with a batch size of 12 and a starting learning rate of 1×10^{-3}. The learning rate is also reduced by a factor of 10 following epochs 30, 60, and 90. We construct 5,000 pairs of data for each category using the data synthesis strategy shown in Fig. 2 and Fig. 3 of the Main Paper to train our fine garment estimator and the geometry-aware boundary predictor. Although our experiments were conducted with 5,000 pairs of data, it is important to note that our strategy is capable of synthesizing larger-scale datasets, given sufficient computing resources. For garment shape registration, it is generally better to use different weight schedulers to optimize various types of clothing. Please refer to our code for details.

B DETAILS OF GARVERSELOD

GarVerseLOD spans a wide range of 3D garment models, containing 5 common garment categories, i.e., dress, skirt, coat, top, and pant. Tab. 4 shows the statistical data for each garment category. The total size represents the number of garments created by artists, not the size of garments our strategy can synthesize. We have recruited eight professional artists to create corresponding 3D garments using Blender based on the collected reference images. The topological consistency of garments enables us to generate new samples by interpolating between two garments within each database. Theoretically, we can generate more garments than the product of the sizes of the individual databases. All eight artists are required to craft 3D models by deforming the predefined template mesh. Each artist possesses over five years of modeling experience, and on average, each garment takes around average 25 minutes to complete. Fig. 12 illustrates the predefined template meshes for each category. Fig. 18, Fig. 19, Fig. 20 and Fig. 21 respectively illustrate our four datasets: 1) Garment Style Database; 2) Local Detail Database; 3) Garment Deformation Database and 4) Fine Garment Dataset.

Table 4. Data statistics for each basic database. The total size refers to the number of garments crafted by artists.

Category	Dress	Coat	Skirt	Top	Pant
Garment Style Database	863	760	538	350	358
Local Detail Database	86	62	55	38	36
Garment Deformation Database	622	605	456	582	589
Total	1,571	1,427	1,049	970	983

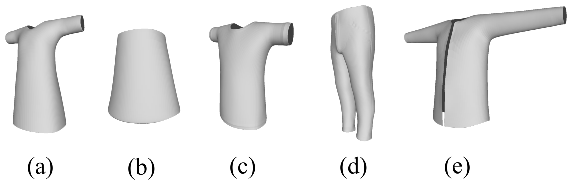

Fig. 12. Predefined templates for each garment category, including (a) dress, (b) skirt, (c) top, (d) pant, and (e) coat.

Garment Deformation Crafting. As shown in Fig. 13, we first collect real images of clothed humans with diverse poses, garment styles, and deformations from the Internet, covering the 5 garment

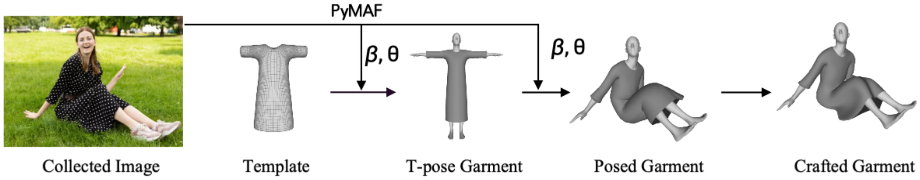

Fig. 13. Given a "Collected Image", we utilize PyMAF [Zhang et al. 2023b, 2021] to estimate SMPL body. Eight artists are then tasked with creating "T-pose Garment" shapes by deforming a predefined "Template" to match the T-pose body predicted by PyMAF. Then the SMPL's Linear Blend Skinning (LBS) is extended to the T-pose garment to obtain the "Posed Garment". Finally, the artists are further instructed to refine the posed garment to get the "Crafted Garment" while ensuring that garment deformations closely match the collected images. "Posed Garment" represent the shape of clothing influenced by human pose, while "Crafted Garment" capture the state of garments affected by various complex factors—not only pose but also other environmental influences, such as garment-environment interactions and external forces like wind.

categories. Secondly, we employ PyMAF [Zhang et al. 2021] to estimate the human shape β and pose θ from the images and discard the inaccurate estimation results manually. Thirdly, we recruited 8 artists to construct 3D clothing models manually to match the reference images as much as possible, following the specific procedure below: 1) The artists are required to create the coarse **T-pose Garments** according to the Collected Images, by deforming the predefined category-specific templates to match T-pose SMPL meshes generated by PyMAF; 2) Then, the SMPL's Linear Blend Skinning (LBS) is extended to the T-pose garments programmatically to capture garment deformations resulting from human poses, obtaining the **Posed Garments**; 3) Finally, the artists are asked to deform the Posed Garments to create the final **Crafted Garments**, ensuring that deformations match the collected images as closely as possible. In-the-wild images naturally capture the complex real-world physical conditions that occur in a single snapshot. By basing manual modeling on reference images, our data encompass diverse clothing-states observed in real-world scenarios. Note that **Posed Garments** represent the shape of garments after being affected by human pose, while **Deformed Garments** (i.e, Crafted Garments) capture the state of garments affected by complex factors (not only affected by pose, but also by other complex environmental factors, such as garment-environment interactions and external forces like wind).

Notation table. Tab. 5 provides a summary of the notations used in the Main Paper.

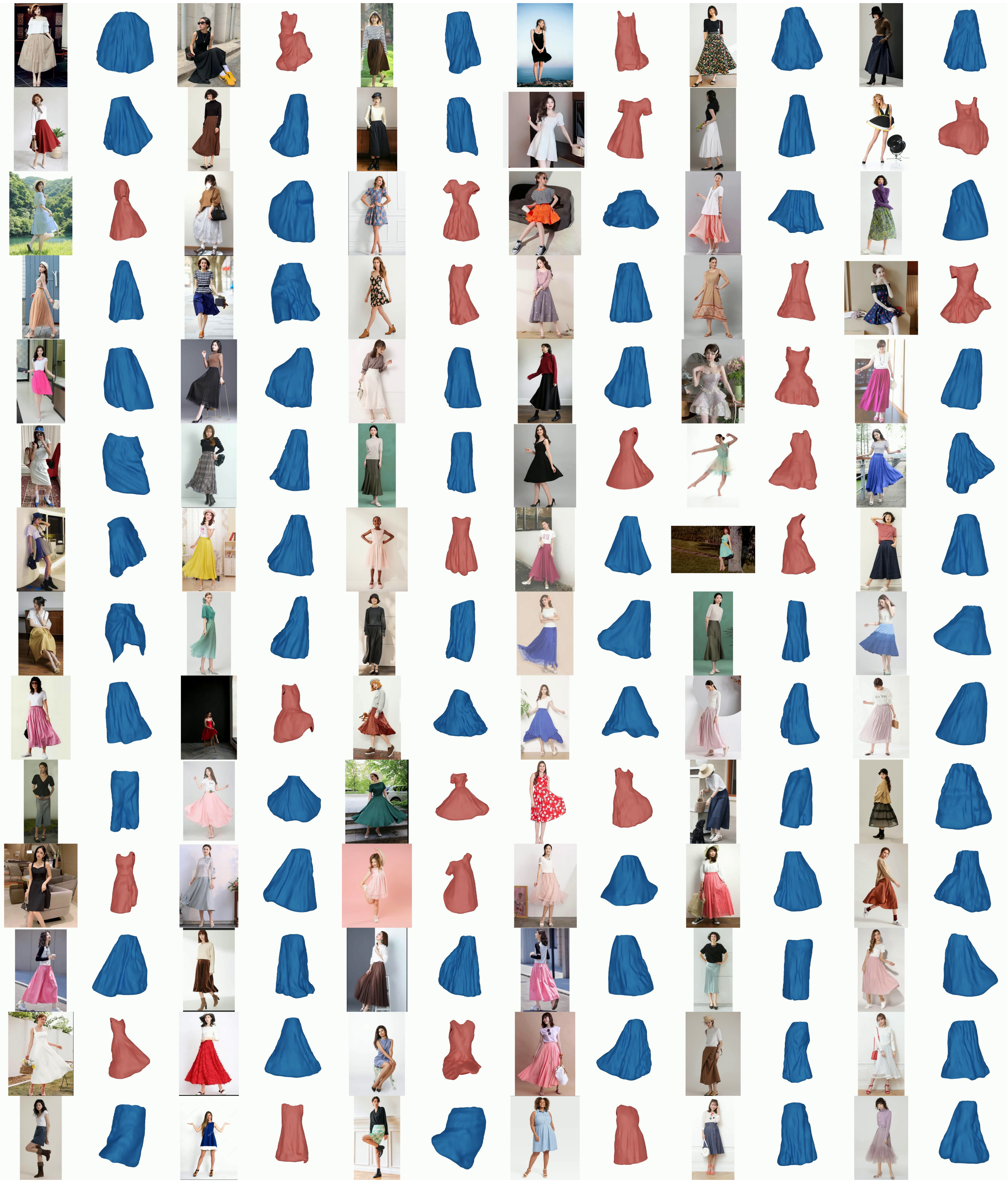

Fig. 14. More Results on Loose-fitting Garments.

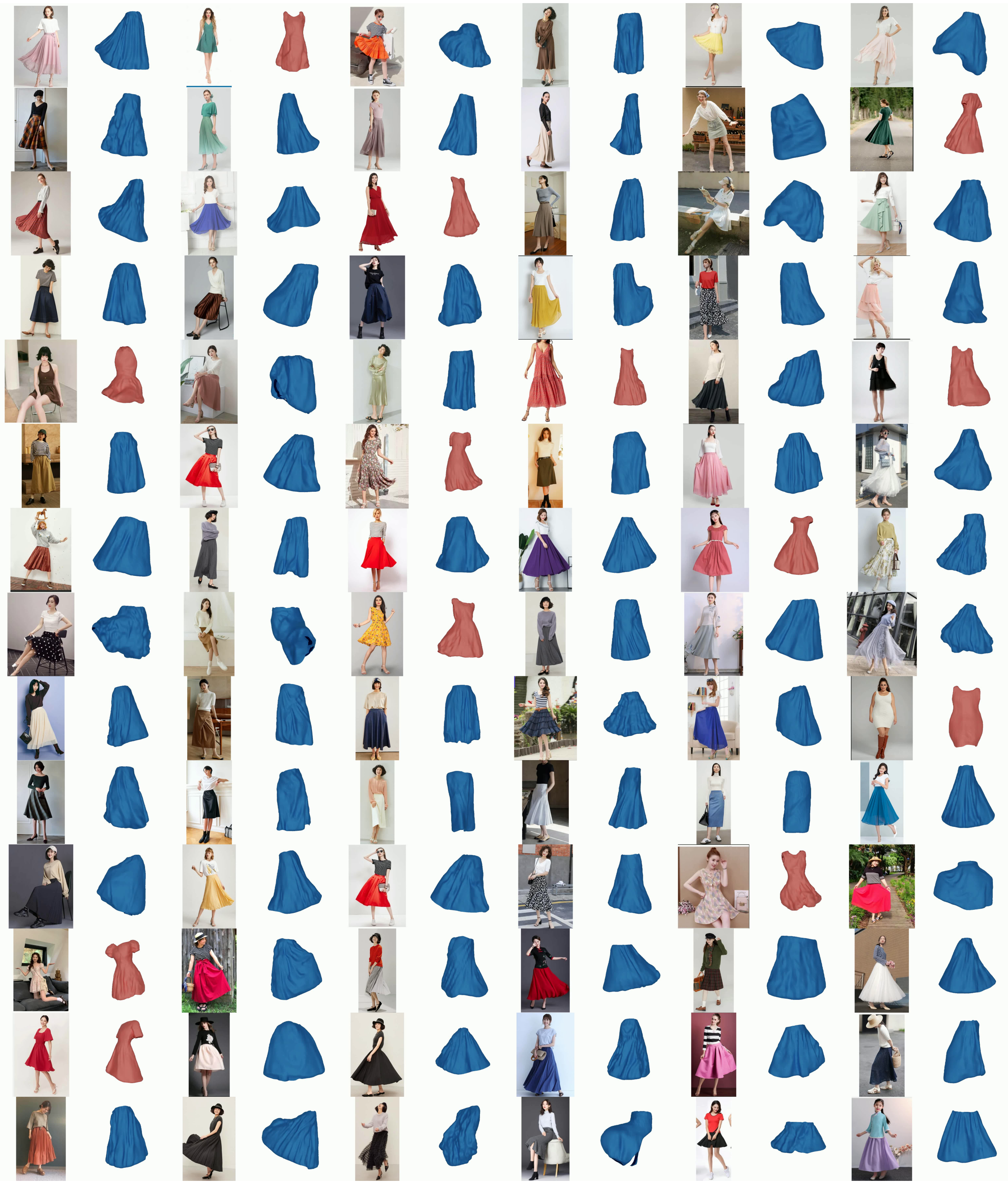

Fig. 15. More Results on Loose-fitting Garments.

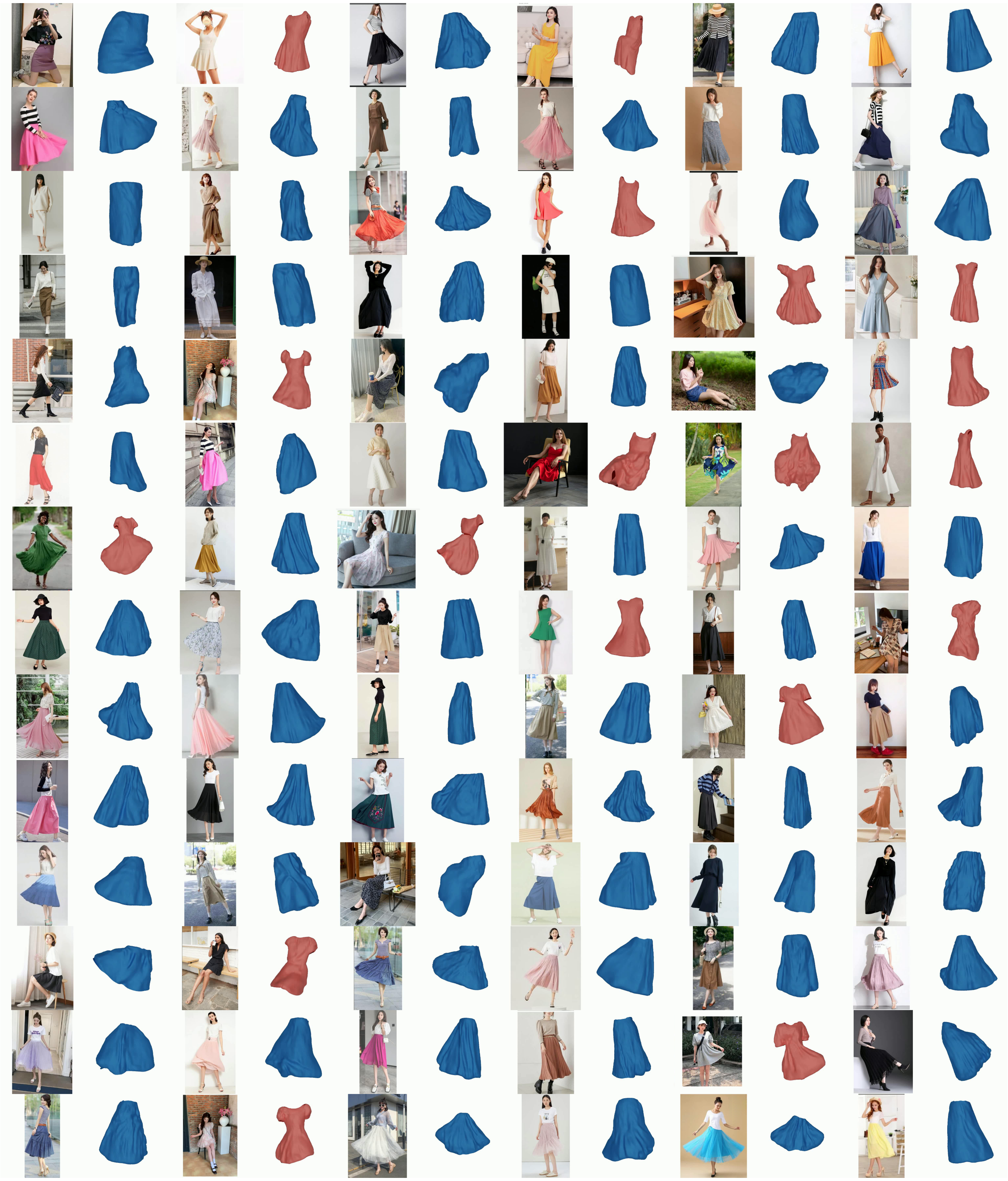

Fig. 16. More Results on Loose-fitting Garments.

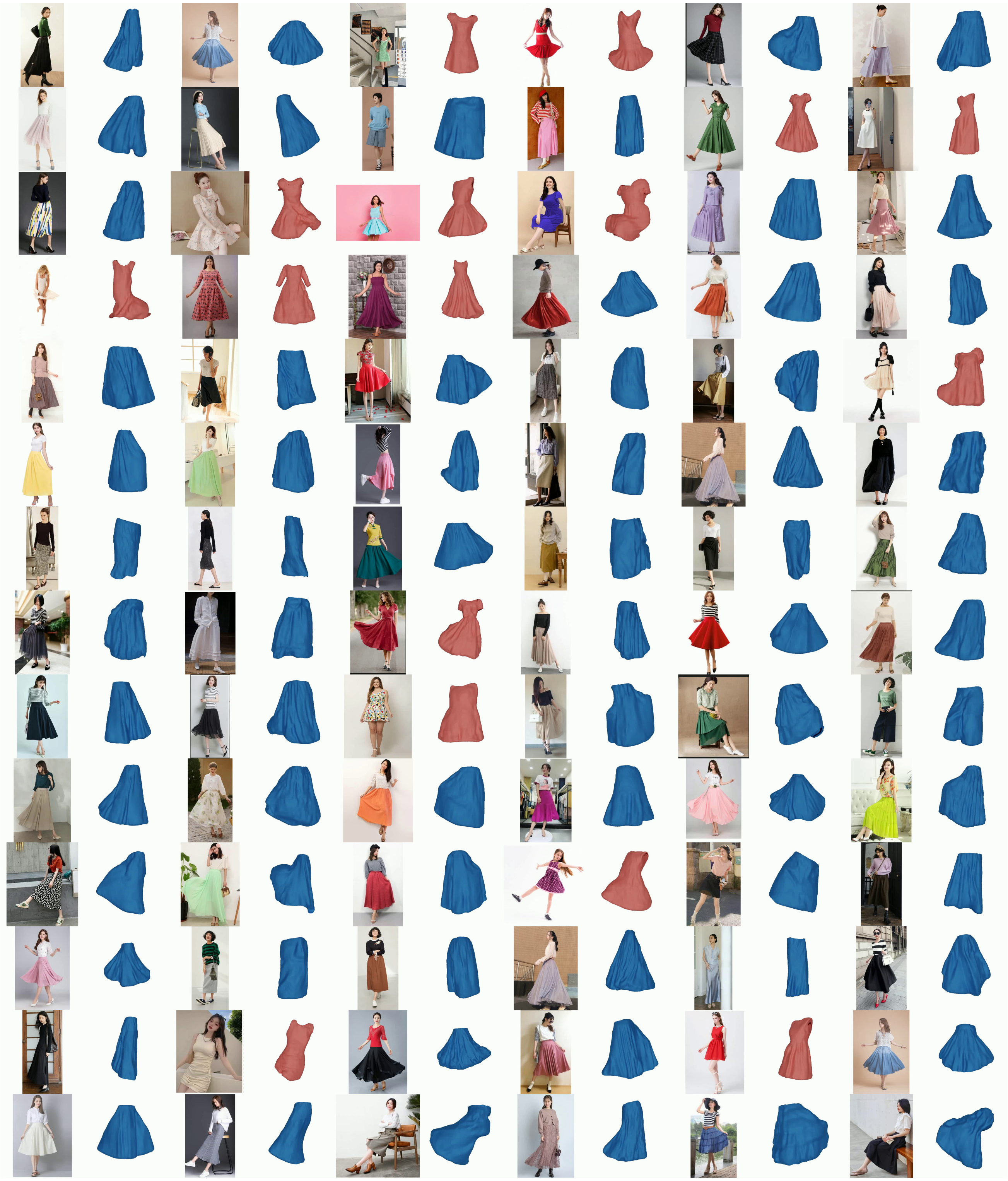

Fig. 17. More Results on Loose-fitting Garments.

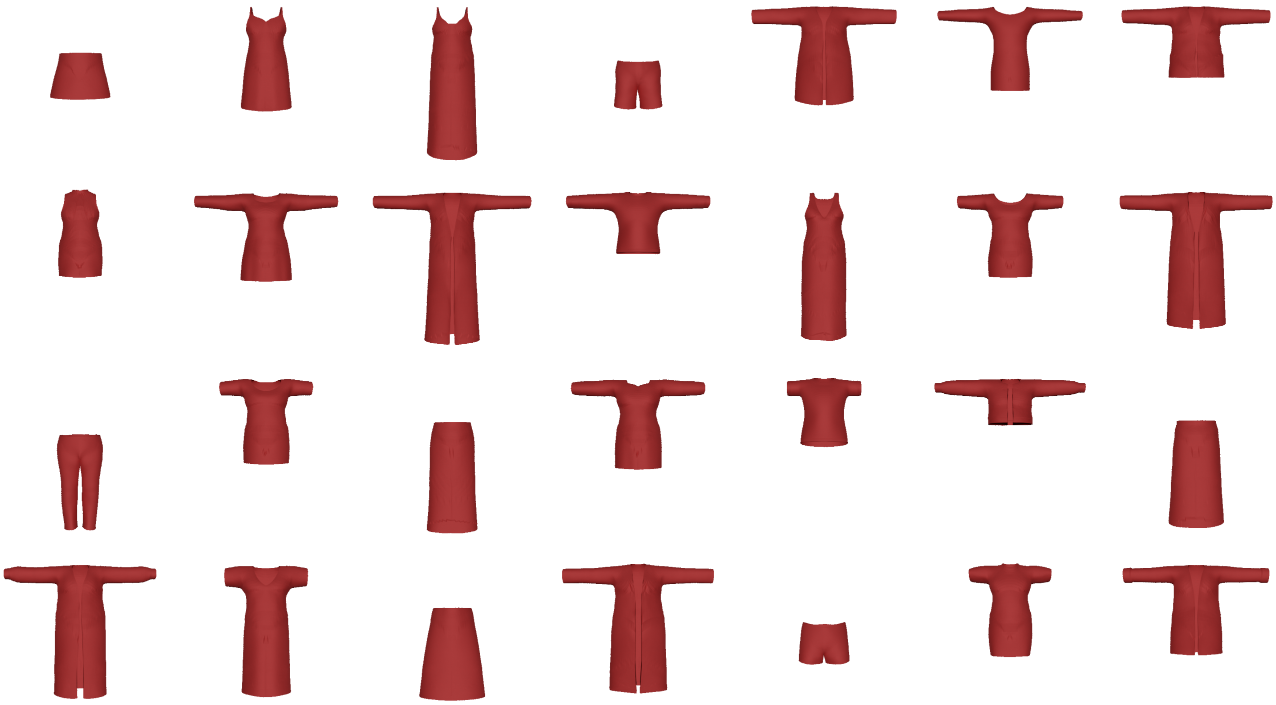

Fig. 18. An illustration of our **Garment Style Database**.

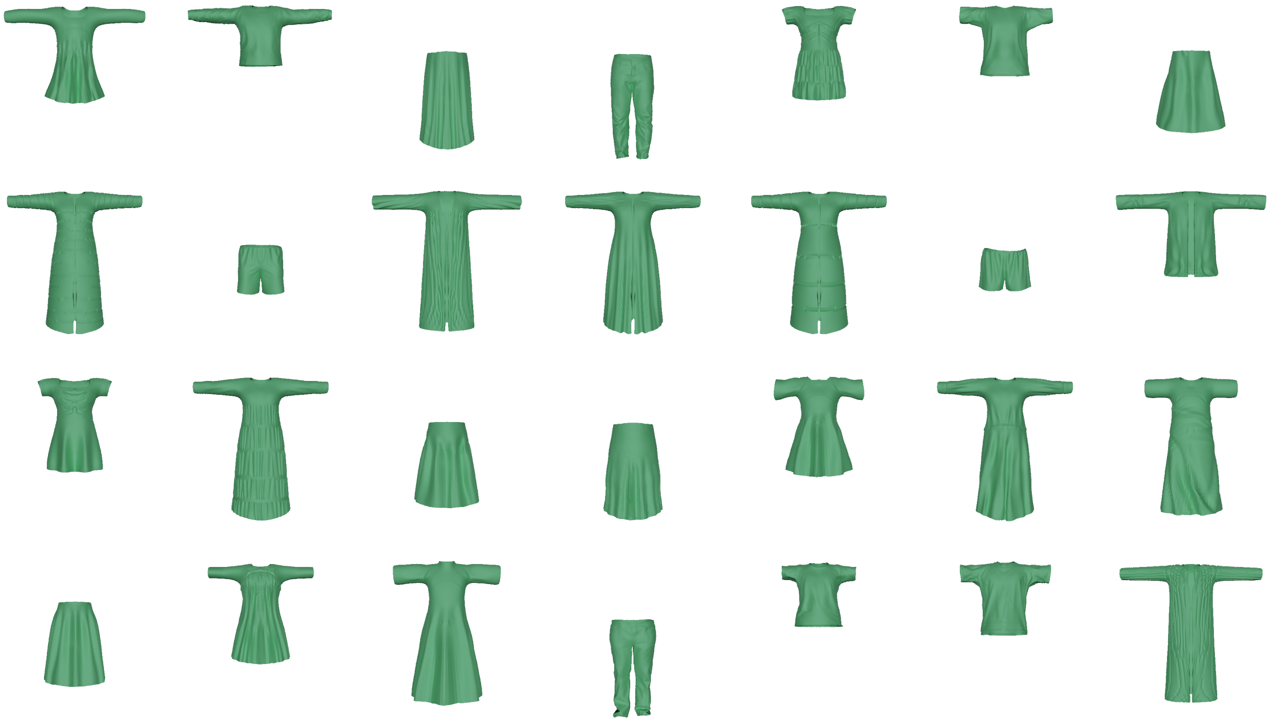

Fig. 19. An illustration of our **Local Detail Database**.

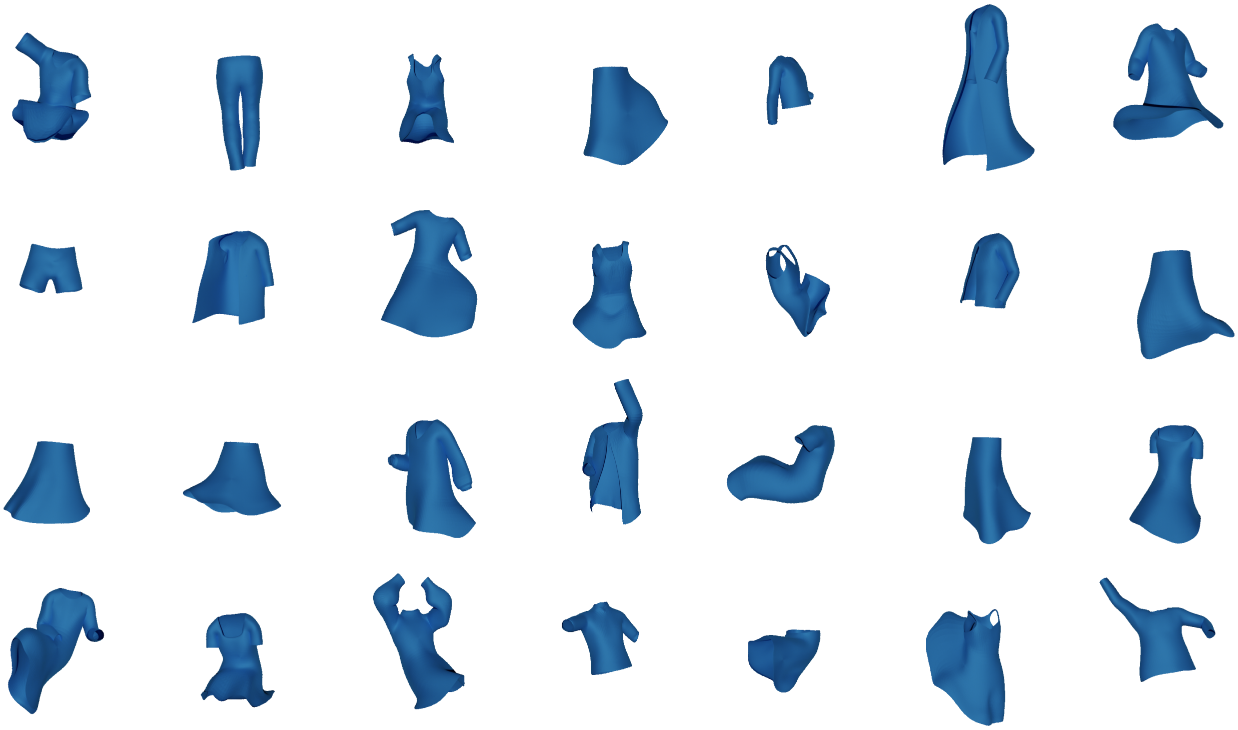

Fig. 20. An illustration of our **Garment Deformation Database**.

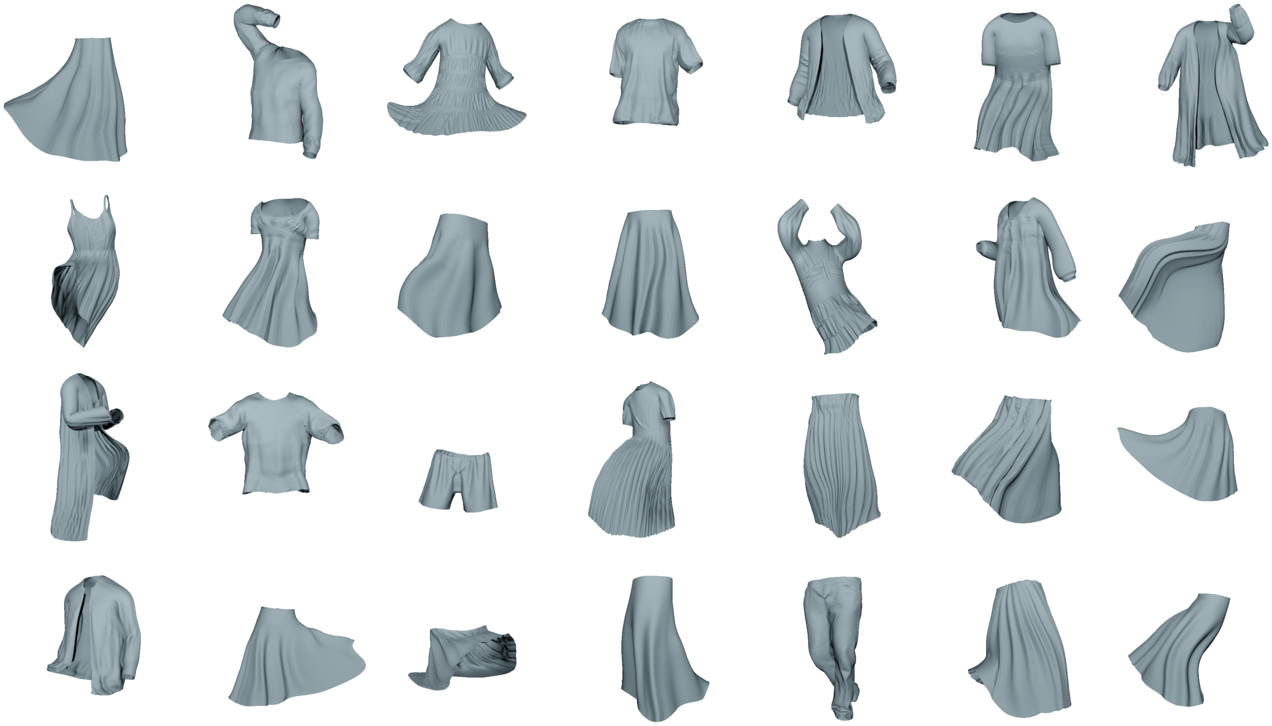

Fig. 21. An illustration of our **Fine Garment Dataset**.

Table 5. Explanation of notations used in the Main Paper.

Notation	Description
LOD	Levels of Details
PCA	Principal Component Analysis
M_C	Coarse garment sampled from the Garment Style Database
L_C, L_F	Garment pair that describes the local geometric detail
M_L	Garment after applying the local details from (L_C, L_F) to M_C
D_T, D_F	Garment pair that depicts global deformation
T	Deformation offsets of (D_T, D_F) in the rest-pose space
LBS	Linear Blend Skinning
M_D	Garment after transferring the deformation from (D_T, D_F) to M_L
$G(\cdot)$	Statistical Garment Model worn on the mean shape of SMPL
$\mathbf{T}_\mathrm{g}$	Garment Template (i.e., The garment mean shape)
$B_\mathrm{g}(\cdot)$	Garment Shape Blend Shape (GSBS) in T-posed space
α	The coefficients of $G(\cdot)$, which control the GSBS
$T_\mathrm{B}(\cdot)$	T-posed Body Mesh
$\mathbf{T}_\mathrm{b}$	Body Template (i.e., SMPL‘s mean shape)
$B_\mathrm{s}(\cdot)$	Body Shape Blend Shape (BSBS) of SMPL
$B_\mathrm{p}(\cdot)$	Body Pose Blend Shape (BPBS) of SMPL
β, θ	The shape and pose parameters of SMPL
$M_\mathrm{B}(\cdot)$	Posed Body Mesh
$W(\cdot)$	Skinning Function
$\mathcal{W}$	Skinning Weights
$J(\cdot)$	Joint Locations
$\widetilde{B}_\mathrm{s}(\cdot)$	Garment displacements influenced by the BSBS, i.e., $B_\mathrm{s}(\cdot)$
$\widetilde{B}_\mathrm{p}(\cdot)$	Garment displacements influenced by the BPBS, i.e., $B_\mathrm{p}(\cdot)$
$w(\cdot)$	Weights for computing garment displacements and skinning
$T_\mathrm{G}(\cdot)$	T-posed garment after applying $\widetilde{B}_\mathrm{s}(\cdot)$ and $\widetilde{B}_\mathrm{p}(\cdot)$ to $G(\cdot)$
$\widetilde{\mathcal{W}}$	Garment skinning weights extended from SMPL
$M_P(\cdot)$	Posed Garment Mesh
M_I	Fine garment predicted by the pixel-aligned network
p	Arbitrary point in 3D space
$I_F(\cdot)$	Pixel-aligned Features
$\pi(\cdot)$	Projection Function
$F(\cdot)$	Feature Extraction Function
$z(\cdot)$	Depth value in the camera coordinate space
$f(\cdot)$	Implicit Function (MLP for decoding the occupancy of p)
s	The occupancy status of p to the garment surface
ψ_{enc}	Triplane Encoder
ψ_{dec}	MLP-based decoder for decoding the occupancy of p
$G_F(\cdot)$	Geometry-aware Features
F_{xy}, F_{xz}, F_{yz}	3D axis-aligned features of three orthogonal planes
$f_i(\cdot)$	Implicit Function of the i-th boundary, i.e., ψ_{dec}
o_i	The occupancy status of p to the i-th boundary
$L_{boundary}$	Boundary Fitting Loss
L_c	Chamfer Distance Loss [Ravi et al. 2020]
L_{lap}	Laplacian Smooth Regularization [Ravi et al. 2020]
L_{edge}	Edge Length Regularization [Ravi et al. 2020]
L_{normal}	Normal Consistency Regularization [Ravi et al. 2020]
$\lambda_c, \lambda_{lap}, \lambda_{edge}, \lambda_{normal}$	Loss Weight
L_{nicp}	Registration Loss (i.e., loss for nicp)
L_d	Distance Cost: Deformed Shape vs. GT [Amberg et al. 2007]
L_b, L_s	Landmark Cost, Stiffness Term [Amberg et al. 2007]
L_{reg}	Mesh Regularization Terms
$\lambda_d, \lambda_b, \lambda_s, \lambda_{reg}$	Loss Weight