Registering Explicit to Implicit: Towards High-Fidelity Garment mesh Reconstruction from Single Images

Heming Zhu[1,2] Lingteng Qiu[1,2] Yuda Qiu[1] Xiaoguang Han[1,3] ✉
[1] SSE, CUHKSZ [2] SRIBD [3] FNii, CUHKSZ
hanxiaoguang@cuhk.edu.cn

Figure 1. Given a single in-the-wild clothed human image, **ReEF** can generate high-fidelity layered garment meshes. The appearances of the reconstructed garments are well aligned with the input image. Moreover, the produced garments can be placed on other virtual characters.

Abstract

*Fueled by the power of deep learning techniques and implicit shape learning, recent advances in single-image human digitalization have reached unprecedented accuracy and could recover fine-grained surface details such as garment wrinkles. However, a common problem for the implicit-based methods is that they cannot produce separated and topology-consistent mesh for each garment piece, which is crucial for the current 3D content creation pipeline. To address this issue, we proposed a novel geometry inference framework **ReEF** that reconstructs topology-consistent layered garment mesh by **re**gistering the **e**xplicit garment template to the whole-body implicit **f**ields predicted from single images. Experiments demonstrate that our method notably outperforms the counterparts on single-image layered garment reconstruction and could bring high-quality digital assets for further content creation.*

1. Introduction

High-quality human-related 3D contents are highly demanded by various real-world applications, including virtual live-streaming, gaming, and filming. However, producing visually plausible 3D digital human assets has always been a laborious task, which may take hours even for an expert modeler.

In contrast, in-the-wild images are easily accessible with commercial cameras and from the internet. Therefore, recent researches have extensively studied human digitalization from single in-the-wild images, aiming at assisting people without expertise to generate visually plausible 3D human-related contents efficiently.

Compared with the recent advances in single image body [3, 13, 20–22, 24, 27, 33, 35, 37] and clothed human reconstruction [1, 2, 7, 25, 32, 36, 38, 42, 44], research on single image layered garment reconstruction is quite sparse. The main challenges towards a high-fidelity garment reconstruction are as two folds: **generating garment styles** and **recovering surface details**. To generate garments with different styles, Multi-Garment Net(MGN) [5], BCNet [17] and SMPLicit [10] adopted either explicit parametric models or implicit parametric models trained on the digital wardrobes but failed to recover the garments with novel garment styles from the image. To generate novel garment styles from the input images, Deep Fashion3D [53] proposed to depict garment styles with the feature lines predicted from the input image. Nevertheless, it fails to generate garment styles well aligned with the input images due to the inac-

curacy of the boundary prediction. As for recovering surface details, MGN [5] and SMPLicit [10] can only produce smoothed garment mesh with limited surface details. BCNet [17] carves fine-grained details onto the garment template with an image-guided graph convolutional network though it fails to produce large-scale wrinkle deformations. Although Deep Fashion3D [53] can generate large-scale surface deformations based on Occupancy Network [31], the generated surface details may deviate from the input image as it only adopts global image features. Therefore, none of existing methods can recover garment styles and surface details aligning with the appearances from the input image.

The recent emergence of the pixel-aligned implicit [38, 39, 49] framework has made it possible to reconstruct clothed humans with image-aligned appearances. On the other hand, it poses a question on how to exploit the power of the pixel-aligned framework to produce layered garment meshes that faithfully reflect the image appearances.

To this end, we propose **ReEF**, a novel geometry inference framework that can produce high-fidelity layered garments by **[re]**gistering the **[e]**xplicit garment template meshes to the full-body implicit **[f]**ields predicted from single images. However, due to the diversity of the real-world garment geometry, it is non-trivial to establish correspondence between the garment template meshes and the clothing on an individual frame of the implicit clothed human. To address this issue, we proposed novel methods to generate boundary fields and semantic fields to align the explicit garment template with the implicit clothed body. On top of the alignment, separated garment meshes with class-specific topology could be instantiated from the implicit fields with a dedicated designed optimization system. Experiments demonstrated that **ReEF** is capable of producing high-quality garment meshes from single images that could serve as off-the-shelf assets for various downstream applications, e.g., animation and simulation.

The main contribution of this work can be summarized as follows:

- We proposed a novel geometry inference framework that reconstructs high-fidelity and topological-consistent garment meshes from single images by registering the explicit garment templates to an individual frame of the implicit clothed human body.
- We contribute to a novel learning-based method that predicts the implicit garment boundary fields with pixel-aligned features and curve-aligned features. The predicted garment boundaries can be well aligned to the appearances from the input image and the implicit clothed body.
- We conducted experiments on both synthetic datasets and in-the-wild datasets. The experiments demonstrated that our method could generate high-quality layered garments with accurate styles and expressive surface details.

2. Related Work

2.1. Single View 3D Human Digitalization.

Human digitalization from a single RGB image is inherently challenging due to the scarcity of information contained in the input regarding the diversity of the shape space. To make the ill-posed problem of single-view 3D human digitalization trackable, SCAPE [3] and SMPL [27] are proposed, which provide strong priors for later human-centric digitization tasks. By simplifying the problem to low-dimensional body parameter estimation, [3, 13, 19–22, 24, 27, 33, 35, 37] achieved human body and pose estimation from a single image. However, these works based on parameterized models [3, 13, 19, 20, 24, 27, 33, 35, 37] are restricted to naked human body reconstruction. Since the garments and surface details are not modeled, the generated shapes are not suitable for visualization applications.

Thanks to the recent rising of 3D deep learning, many works on single-view human digitalization have been proposed to create high-quality 3D clothed human models [1, 2, 7, 25, 32, 36, 38, 42, 44]. The works can be roughly divided into two streams: parametric methods and non-parametric methods. Parametric methods [2, 41, 46, 48, 52] explicitly model the clothed human as the body parameters and the offset to the naked 3D parametric human body models. Though it can generate plausible results even from a single in-the-wild image, it fails to generate loose garments that are not close to the body.

Contrary to parametric-based methods, non-parametric models do not explicitly lean on the parametric human body and could reconstruct the clothed human body with arbitrary topologies. Siclope [32] reconstructed clothed human from multi-view silhouette predicted from a single front-view image. DeepHuman [52] achieved single image human reconstruction with an image-guided volume-to-volume translation network. Although both methods could generate human shapes with detailed garments of arbitrary topology, the details generated are relatively coarse or can not faithfully reproduce the input portrait's appearances. Saito et al. [26, 38, 39] addressed this issue through pixel-aligned implicit function and achieved high-fidelity reconstruction where the geometry generated can be pixel-wise aligned to the input images. Under the pixel-aligned framework, later works attacked the robustness of the model [47, 49], or animating human encoded in the implicit space [14, 15]. However, the above methods fails to provide clothing mesh separated from the human body.

2.2. Single View 3D Garment Reconstruction.

Compared with clothed 3D bodies, layered reconstruction of body and clothing provides easy-to-use assets for

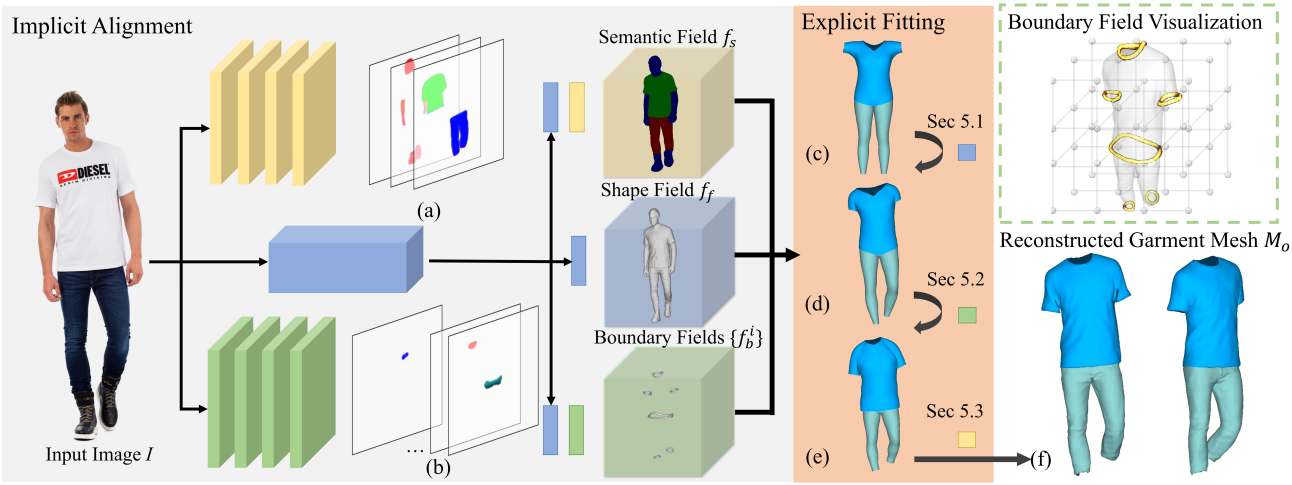


Figure 2. The pipeline of our proposed approach. (a) The semantic attention maps $\{H_s^i\}$. (b) The boundary attention maps $\{H_b^i\}$. (c) The explicit template mesh M_t. (d) The pose deformed template mesh M_p. (e) The boundary deformed template mesh M_l. (f) The output layered garments M_o.

downstream tasks like animation and content creation. However, as garments are shapes with highly diversified topologies and complicated high-frequency surface details, separating garments from the human scan often involves laborious manual efforts. DeepWrinkles [23] and Jin et al. [18] proposed synthesizing high-quality clothing wrinkle deformations as displacement maps or normal maps on the UV space of a fixed garment template, but they are not capable of handling large surface deformations and only support limited garment types.

Closer to our work are Multi-Garment Net(MGN) [5], SMPLicit [10], BCNet [17], and Deep Fashion3D [17]. MGN pioneers by learning a per-category parametric model from a large-scale digital wardrobe. Layered garments could be inferred with MGN with a few images as input. SMPLicit [10] introduced a generative model that supports reconstructing layered garments from a single parsed portrait. However, neither MGN nor SMPLicit can generate high-frequency details from single input images.

To generate layered garments with vivid surface details, Deep Fashion3D [53] adopted Occupancy Network [31] to reconstruct high-frequency surface details from the input image. The surface details generated by Occupancy Network are then transferred to smoothed template mesh with nonrigid-ICP. However, as Occupancy Network lean solely on global image features to produce surface details, it could not loyally recover the appearances from the input portrait. BCNet [17] firstly generates a coarse template mesh with PCA and then emboss surface details with an image-guided graph attention network but fails to produce large-scale wrinkle deformations.

2.3. 3D Shape Registration.

3D Shape registration is a fundamental problem that has been extensively studied in the last decades, targeting setting up correspondences between predefined templates and novel observations. Previous works have tackled registering template mesh to explicit shapes(point clouds or meshes) [9] or registering implicit templates to implicit shapes [11, 51]. Closer to our work, Clothcap [34], MetaAvatar [45], SCALE [29], and SCANimate [40] could register garment templates or human body template to the clothed human scans, though they could only work on scan sequences. POP [30] supports registering the animatable articulated dense point cloud to a single clothed human scan. LoopReg [4] could be adopt for registering the parametric SMPL body model to a clothed human scan bridged by a space-diffused SMPL. MGN [5] and Sizer [43] register the garment templates to human scans though they can only handle tight clothing. Deep Fashion3D [53] registered a boundary-deformed garment template to the reconstructed garment mesh with non-rigid ICP. However, none of the methods mentioned above could register garment template meshes to whole-body implicit fields predicted from a single in-the-wild image.

3. Overview

As illustrated in Figure 2., given a single in-the-wild image I, the goal of **ReEF** is to generate high-fidelity garment mesh M_o with class-specific triangulation by registering the explicit garment template mesh M_t to the predicted whole-body implicit field f_f. To this end, we decompose the whole registration process into two stages: **aligning explicit to implicit**(Section 4) and **fitting explicit to implicit** (Section 5). In the first stage, we will align explicit garment template

mesh M_t to the implicit target f_f with implicit boundary fields $\{f_b^i\}$ and the implicit semantic field f_s predicted from the input image I. On top of the alignment, in the second stage, we will deform the explicit garment template M_t to fit the implicit target f_f with a dedicated designed optimization system.

4. Aligning Explicit to Implicit

Setting up accurate correspondence between the template and the target is the key to achieve a successful registration. In the following section, we will carefully introduce how to align the garment template mesh to the implicit clothed human body from the following aspects: the definition of explicit templates(Section 4.1), the generation of the implicit targets(Section 4.2), and the alignment between the explicit mesh and the implicit fields(Section 4.3).

4.1. Explicit template

We designed class-specific garment template meshes M_t on top of the SMPL [27] body following the previous works [5, 17, 53]. The designed garment template meshes M_t covers 12 common clothes categories, including long/short/no sleeve uppers, long/short/no sleeve dresses, long/short/no sleeve open coats, long/short pants, and skirts. Notably, we define the outermost curves of each garment template M_t as the garment boundaries$\{L_t^i\}$. It is worth mentioning that for garment templates that belong to the open coats categories, the necklines, the center-front lines, and hemlines are treated as different boundaries though they belong to the same curve. Please see the appendix for more details about the explicit garment template.

4.2. Implicit Target

We adopt the pixel-aligned implicit framework to generate the implicit target f_f (i.e., the implicit clothed human), which is superior to its counterparts by producing results that well match the input image.

Pixel-aligned implicit framework The pixel-aligned implicit framework [38, 39] is built upon the implicit shape representation, where a 3D shape can be represented as the occupancy status within a bounded volume. Conditioned on an input image I, the pixel-aligned implicit function could predict the occupancy status of the queried coordinate $X \in R^3$ with:

$$f(X, I) = g(X, \phi_{local}(I, \pi(X))) \tag{1}$$

where $\pi(X) \in (X_x, X_y)$ indicates the projected 2D position on image space. $\phi_{local}(I, \pi(X))$ denotes the image features fetched from the projected position.

Implicit Target Generation. Inspired by the PIFu [38] and PIFuHD [39], the learning process of fine-grained information, e.g., surface details and color, would be more trackable if conditioned on a coarse shape information descriptor. Therefore, we firstly defined the coarse shape field f_c similar to the coarse branch in PIFuHD:

$$f_c(X, I) = g_c(X, \phi_c(I, \pi(X))) \tag{2}$$

where ϕ_c indicates the image features extracted from the down-sampled input.

To emboss the coarse shape fieldf_c with fine-grained details, a fine shape field adopted as the implicit target f_f, is built on top of the coarse shape module. It takes the coarse shape embedding $\Omega_c(X)$ and fine-level image features ϕ_f to predict the occupancy status for the fine-grained shape:

$$f_f(X, I') = g_f(\Omega_c(X), \phi_f(I', \pi'(X))) \tag{3}$$

where I' denotes the cropped input image at the original resolution and $\pi'(X)$ denotes the projected position of the sample points on the cropped image.

4.3. Boundary Alignment

We propose to set up boundary correspondence between the explicit template mesh M_t and the implicit target f_f for registration, as the boundaries possess the most prominent geometrical features of the garment shape. To obtain the garment boundaries on a 3D clothed human, one could turn to scan surface parsing or image-guided curve regression. However, as illustrated in Section. 6, the garment boundaries generated with surface parsing would be heavily corrupted due to the occlusion of the human body and other accessories. Although the image-guided curve regression can always deliver complete boundary curves, it fails to produce accurate boundary which aligns with the implicit target f_f. To this end, we propose a novel method that predicts a set of garment boundary fields $\{f_b^i\} \in (-1, 1)$ from the input image I, each representing a type of garment boundaries, e.g., collars, cuffs, hemlines.

Garment boundary fields. The garment boundaries are thin 3D spatial curves that are inherently hard to be captured by the implicit functions. Hence, instead of modeling each boundary curve directly with the implicit function, as illustrated in Figure.2, we propose to model each garment boundary as an implicit cylinder with a signed distance field:

$$f_b^i(X) = d(X - l_b^i) - \epsilon_b \tag{4}$$

where the $d(X - l_b^i)$ denotes the distance from the query point $X \in R^3$ to i^{th} garment boundary. ϵ_b indicates the radius of the boundary cylinder that is set to $1e^{-3}$ empirically.

Vanilla approach. We design a vanilla approach to predict the garment boundary fields $\{f_b^i\}$ from the input image I. To make sure that the predicted boundary fields $\{f_b^i\}$ are aligned with the target shape field f_f, we jointly trained the garment boundary fields $\{f_b^i\}$ and the target shape fields f_f conditioned on the same coarse shape embedding $\Omega_c(X)$:

$$f_b^i(X, I) = g_b^{vanilla}(\Omega_c(X)) \tag{5}$$

Although the vanilla approach can produce garment boundary fields $\{f_b^i\}$ that align with the target shape field f_f, it may not reflect the boundary appearances of the input image due to the lack of guidance received from the image space.

Curve-aligned boundary generation. We thus propose a curve-aligned boundary generation module to generate garment boundary fields $\{f_b^i\}$ that accurately reflect the boundaries' appearances from the input image I. Compared with the vanilla approach which solely bases on pixel-aligned coarse shape feature $\Omega_c(X)$, our proposed curve-aligned boundary generation module may receive extra guidance, i.e., curve-aligned features, from the image space.

To produce curve-aligned features for boundary field generation, we designed garment boundary attention maps which depict the likelihood of each garment boundary on image space. The garment boundary attention maps $\{H_b^i\}$ are generated from the input image I with HigherHRNet [8] and could receive supervision from the ground-truth boundary heatmap. Conditioned on the curve-aligned features produced by boundary attention maps, the garment boundary fields can be generated with:

$$f_b^i(X, I) = g_b(\Omega_c(X), \phi_h(I, \pi(X))) \tag{6}$$

where $\phi_h(I, \pi(X))$ denotes the curve-aligned features sampled from the boundary attention maps$\{H_b^i\}$.

4.4. Semantic Alignment.

Apart from the boundary correspondence, semantic correspondence between the explicit template M_t and the implicit target f_f are required to mute the influences of non-relevant regions on the implicit targetf_f. To this end, we designed semantic implicit fields $\{f_s^i(X, I)\}$, which denotes the occupancy likelihood for each kind of clothing(i.e., upper body clothing and lower body clothing) in 3D space. Notably, similar to the generation of garment boundary fields$\{f_b^i\}$, semantic attention maps $\{H_s^i\}$ predicted from the input images I are adopted for additional 2D guidance:

$$f_s^i(X, I) = g_s(\Omega_c(X), \phi_s(I, \pi(X))) \tag{7}$$

where $\phi_s(I, \pi(X))$ denotes the semantic attention map features fetched from the projected position $\pi(X)$. The semantic label for each 3D query point X can be predicted by aggregating the implicit semantic fields of the possible labels:

$$f_s(X, I) = \arg\max_i(f_s^i(X, I)) \tag{8}$$

5. Explicit Fitting

In the previous section, we have bridged the gap between the explicit garment template M_t and the implicit target(i.e., clothed human) with boundary correspondence $\{f_b^i\}$ and semantic correspondence $\{f_s^i\}$ predicted from the input image I. On top of the established correspondences, we proposed an explicit fitting pipeline, which progressively deforms the garment template mesh M_t to be aligned with the implicit target f_f. The proposed explicit fitting pipeline consists of four phases, namely, template initialization(Section 5.1), boundary fitting(Section 5.2), template fitting(Section 5.2) and post processing(Section 5.4).

5.1. Template Initialization

As the explicit garment template mesh M_t is built on top of the SMPL parametric human body [27], accurate body pose and shape estimation may benefit the registration process by setting up a good initialization, i.e., the pose deformed garment template mesh M_p. However, estimating accurate 3D pose from a single in-the-wild image is inherently challenging due to the depth ambiguity, unknown camera parameters, and the scarcity of annotated in-the-wild datasets. In contrast, the state-of-the-art 2D pose estimation has reached relatively high accuracy on in-the-wild images. To this end, we propose to optimize the SMPL body parameters $SMPL(\theta, \beta)$ to be aligned with the implicit shape field under the additional guidance of selected 2D joints J_{gt} predicted by off-the-shelf singe image pose estimator [6]:

$$\begin{aligned} V_{pred}, J_{pred} =& SMPL(\theta, \beta) \\ \mathcal{L}_{body} =& MSE(J'_{pred}, J_{gt}) + \eta_{reg} Reg(\theta) \\ &+ \eta_{shape} CD(V_{lres}, V_{pred}) \end{aligned} \tag{9}$$

where Reg denotes the pose regularization function adopted to reduce undesired poses and V_{lres} indicates the low-resolution mesh vertices extracted from the coarse field f_c.

5.2. Boundary Fitting

In Section 4.3, we have established the correspondence between the boundaries$\{l_t^i\}$ of the template mesh M_t and the garment boundaries of the implicit target f_f with the boundary fields $\{f_b^i\}$. Based on the boundary correspondence, we may deform the boundaries $\{l_p^i\}$ of the pose deformed template mesh M_p to be aligned with the garment boundaries of the implicit target f_f:

$$\mathcal{L}_b = f_b^i(l_p^i) + \eta_{ea} Avg(e_b^i) + \eta_{ed} Var(e_b^i) \tag{10}$$

where e_b^i denotes the boundary edge lengths of the garment template mesh M_p. The optimized garment boundaries $\{l_a^i\}$ are set as the hard constraints for Bi-Harmonic deformation [16]. So far, a plain garment template mesh M_l is produced with the garment boundaries $\{l_a^i\}$ aligned with the garment boundaries of the implicit target f_f.

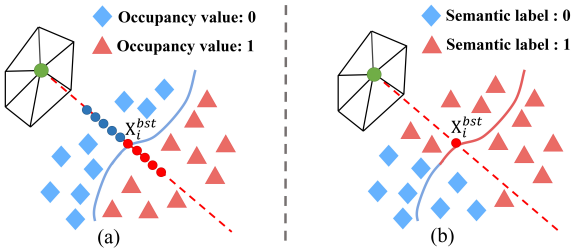


Figure 3. A illustration of our proposed active area probing scheme, our proposed active area probing scheme could aggregate the implicit shape information and implicit semantic information to guide the deformation of the garment template mesh.

5.3. Shape Fitting

By the end of the boundary fitting stage, a plain garment template mesh M_l is generated whose boundaries are aligned to the garment boundaries of the implicit target f_f. To emboss the boundary-aligned plain template mesh M_l with fine-grained details, we compiled an optimization system with the following design goals: Firstly, the resulting garment M_o mesh should stick tightly with the corresponded part on the implicit target f_f. Secondly, the boundaries of the resulting mesh $\{l_o^i\}$ should remain aligned with the garment boundaries of the implicit target f_f. Thirdly, the resulting mesh M_o should not penetrate with the predicted human body mesh M_{smpl}.

However, as the implicit target f_f encodes the whole clothed human body, directly fitting the explicit template M_p to the implicit target f_f may be affected by non-relevant regions, i.e., hairs, skins, or other clothes. To address this issue, we propose an operation called activate area probing, which predicts the activated template vertices to be deformed, and the approximated distances D_{act} between the activated vertices to the corresponding areas on the implicit target f_f.

Activate Area Probing. Given a vertex X_i on the explicit template M_l, we cast rays in both directions along the vertex normal to sample k points on each direction. For each vertex X_i, 2k + 1 points $\{X_i^0, X_i^1, ..., X_i^{2k-1}, X_i\}$ in total are fed to both implicit target field f_f and implicit semantic fields $\{f_s^i\}$ in batches. The approximate distance can be calculated as the distance from the template vertex X_i to the closet sample point X_i^{bst} that penetrates the surface threshold ϵ (i.e. 0.5). The activation status $B_i \in \{0, 1\}$ for the vertex X_i is set to active only when the cast rays reach the iso-surface and the semantic label at the penetration point $f_s(X_i^{bst})$ is consistent with the current template. Finally, the approximated distance from the activate areas on of the explicit template mesh M_t to the implicit target f_f can be calculated with:

$$D_{act}(M_o) = Avg(B_i \dot{M}SE(X_i, X_i^{bst}))/Sum(B_i) \quad (11)$$

With the proposed activate region loss $D_{act}(M_o)$, we can update the explicit template mesh M_o to fit corresponding areas on the implicit target f_f with the following loss function:

$$\begin{aligned}\mathcal{L}_o = D_{act}(M_o) - \eta_{pen} TSDF(M_{smpl})(M_o) \\ + \eta_b \mathcal{L}_b + \eta_{lap} \mathcal{L}_{lap}\end{aligned} \quad (12)$$

where $TSDF(M_{smpl})$ indicates the truncated signed distance function of the posed human body mesh M_p adopted for penalizing the garment-body penetration, and $\mathcal{L}_{lap}$ denotes the laplacian of the deformed template mesh. By the end of the explicit fitting stage, we will obtain a high-fidelity garment mesh M_o well align with the input image I.

5.4. Post Processing

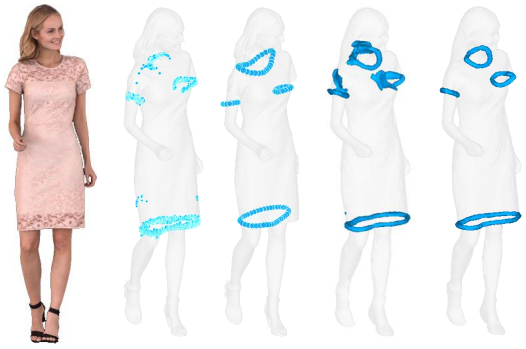
While the reconstructed garment meshes M_o could well recover the garment styles and surface details from an in-the-wild input image, like most existing image-based reconstruction methods, it may fail to reconstruct folded structures like the collars. Therefore, we manually created a collar warehouse containing various real-world collars built upon the garment template and trained a light-weight image classification network to choose the collar type with the closest appearance to the image. Thanks to the topology-consistent nature of our generated garment mesh, the collar can be attached to the garment template through vertex correspondence. The collar's geometry is further tuned with Bi-Harmonic deformation to be collocated with the reconstructed garment mesh.

6. Experiment Results

6.1. Implementation Details

Data Preparation We adopt RenderPeople [12] data to train our proposed model, which contains 400 photo-realistic 3D clothed humans with high-resolution textures and surface semantic parsing. We split the whole dataset into a training set of 360 subjects and a testing set of 40 subjects. All of the textured scans are rendered following the settings in PIFuHD [39]. It is worth mentioning that although the semantic parsing provided by RenderPeople could help to identify the garment boundaries automatically, they may be heavily corrupted due to the occlusion of the human body and accessories. Therefore, we hired professional artists to annotate the garment boundaries on the scan surfaces. More importantly, the artists may link the incomplete boundary segments into smoothed closed curves with their expertise in garments' shape.

Network Training The coarse shape, boundary, and semantic field generation modules are trained with the input image rescaled to 512×512. The target shape field generation module is trained with random cropped images at the original resolution with window size as 512×512. We jointly train the coarse shape generation module, the boundary field



(a) (b) (c) (d) (e)

Figure 4. Qualitative comparison of the garment boundaries produced under different ablation settings. The input image (a) is followed by the garments generated with (b) **PCT**, (2) **GCN**, (3) **w/o HM** and (4) **Ours**.

generation module, and the semantic field generation module with a learning rate of 1×10^{-4} for six epochs from scratch. The fine shape module is trained conditioned on a fixed coarse shape module with a learning rate of 1×10^{-4}. It takes roughly 72 hours to train all the modules mentioned above on two GTX 3090 GPUs. Please refer to the appendix for more implementation details about the network training and the explicit fitting.

6.2. Ablation Studies

In this section, we compile a set of ablation experiments to verify the algorithmic components' effectiveness for our boundary field generation module. Please refer to the appendix for more details on the ablations for the explicit fitting stage.

We orchestrated both quantitative and qualitative comparisons between our proposed model and the alternatives that take other candidate design choices: 1) Predict the garment boundaries by parsing the dense point cloud sampled on the explicit surface with Point Transformer [50], termed as **PCT**. 2) Predict the garment boundaries by regressing the explicit curves with an image-guided graph convolution network, termed as **GCN**. 3) Predict the garment boundaries with pixel-aligned coarse shape features identical to the vanilla approach mentioned in Sec.4.3, termed as **w/o HM**. 4) The proposed full model, termed as **Ours**. Specifically, we extracted explicit mesh from the garment boundary fields generated with setting **w/o HM** and **Ours** with Marching Cubes [28] for later comparison. Table.1 shows the quantitative comparisons between the design alternatives and the proposed one. As seen, the proposed approach exhibits the best accuracy among all the settings.

Figure.4 shows the visualization results generated under different experiment settings. As clothed human bodies are highly diversified shapes with various surface details, **PCT** may produce corrupted garment boundaries and noisy parsing. Though **GCN** can produce complete curves, the boundary curves generated with **GCN** are largely deviates the clothing boundary. Due to the lack of the guidance from the image space, **w/o HM** may generate garment boundaries with undesired shapes. **Ours** can produce clean garment boundaries that are well aligned with the boundary appearances from the images.

Methods	**PCT**	**GCN**	**w/o HM**	**Ours**
CD($\times 10^{-3}$)	6.5329	9.18467	6.3786	**1.1073**

Table 1. The quantitative comparison between the proposed model and the ablation alternatives.

6.3. Comparison Experiments

We compared our method with the state-of-the-art single image garment reconstruction methods of which the codes are publicly available, i.e., Multi-Garment Net [5], BCNet [17], and SMPLicit [10], both quantitatively and qualitatively.

Quantitative Comparison We test our method and the state-of-the-art methods with the rendered images from our synthetic testing set. Notably, the garment meshes generated by different methods are aligned to the ground truth garment meshes with the underlying SMPL body. After aligning the results to the ground truth garment mesh, we compute the Chamfer Distance (CD) between the reconstructed mesh and ground truth for accuracy measurement. As illustrated in Table.2, our method outperforms the comparison counterparts in reconstruction accuracy by a large margin.

Methods	MGN	SMPLicit	BCNet	Ours
CD($\times 10^{-3}$)	1.1424	1.3408	0.9725	**0.5477**

Table 2. The quantitive comparison between our model with the state of the art garment reconstruction methods.

Qualitative Comparison Figure.6 provides qualitative comparisons on the results generated with in-the-wild images collected from the internet. Compared to the other methods, our method is superior in reconstructing accurate garment styles and reproducing the surface details faithfully. Figure.7 demonstrates the qualitative comparison between our method and BCNet [17]. While BCNet [17] fails to produce garments with the correct style, our proposed method could reconstruct the garment meshes with boundaries and surface details highly identical to the image input.

6.4. Gallery on in-the-wild images

Figure.5 shows the results generated by our proposed method on in-the-wild images. The results demonstrate that our method could produce high-quality garments with fine grained details and correct garment styles.

Figure 5. The results generated by our method on in-the-wild images. Each image is followed by the reconstructed layered garment mesh.

Figure 6. Qualitative comparison between ours and the state of the arts. For each row, the input image (a) is followed by the results generated by (b) Multi-Garment Net [5], (c) SMPLicit [10], (d) BCNet [17] and (e) our method.

Figure 7. The qualitative comparison between our proposed method and BCNet [17] on dress reconstruction.

7. Conclusion and Limitation

Layered garment reconstruction from a single in-the-wild image is inherently a challenging problem due to the highly diversified garment shapes and the high-frequency details. To this end, we proposed a novel pipeline which faithfully recovers high-quality garments from a single image by registering the explicit mesh to the implicit fields. A novel garment boundary field generation model is proposed to align the explicit template mesh to the implicit target. Based on the seamless alignment between the explicit and the implicit, the proposed pipeline may produce high quality garment meshes through fitting template mesh to the corresponding parts on the implicit target.

Limitations Although our methods may generate high quality garments from a single image. It only supports reconstructing clothing in common clothes categories. In the future, we will attack the problem of the generation of clothing with complex topology and multi-layered clothing.

Acknowledgement The work is supported by the Basic Research Project No.HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, National Key R&D Program of China with grant No.2018YFB1800800 by Shenzhen Outstanding Talents Training Fund 202002, and by Guangdong Research Projects No.2017ZT07X152 and No.2019CX01X104. It is also supported by NSFC-62172348, 61902334 and Shenzhen General Project (JCYJ20190814112007258). We thank the ITSO in CUHKSZ for their High-Performance Computing Services.

References

[1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2019. 1, 2

[2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 1, 2

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. *ACM Transactions on Graphics*, 24(3):408–416, 2005. 1, 2

[4] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Neural Information Processing Systems (NeurIPS)*, December 2020. 3

[5] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 1, 2, 3, 4, 7, 8

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018. 5

[7] Xiaowu Chen, Yu Guo, Bin Zhou, and Qinping Zhao. Deformable model for estimating clothed and naked human shapes from a single image. *The Visual Computer*, 29(11):1187–1196, 2013. 1, 2

[8] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 5

[9] Haili Chui and Anand Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 89(2-3):114–141, 2003. 3

[10] 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 (CVPR)*, pages 11875–11885, June 2021. 1, 2, 3, 7, 8

[11] Yu Deng, Jiaolong Yang, and Xin Tong. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10286–10296, 2021. 3

[12] Renderpeople GmbH. Renderpeople. https://renderpeople.com/, 2019. 6

[13] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and H-P Seidel. A statistical model of human pose and body shape. In *Computer graphics forum*, volume 28, pages 337–346. Wiley Online Library, 2009. 1, 2

[14] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11046–11056, October 2021. 2

[15] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[16] Alec Jacobson, Ilya Baran, Jovan Popović, and Olga Sorkine-Hornung. Bounded biharmonic weights for real-time deformation. *Communications of the ACM*, 57(4):99–106, 2014. 5

[17] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. *arXiv preprint arXiv:2004.00214*, 2020. 1, 2, 3, 4, 7, 8

[18] Ning Jin, Yilin Zhu, Zhenglin Geng, and Ronald Fedkiw. A pixel-based framework for data-driven clothing. *arXiv preprint arXiv:1812.01677*, 2018. 3

[19] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 2

[20] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 1, 2

[21] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11127–11137, October 2021. 1, 2

[22] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. Spec: Seeing people in the wild with an estimated camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11035–11045, October 2021. 1, 2

[23] Zorah Lahner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 667–684, 2018. 3

[24] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6050–6059, 2017. 1, 2

[25] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *International Conference on 3D Vision (3DV)*, sep 2019. 1, 2

[26] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *European Conference on Computer Vision*, pages 49–67. Springer, 2020. 2

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned

multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015. 1, 2, 4, 5

[28] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 7

[29] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J. Black. SCALE: Modeling clothed humans with a surface codec of articulated local elements. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 3

[30] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021. 3

[31] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2, 3

[32] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4480–4490, 2019. 1, 2

[33] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 International Conference on 3D Vision (3DV)*, pages 484–494. IEEE, 2018. 1, 2

[34] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics (SIGGRAPH)*, 36(4), 2017. 3

[35] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J Black. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics (TOG)*, 34(4):120, 2015. 1, 2

[36] Albert Pumarola, Jordi Sanchez, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3DPeople: Modeling the Geometry of Dressed Humans. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[37] Helge Rhodin, Nadia Robertini, Dan Casas, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *European conference on computer vision*, pages 509–526. Springer, 2016. 1, 2

[38] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019. 1, 2, 4

[39] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 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*, pages 84–93, 2020. 2, 4, 6

[40] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 3

[41] Feitong Tan, Hao Zhu, Zhaopeng Cui, Siyu Zhu, Marc Pollefeys, and Ping Tan. Self-supervised human depth estimation from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 650–659, 2020. 2

[42] Sicong Tang, Feitong Tan, Kelvin Cheng, Zhaoyang Li, Siyu Zhu, and Ping Tan. A neural network for detailed human depth estimation from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7750–7759, 2019. 1, 2

[43] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *European Conference on Computer Vision (ECCV)*. Springer, August 2020. 3

[44] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2018. 1, 2

[45] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. In *Advances in Neural Information Processing Systems*, 2021. 3

[46] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica K. Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular RGB video. *CoRR*, abs/2009.10711, 2020. 2

[47] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. ICON: Implicit Clothed humans Obtained from Normals. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 2

[48] Shan Yang, Zherong Pan, Tanya Amert, Ke Wang, Licheng Yu, Tamara Berg, and Ming C. Lin. Physics-inspired garment recovery from a single-view image. *ACM Trans. Graph.*, 37(5), nov 2018. 2

[49] Zheng Zerong, Yu Tao, Liu Yebin, and Dai Qionghai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction, 2021. 2

[50] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021. 7

[51] Zerong Zheng, Tao Yu, Qionghai Dai, and Yebin Liu. Deep implicit templates for 3d shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1429–1439, 2021. 3

[52] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[53] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *European Conference on Computer Vision (ECCV)*, pages 512–530. Springer, August 2020. 1, 2, 3, 4