Garment4D: Garment Reconstruction from Point Cloud Sequences

Anonymous Author(s)
Affiliation
Address
email

Abstract

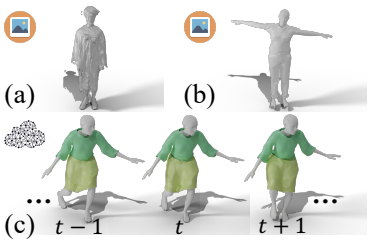
Learning to reconstruct 3D garment models is important for dressing 3D human bodies with different shapes and poses. Previous works typically rely on 2D RGB images as input, which, however, suffers from the scale and pose ambiguities. In this paper, we propose a principled framework, **Garment4D**, that uses 3D point cloud sequences of dressed humans for garment reconstruction. Garment4D has three dedicated steps: sequential garments registration, canonical-posed garment shape estimation, and posed garment reconstruction. The main challenges are two-fold: **1)** effective 3D feature learning for fine details, and **2)** garment dynamics capturing caused by the interaction between garment and human body, especially for loose garments like skirts. To unravel these problems, we introduce a novel proposal-guided hierarchical feature network and iterative GCN, which integrate both high-level semantic features and low-level geometric features to learn effective 3D features for fine details. Furthermore, we propose a temporal transformer for capturing smooth garment motions. Unlike non-parametric methods, the reconstructed garment meshes of our method are separated from human body, and hence have strong interpretability, which is desirable for downstream tasks. As the first attempt at this task, high-quality reconstruction results are qualitatively and quantitatively illustrated through extensive experiments. We hope our work could pave way for more robust and dynamic 3D perception.

1 Introduction

Garment reconstruction is a crucial technique in many applications, *e.g.* virtual try-on [1], VR/ AR [2] and visual effects [3]. Extensive efforts [4–15] have been put into reconstructing a human body and its garments as a whole with the help of implicit representation or volumetric representation. However, it is desirable to have a controllable garment model itself in many applications. Here, we focus on the parametric modeling of 3D garment, which has advantages in the following two ways. First, we can separate garments from human body. Second, the topology of the reconstructed meshes can be controlled and therefore allows downstream tasks that require high interpretability.

Different from previous parametric methods for garment reconstruction [7, 8, 16, 17] that take 2D RGB images as input, we choose to approach the problem from the perspective of 3D, specifically point cloud sequences (shown in Fig. 1), for the following three reasons. Firstly, 3D inputs eliminate scale and pose ambiguities that are difficult to avoid when using 2D images. Secondly, exploiting temporal information is important for garment dynamics capturing, at which there are few attempts. Thirdly, recent development in 3D sensors (*e.g.* LiDAR) has reduced the cost and difficulties of obtaining point clouds, which makes it essential to leverage 3D point clouds for research problems and commercial applications. However, garment reconstruction from point cloud sequences has yet

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.



Method Type	(a) Template-Free	(b) SMPL+D	(c) Garment4D
Input Modality	image	image	pcd seq.
No Ambiguity	✗	✗	✓
Separable	✗	✗*	✓
Interpretability	✗	✓	✓
Loose Garments	✓	✗	✓

Figure 1 & Table 1: Comparison Between Popular Garment Reconstruction Methods. Two major types of garment reconstruction methods, *i.e.* **(a)** template-free and **(b)** SMPL+D-based methods, are compared to **(c)** Garment4D in terms of input modality, scale/ pose ambiguity, ability to separate garments from body, mesh interpretability and ability to deal with loose garments. *Some modified version of SMPL+D [16] could separate garments from body. 'pcd seq.' stands for point cloud sequence.

been properly explored by previous works, and we contribute an early attempt at this meaningful task in this paper. Our advantages against other garment reconstruction methods are listed in Tab. 1.

The proposed garment reconstruction framework, Garment4D, consists of three major steps: **a)** sequential garments registration, **b)** canonical-posed garment shape estimation and **c)** posed garment reconstruction. In the first step, we have several sequences of garment meshes which have different mesh topology across the sequences but share same topology inside one sequence. For each type of garments (*e.g.* T-shirt, Trousers, Skirt), we define a template mesh as the registration target for the canonical-posed garments, which is solved by an optimization-based method. Thereafter, inside each sequence, a barycentric interpolation method is used to register the posed garment meshes to the corresponding registered canonical-posed garment meshes. For the second step, following the practice of previous parametric methods [16, 17], we estimate the corresponding canonical-posed garment mesh for each sequence by a semantic-aware garment PCA coefficients encoder, which takes the dressed human point cloud sequences as input to predict a set of PCA coefficients.

The third step is where the challenges come in for two reasons. Firstly, due to unordered and unstructured nature of point clouds, it is non-trivial to efficiently learn low-level geometric features directly from unorganized 3D points. Secondly, it is challenging to capture garment dynamics caused by interactions between the human body and its garments, especially for the non-rigid deformation of loose garments (*e.g.* skirts), which depends on both current human pose and previous human motions. To address the challenges, we first apply the Interpolated Linear Blend Skinning (LBS) to the estimated canonical-posed garments from the previous step as proposals of the posed garments. Unlike [7, 8, 16] that depends on SMPL+D model to skin the garment, the proposed interpolated LBS has the ability to perform skinning on loose garments without artifacts. Moreover, our method does not require learning which is different from that of [17]. After obtaining the proposals, we present a proposal-guided hierarchical feature network along with iterative Graph Convolution Network (GCN) for efficient geometric feature learning. Meanwhile, a temporal transformer is utilized for temporal fusion which facilitates capturing smooth garment dynamics.

Our contributions can be summarised as:

1) We propose a practical pipeline for the reconstruction of various garments (including challenging loose garments) that are driven by large human movements, which does not rely on any particular human body parametric models or extra annotations other than human and garment meshes.

2) Many novel learning modules, such as proposal-guided hierarchical feature network, iterative GCN, and temporal transformer are proposed, which effectively learns spatio-temporal shape and pose features from 3D point cloud sequences for garment reconstruction.

3) We establish a comprehensive dataset consisting of $2,300$ 3D point cloud sequences by adapting CLOTH3D [18], which serves as the first benchmark dataset for evaluating posed garment reconstruction based on point cloud sequences.

In addition, this work presents an early attempt at exploiting temporal information for capturing garment motions. To the best of our knowledge, we are the first to explore garment reconstruction from point cloud sequences.

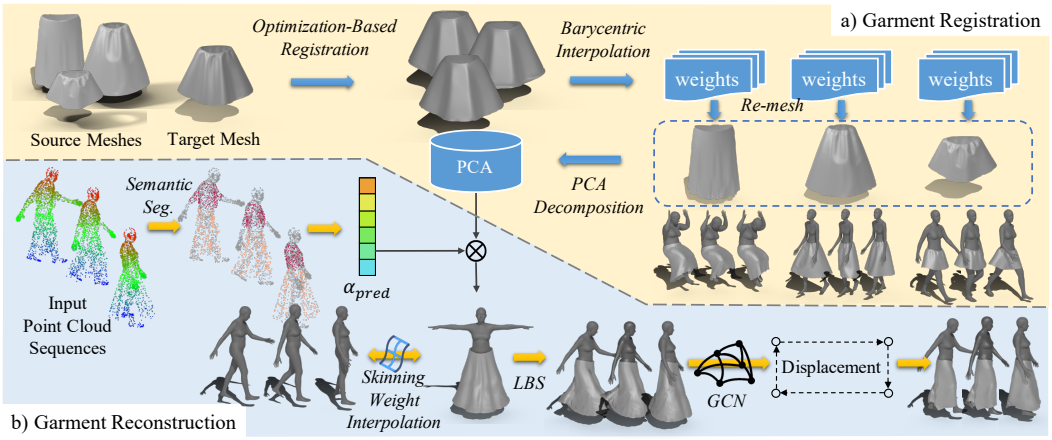


Figure 2: Garment4D Pipeline Overview. a) The upper part illustrates the process of garment registration. Source meshes are registered to target meshes and re-meshed using barycentric interpolation weights. **b)** The lower part shows the framework of garment reconstruction network. The first stage takes point cloud sequences as input and predict PCA coefficients to estimate the canonical-posed garment. The second stage predicts displacements guided by the interpolated LBS proposals.

2 Our Approach

We introduce our approach, Garment4D, in this section. Firstly, our garment model are formulated in Subsection 2.1. Afterwards, we successively elaborate the three major steps of Garment4D: 2.2 sequential garments registration, 2.3 canonical-posed garment shape estimation, and 2.4 posed garment reconstruction. In Subsection 2.5, our training loss functions are introduced. An overview of the proposed Garment4D is shown in Fig. 2.

2.1 3D Garment Model

We design the garment model by the following three steps: a) canonical-posed garment model using PCA, b) skinning according to human poses, and c) displacement. Formally, the canonical-posed garment $T \in \mathbb{R}^{N\times 3}$ can be formulated as

$$T(\alpha) = G + C\alpha, \tag{1}$$

where N is the number of vertices of the garment meshes, $G \in \mathbb{R}^{N\times 3}$ represents the mean shape of current type of garment, $C \in \mathbb{R}^{N\times a\times 3}$ is the PCA components that form garment shape sub-spaces, $\alpha \in \mathbb{R}^{a}$ re presents the PCA coefficients of the corresponding garment.

Interpolated LBS. The skinning step is implemented by Linear Blend Skinning. One critical problem is to get a set of skinning weights of garment vertices that can provide reasonable proposals for posed garments. Previous methods [7, 8, 16] approximate the garment skinning weights by using the weight of the closest human mesh vertex for each garment vertex. However, these methods can result in mesh artifacts, and hence limiting the quality of reconstructed garments for the following two facts. Firstly, the resolutions of our garment meshes is higher than that of human meshes. Secondly, loose garments like skirts do not always share same topology with human body. To avoid artifacts and jagged meshes, we propose to interpolate the skinning weights from K nearest neighbors of human mesh vertices $T_h \in \mathbb{R}^{M\times 3}$. Laplacian smoothing is also performed to further ensure the smoothness of the skinning weights. For those tight garments like shirts and trousers, neighboring interpolations can provide reasonable reconstruction results. For those challenging loose garments like skirts and dresses, the interpolated results are mostly not satisfactory. To this end, we further refine the results of interpolated LBS by predicting displacements $D \in \mathbb{R}^{N\times 3}$. In general, the whole garment model can be formulated as

$$M(\alpha, \theta, K, \mathcal{W}_h, D) = W(T(\alpha), J, \theta, \mathcal{W}_g(\alpha, T_h, \mathcal{W}_h, K)) + D, \tag{2}$$

$$\mathcal{W}_g(\alpha, T_h, \mathcal{W}_h, K) = \mathcal{W}_g^K(T(\alpha), T_h, \mathcal{W}_h, K) I_w(T(\alpha), T_h, K), \tag{3}$$

where $\theta \in \mathbb{R}^{|J|\times 3}$ represents the axis-angle of each joint, $\mathcal{W}_h \in \mathbb{R}^{M\times |J|}$, $\mathcal{W}_g \in \mathbb{R}^{N\times |J|}$ are the skinning weights of human and garment mesh vertices, J is the joint locations of human body,

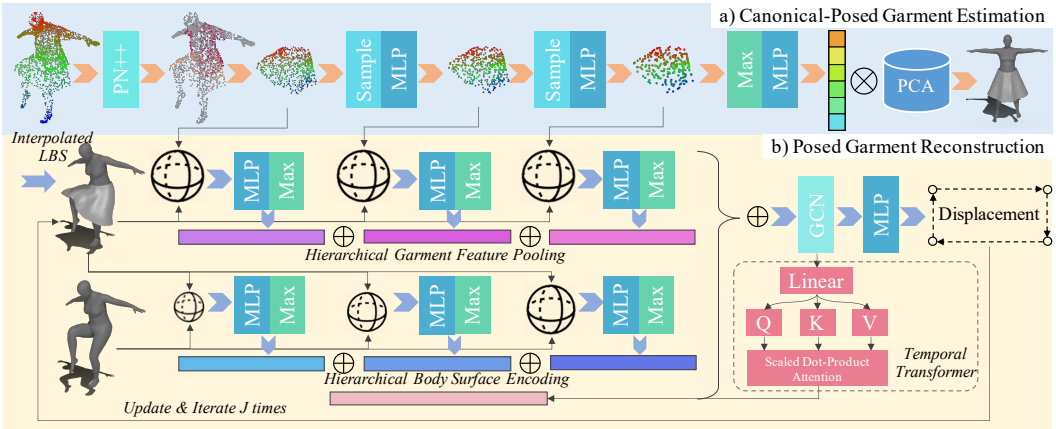


Figure 3: Details of Garment Reconstruction Network. a) The canonical-posed garment estimation takes point clouds as input. Semantic segmentation is performed first to get the garment point clouds, upon which down-sampling and feature extraction are further applied to predict the PCA coefficients. **b)** Using the interpolated LBS, proposals of posed garments are obtained, guided by which, a hierarchical feature network along with iterative GCN and temporal transformer are used to iteratively predict per-vertex displacements.

$\mathcal{W}_g^K \in \mathbb{R}^{M\times|J|\times K}$ is the garment skinning weights of K nearest body vertices, $I_w \in \mathbb{R}^{M\times K}$ is the inverse distance weights used for interpolation.

2.2 Garment Registration

In order to perform PCA decomposition to the same type of garments, we unite the mesh topology by registering the raw meshes $R \in \mathbb{R}^{X\times 3}$ to a template mesh $T \in \mathbb{R}^{N\times 3}$. A visualization of the registration process is illustrated in Fig. 2 (a). Suppose different observations for the same garment share the same topology in one point cloud sequence, we only register one frame of the sequence to the template mesh. For the rest frames, we use barycentric interpolation to re-mesh each observed garment to its template topology.

Considering that the source meshes have large shape variance and diverse sleeve/ hem length, we use a boundary-aware optimization method to align the source meshes to the target. The registration loss function L_{reg} can be formulated as

$$L_{reg} = \lambda_c \text{CD}(R', T) + \lambda_e \text{EL}(R') + \lambda_n \text{NC}(R') + \lambda_b \text{BC}(R', T), \tag{4}$$

where $\text{CD}(R', T)$ is the Chamfer Distance (CD) between source and target mesh vertices, $\text{EL}(R')$ normalized edge length of the source mesh, $\text{NC}(R')$ maintains the normal consistency between neighbouring faces, $\text{BC}(R', T)$ minimizes the CD between corresponding boundaries of source and target meshes (*e.g.* cuffs and neckline), and λ_* represents the corresponding weight for each loss term. The BC term makes the optimization process aware of the boundaries of the meshes, which is essential for aligning garments with large shape variance.

For each sequence, one frame of registered garment mesh $R' \in \mathbb{R}^{X\times 3}$ is then used to calculate barycentric interpolation weights against the vertices of template mesh T, which is further used to re-mesh garments observed by other frames. Specifically, for each vertex x of T, we can find the nearest face f in R'. Afterwards, we project x to f and then calculate the barycentric coordinate (w_1, w_2, w_3) of the projected point x' on f, which satisfies $x' = w_1 f_1 + w_2 f_2 + w_3 f_3$. Using the barycentric coordinates, it is straight-forward to re-mesh all the other frames of garment meshes to the topology of the template mesh T.

2.3 Canonical-Posed Garment Estimation

To estimate canonical-posed garments $T(\alpha)$, we predict the PCA coefficients α by using the observed 3D points belonging to garments from input point cloud sequences $P \in \mathbb{R}^{T\times p\times 3}$, where T is the sequence length. As shown in Fig. 3 (a), we first perform point-wise semantic segmentation to select the 3D points of interest (*i.e.* garments) for each frame, which results in a sequence of garment point

clouds $P_g \in \mathbb{R}^{T \times p_g \times 3}$. Subsequently, P_g is employed to predict the PCA coefficients α for each sequence, which is further used to reconstruct target canonical-posed garments by $T(\alpha) = G + C\alpha$.

2.4 Posed Garment Reconstruction

Proposal-Guided Hierarchical Feature Network. As introduced in Subsection 2.1, the skinned garments $M_p = W(T(\alpha), J, \theta, \mathcal{W}_g)$ of the proposed garment model can provide reasonable initial proposals for reconstructing posed garments. However, due to complex human motions and garment flexibility, especially for intense movements and loose skirts, the initial garment poses can be far different from real poses. To address these problems, we propose a novel proposal-guided hierarchical feature network, consisting of two major parts: 1) hierarchical garment feature pooling, and 2) hierarchical body surface encoder, which predicts the displacements D to refine the initial garment poses. Detailed architectures of these two modules are demonstrated in Fig. 3 (b). The proposed hierarchical garment feature pooling can effectively capture local geometric details from garments. Furthermore, hierarchical body surface encoder facilitates predicting correct garment poses by considering interactions between garment and human body. By estimating proper contacts between garments and skin, our body surface encoder highly alleviates the interpenetration problems.

- **Hierarchical garment feature pooling** aims at pooling accurate geometric features and rich semantic information for each vertex of mesh proposals. Specifically, for each vertex of M_p, several query balls with different radiuses are constructed to sample points from the segmented point clouds of garments. To make use of the rich semantic information extracted when predicting the PCA coefficients α, different levels of down-sampled garment point clouds $p_g = \{p_{g1}, p_{g2}, p_{g3}\}$ and the corresponding features $f_g = \{f_{g1}, f_{g2}, f_{g3}\}$ from the previous stage are used for the query process with radiuses $r = \{r_1, r_2, r_3\}$ and sample number $n = \{n_1, n_2, n_3\}$. Then for each vertex, the pooled garment features are further processed by MLP, max pooling and concatenation to aggregate low-level geometric features and high-level semantic information.

- **Hierarchical body surface encoder** makes each vertex of the garment mesh aware of its neighbouring human body surfaces, which is important for correct interaction between garment and human body, especially under the circumstances of large body movements. For each vertex of M_p, neighbouring human mesh vertices are sampled using different radiuses of query balls. Both coordinates and normal directions of sampled human mesh vertices are encoded using MLP.

Iterative GCN. Concatenating the per-vertex features obtained from the above feature pooling and surface encoder, along with the coordinates of the vertex, we have aggregated necessary information for the following iterative Graph Convolution Network (GCN) to process and output the displacement prediction D. The loose garments skinned by the proposed skinning method are normally not in accordance with the real situation, which will be shown in experiments. Therefore, we choose to iteratively predict displacements and gradually approach the ground truth. As shown in Fig. 3 (b), for the j-th iteration, using the previous accumulated displacement predictions $\sum_{i=1}^{j-1} D_i$ and the proposal M_p, we can calculate current mesh prediction by $M_{pj} = M_p + \sum_{i=1}^{j-1} D_i$. M_{pj} is then used for the proposal-guided hierarchical feature pooling of this iteration. The last GCN layer of each iteration outputs current displacement D_j. Inspired by [19], the GCN layer could be formulated as $H^{(l+1)} = \text{ReLU}(\bar{D}^{-1}\bar{A}H^{(l)}W^{(l)} + B^{(l)})$, where $\bar{A} = A + I$ is the bidirectional mesh graph with self-connections, $\bar{D} = \text{diag}(\sum_j \bar{A}_{ij})$, $W^{(l)}$ and $B^{(l)}$ are the learn-able weights and bias.

Temporal Transformer. To support better temporal information extraction, a transformer structure is integrated into the iterative GCN. For the j-th iteration, we take the features extracted by the $(j-1)$-th iteration, perform temporal fusion by transformer and concatenate to the input of the j-th iteration. Specifically, we denote the output feature of the second last layer of the $(j-1)$-th iteration as $F_{j-1} \in \mathbb{R}^{T \times M \times K}$. Query, key and value are obtained by applying MLP to F_{j-1}, which could be denoted as $Q, K, V = \text{MLP}(F_{j-1})$, where $Q, K, V \in \mathbb{R}^{T \times MK}$. Then the value is updated as $V' = \text{softmax}(QK^T/\sqrt{T})V$. V' is then concatenated to the pooled features of the j-th iteration.

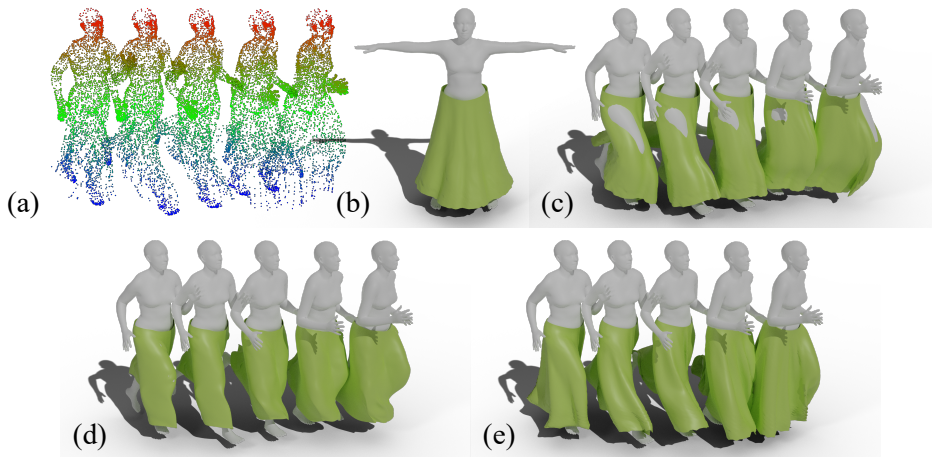


Figure 4: Visualization of the Garment Reconstruction Process from Point Cloud Sequence. **(a)** Input point cloud sequence; **(b)** The canonical-posed garment estimation; **(c)** Results of the interpolated LBS; **(d)** Final reconstruction results; **(e)** Ground truth.

2.5 Loss Functions

Loss functions corresponding to two parts of network are introduced in this section. For the canonical-posed garment estimation, the loss function consists of five terms, which could be formulated as

$$\begin{aligned} L_c =& \lambda_1 \text{CE}(logits, label_{gt}) + \lambda_2 |\alpha_{pred} - \alpha_{gt}|_2^2 + \lambda_3 \sum_{i=1}^{M} \frac{1}{M} |T_i(\alpha_{pred}) - T_i(\alpha_{gt})|_2^2 + \\ & \lambda_4 \text{IL}(T_h, T(\alpha_{pred})) + \lambda_5 \text{LR}(T(\alpha_{pred}), T(\alpha_{gt}), L). \end{aligned} \tag{5}$$

The first term CE is the cross entropy term for semantic segmentation. The second and third terms are L2 loss terms for PCA coefficients and mesh vertices. The fourth term is the interpenetration loss term, which could be formulated as

$$\text{IL}(T_h, T(\alpha_{pred})) = \sum_{i=1}^{M} \frac{1}{M} \text{ReLU}(-(N_{hi} \cdot (T_i(\alpha_{pred}) - V_{hi}))), \tag{6}$$

where V_{hi} and N_{hi} are the coordinate and normal vector of the nearest human mesh vertex of the i-th garment mesh vertex $T_i(\alpha_{pred})$. The fifth term is the Laplacian regularization term,

$$\text{LR}(T(\alpha_{pred}), T(\alpha_{gt}), L) = \sum_{i=1}^{M} \frac{1}{M} (|LT(\alpha_{pred})|_2 - |LT(\alpha_{gt})|_2), \tag{7}$$

where L is the cotangent weighted Laplacian matrix of the garment mesh.

For the j-th iteration of the posed garment reconstruction, the loss function includes four terms, which are defined as

$$L_{pj} = \lambda_6 \sum_{i=1}^{M} \frac{1}{M} |M_{pred} - M_{gt}|_2^2 + \lambda_7 \text{IL}(M_h, M_{pred}) + \lambda_8 \text{LR}(M_{pred}, M_{gt}, L) + \lambda_9 \text{TC}(M_{pred}). \tag{8}$$

The first term is the L2 loss term between predicted and ground truth garment mesh vertices M_{pred} and M_{gt}. The second term penalizes interpenetration between posed garment prediction M_{pred} and posed human body M_h. The third term performs cotangent weighted Laplacian regularization. The last term adds constraints to the temporal deformations of the predicted garments, which could be formulated as

$$\text{TC}(M_{pred}) = \sum_{t=2}^{T} \frac{1}{T-1} |M_{pred}(t) - M_{pred}(t-1)|_2^2. \tag{9}$$

Adding up weighted losses L_{pj} of each iteration, we get the loss term of the posed garment reconstruction $L_p = \sum_j w_j L_{pj}$.

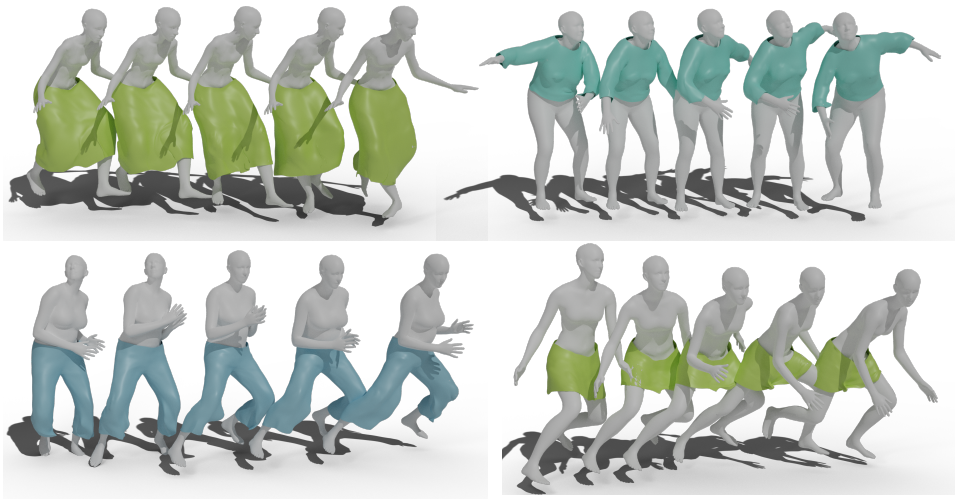


Figure 5: Qualitative Reconstruction Results. Reconstructed sequences of skirts, t-shirt and trousers are shown in different colors. High reconstruction quality is achieved even with large body movements.

3 Experiments

3.1 Dataset and Evaluation Protocols

Dataset. We establish point cloud sequence-based garment reconstruction dataset by adapting CLOTH3D [18] for our experiments. CLOTH3D[1] is a large scale synthetic dataset with rich garment shapes and styles and abundant human pose sequences. Specifically, we sample point sets from 3D human models to produce the point cloud sequence inputs. We select three types of garments, *i.e.* Skirts, T-Shirt, Trousers, for the experiments. For skirts, 453 sequences containing 109629 frames of point clouds are used for training and testing. For t-shirts, 394 sequences with 95540 frames are used. For trousers, 1465 sequences and 357086 frames are used. We split the sequences to training and testing sets at the ratio of $8:2$.

Evaluation Metrics. Three metrics are utilized for evaluation. We first evaluate the per-vertex L2 error of canonical-posed garments, $M_1 = \sum_{s,i} \frac{1}{MS} |T_{s,i}(\alpha_{pred}) - T_{s,i}(\alpha_{gt})|_2$. Secondly, we evaluate the per-vertex L2 error of posed garments, which is defined as $M_2 = \sum_{f,i} \frac{1}{MF} |M^{f,i}_{pred} - M^{f,i}_{gt}|_2$. The last one is the acceleration error, which is used to evaluate the smoothness of the predicted sequences and could be formulated as $M_3 = \sum_{f,i} \frac{1}{MF} |A^{f,i}_{pred} - A^{f,i}_{gt}|_2$, where $A^{f,i}$ stands for acceleration vector of the i-th vertex of f-th frame.

Comparison Methods. Since it is the first work tackling garment reconstruction from point cloud sequences, there is no readily available prior work for direct comparison. Here, for canonical-posed garment estimation, we adapt the PointNet++ [20] structure to predict the PCA coefficients directly from input point clouds. For posed garments reconstruction, we adapt Multi-Garment Net [16] (MGN) on point cloud inputs for comparison.

3.2 Qualitative Results

Step-by-Step Visualization. As shown in Fig. 4, we perform a step-by-step visualization of the whole reconstruction process. The estimated canonical-posed skirt recovered rough shape and length of the skirt from the point cloud sequences. As expected, the interpolation LBS produces reasonable proposals with smooth surfaces and no artifacts. There are two things that make the LBS results look unnatural. One is the interpenetration between garments and huamn body. The other is the lifted skirt floating in the air, not touching the leg, which should be the cause of the lift. These two flaws are inevitable because the skirt is not homotopy with human body. These defects would be

[1] We download the data from http://chalearnlap.cvc.uab.es/dataset/38/description/. The terms and conditions are included in the website, permitted by which, we use the data only for research purpose.

Class	PN++	Ours
Skirt	33.71	**30.23**
T-Shirt	41.05	**28.39**
Trousers	20.80	**18.16**

Table 2: L2 Error of Canonical-posed Garment Estimation. PN++ refers to [20]. $[mm]$

Class	MGN*	Int. LBS	Ours
Skirt	71.66	79.52	**49.39**
T-Shirt	40.54	44.55	**37.83**
Trousers	29.12	31.05	**26.48**

Table 3: L2 Error of Posed Garment Reconstruction. MGN* is adapted from [16]. Int. LBS is the results of interpolated LBS. $[mm]$

Class	MGN*	Ours
Skirt	12.59	**3.273**
T-Shirt	7.173	**4.749**
Trousers	4.135	**2.520**

Table 4: Acceleration Error of Posed Garment Reconstruction. $[m/s^2]$

eliminated finally. The final step of displacement prediction meets the expectations of three aspects. Firstly, the hierarchical garment feature pooling module captures the geometric features of cloth surfaces, which helps the precise shape recovery of posed garments. Secondly, the human body surface encoding module makes the network aware of the body poses and surface positions, which prevents the interpenetration of garments and human body and helps the capture of the deformations caused by body movements, *e.g.* knees lifting up the skirt. Thirdly, the temporal fusion module encourages smooth deformation between consecutive frames.

Reconstruction for Different Garment Types. More visualization of the reconstruction results are illustrated in Fig. 5, which shows Garment4D's ability of handling both loose and tight garments driven by large body movements. For garments that are homotopy to human body, *i.e.* t-shirt and trousers, Garment4D can not only capture correct shape of them, but also the dynamics caused by body movement. Taking the bottom left trousers reconstruction as an example, when the right leg is stretched out, the back side of trousers stick to the back of calf. The front part of trousers naturally form the dynamics being pulled backwards. Moreover, it can be observed that when the leg or arm bends, more of the ankle or forearm are exposed in order to maintain the length of trousers legs or sleeves. For the reconstruction of skirts which are not homotopy to human body, both long and short skirts can be faithfully recovered. The short skirt is tighter and we can observe the natural deformation when the two legs are stretched to sides. For the looser long skirt, clear contacts and deformations can be observed when knees and calves lift. And when the legs are retracted to the neutral position, skirts would naturally fall back but would not stick to leg surfaces.

3.3 Quantitative Results

Canonical-Posed Garment Estimation. Tab. 2 reports the per-vertex L2 error of canonical-posed garment estimation of the adapted PointNet++ [20] and our method. Explicitly making the network aware of the semantics by cropping out the garments of interest out-performs the vanilla PointNet++ in all three garment types. Although body poses also matters in garment estimation, the experiment shows that it is better to focus more on the garments-of-interest at this stage.

Posed Garment Reconstruction. The per-vertex L2 errors of posed garment reconstruction are shown in Tab. 3. Because t-shirt and trousers are homotopy to human body, the interpolated LBS step could give decent reconstruction results. In contrast, for skirts, the results of interpolated LBS is far from the real situation and therefore have high L2 error, which is also clearly analyzed in the above qualitative results. The final step of displacement prediction effectively captures the dynamics of skirts and improves $30.13\ mm$ on top of the results of interpolated LBS. Moreover, our Garment4D outperforms the adapted Multi-Garment Net [16] in all three garment types. Especially, Garment4D surpasses the adapted MGN in skirt by $22.27\ mm$, which further shows the advantage of Garment4D over SMPL+D-based methods at reconstructing loose garments like skirts.

Temporal Smoothness. As reported in Tab. 4, our method also outperforms the adapted MGN in terms of reconstruction smoothness. Even though the same temporal constraints term is included in the loss functions of both methods, unlike Garment4D, the adapted MGN still fails to properly model the motion of garments and aggregate temporal information. Particularly for loose garments like skirts, Garment4D outperforms the adapted MGN by $9.317\ m/s^2$.

3.4 Ablation Study

Hierarchical Body Surface Encoder. As shown in Fig. 7, without the hierarchical body surface encoder (denoted as H.B.S Encoder in figure), the per-vertex L2 error increases by $8.42\ mm$, which is in line with the expectations. Without the knowledge of body surface positions, it is hard for networks to infer garment-body interactions.

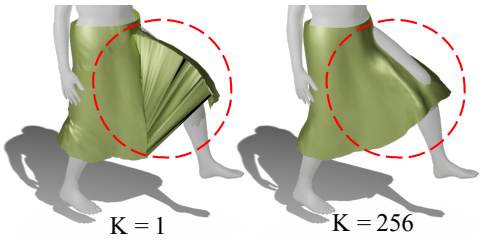


Figure 6: Comparison of LBS results when nearest neighbour K is set to 1 and 256. Severe artifacts appear when $K = 1$, which is circled out.

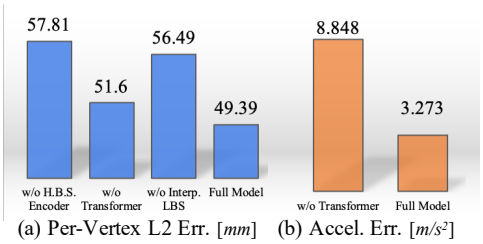


Figure 7: Results of Ablation Study. (a) shows the comparison of Per-vertex L2 error of posed garment reconstruction. (b) reports the acceleration error.

Temporal Transformer. As illustrated in Fig. 7, if we remove the temporal transformer from the network, the per-vertex L2 error increases by $2.21\ mm$ and the acceleration error increases by $5.575\ m/s^2$. The results indicate that the temporal transformer not only helps the single frame reconstruction by aggregating information from neighbouring frames, but also, more importantly, increases the smoothness of the reconstruction sequences.

Interpolated LBS. As circled out in the left side of Fig. 6, if the skinning weights of garment vertices are copied from the nearest body vertex, which is how SMPL+D methods (*e.g.* MGN[16]) deal with the garments, the severe cliff-like artifacts appear between the legs. This is because both halves of skirt vertices find the nearest vertices in the corresponding leg and the tear happens at the left and right boundary. Setting the number of nearest neighbours K to 256, the right side of Fig. 6 demonstrates the smooth transition between the left and right half of the skirt. The quality of the LBS proposals would effect the final reconstruction results, which is shown in Fig. 7. If $K = 1$, the per-vertex L2 error of the final reconstruction results would increase by $7.1\ mm$.

4 Related Work

Parametric Clothed Human Modeling. Most previous garment reconstruction work that use parametric methods tend to take human body and garment as a whole. A common practice, SMPL+D [7, 8, 21–24], is to predict displacements from SMPL [25] vertices. However, SMPL+D has difficulty modeling loose garments or garments that do not share same topology as human body, which limits the expression ability of SMPL+D.

Parametric Garment Modeling. In order to reconstruct garments separately using parametric models, PCA decomposition is used by [16, 17] to model the single layered garment meshes, which is independent of any parametric human body model. For each type of garment, a single PCA model has to be constructed, which is a clear limitation of PCA modeling. To further increase the expressive ability of an individual parametric garment model, [26] maps garments to two types of UV-maps according to whether the garment is homotopy to human body. As for the deformation of the garment models, most existing methods rely on the skinning process of human body. [16] samples the skinning weights of garment vertices from the closest human body vertices, which is not suitable for garment meshes with high resolution or garments not homotopy with human body. [17] predicts skinning weights of garment vertices using neural network. Only relying on the skinning process would lead to overly smoothed results, which is addressed by [27] through additional displacement prediction before skinning.

5 Conclusion

In this work, we present the first attempt at garment reconstruction from point cloud sequences. A garment model, with the novel interpolated LBS, is introduced to adapt to loose garments. A garment mesh registration method is proposed to prepare sequential meshes for the garment reconstruction network. Taking point clouds sequences as input, the corresponding canonical-posed garment is estimated by predicting PCA coefficients. Then a proposal-guided hierarchical feature network is introduced to perform semantic, geometric and human-body-aware feature aggregation. Efficient iterative GCN along with temporal transformer is utilized to further encode the features and predict displacements in an iterative way. Qualitative and quantitative results from extensive experiments demonstrate the effectiveness of the proposed framework.

References

[1] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum*, volume 38, pages 355–366. Wiley Online Library, 2019. 1

[2] Pascal Volino, Christiane Luible, and Nadia Magnenat-Thalmann. Virtual clothing. *Wiley Encyclopedia of Computer Science and Engineering*, 2007. 1

[3] Simon Spielmann, Volker Helzle, and Rahul Nair. On-set depth capturing for vfx productions using time of flight. In *ACM SIGGRAPH 2013 Talks*, pages 1–1. 2013. 1

[4] 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

[5] Aaron S Jackson, Chris Manafas, and Georgios Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[6] 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.

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 1, 2, 3, 9

[8] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018. 1, 2, 3, 9

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

[10] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 4480–4490, 2019.

[11] 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/CVF International Conference on Computer Vision*, pages 7750–7759, 2019.

[12] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2232–2241, 2019.

[13] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. Facsimile: Fast and accurate scans from an image in less than a second. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5330–5339, 2019.

[14] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. *arXiv preprint arXiv:2103.06871*, 2021.

[15] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. *arXiv preprint arXiv:2007.11432*, 2020. 1

[16] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5420–5430, 2019. 1, 2, 3, 7, 8, 9

[17] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision*, pages 18–35. Springer, 2020. 1, 2, 9

[18] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: Clothed 3d humans. In *European Conference on Computer Vision*, pages 344–359. Springer, 2020. 2, 7

[19] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 5

[20] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017. 7, 8

[21] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 9

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

[23] Akihiko Sayo, Hayato Onizuka, Diego Thomas, Yuta Nakashima, Hiroshi Kawasaki, and Katsushi Ikeuchi. Human shape reconstruction with loose clothes from partially observed data by pose specific deformation. In *Pacific-Rim Symposium on Image and Video Technology*, pages 225–239. Springer, 2019.

[24] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4491–4500, 2019. 9

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015. 9

[26] Zhaoqi Su, Tao Yu, Yangang Wang, Yipeng Li, and Yebin Liu. Deepcloth: Neural garment representation for shape and style editing. *arXiv preprint arXiv:2011.14619*, 2020. 9

[27] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7365–7375, 2020. 9

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 4.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Because reconstruction is a well established computer vision task.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] However, we will release the related assets upon paper acceptance.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please kindly refer to the supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We empirically find that the results are stable especially with large dataset. Moreover, the time and computational resources required for performing the experiments makes repeated runs impractical.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please kindly refer to the supplemental material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 3.1.
   (b) Did you mention the license of the assets? [Yes] See Section 3.1.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The extension to the existing dataset will be released upon paper acceptance.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section 3.1.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] It is a synthetic dataset, which means no personally identifiable information is used.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]