Learning Anchored Unsigned Distance Functions with Gradient Direction Alignment for Single-view Garment Reconstruction

Fang Zhao[1]   Wenhao Wang[2]   Shengcai Liao[1*]   Ling Shao[1,3]

[1]Inception Institute of Artificial Intelligence   [2]ReLER, University of Technology Sydney

[3]Mohamed bin Zayed University of Artificial Intelligence

Abstract

While single-view 3D reconstruction has made significant progress benefiting from deep shape representations in recent years, garment reconstruction is still not solved well due to open surfaces, diverse topologies and complex geometric details. In this paper, we propose a novel learnable Anchored Unsigned Distance Function (AnchorUDF) representation for 3D garment reconstruction from a single image. AnchorUDF represents 3D shapes by predicting unsigned distance fields (UDFs) to enable open garment surface modeling at arbitrary resolution. To capture diverse garment topologies, AnchorUDF not only computes pixel-aligned local image features of query points, but also leverages a set of anchor points located around the surface to enrich 3D position features for query points, which provides stronger 3D space context for the distance function. Furthermore, in order to obtain more accurate point projection direction at inference, we explicitly align the spatial gradient direction of AnchorUDF with the ground-truth direction to the surface during training. Extensive experiments on two public 3D garment datasets, i.e., MGN and Deep Fashion3D, demonstrate that AnchorUDF achieves the state-of-the-art performance on single-view garment reconstruction. Code is available at `https://github.com/zhaofang0627/AnchorUDF`.

1. Introduction

3D garment reconstruction has a wide range of applications in clothed human digitization, virtual try-on, online shopping and so on. Recently, image based 3D reconstruction has made significant progress benefiting from shape representation learning with deep neural networks [14, 16, 19, 35, 26]. Compared to voxels, points and meshes, implicit functions, which define a surface as a level set of a function, can represent 3D surfaces at arbitrary resolu-

tion and produce fine-scale detailed surfaces in a memory-efficient way, and have been successfully applied to single-view human reconstruction [32, 33]. However, recovering 3D garment shape from a single image is still a challenging task because garments have open surfaces and diverse topologies in addition to complex geometric details like clothed humans.

Existing methods usually use parametric models to provide shape priors of garments [11, 22, 40]. BCNet [22] introduces a layered garment representation on top of the SMPL body model [25] and a generic skinning weights generating network to improve the expression ability of the garment model. Deep Fashion3D [40] proposes adaptable template meshes to fit garment shapes and incorporates implicit representations to refine surface details. These methods rely on pre-defined category-specific templates and have poor scalability to new garment categories.

In this paper, we propose Anchored Unsigned Distance Function (AnchorUDF), a learnable unsigned distance function that enriches 3D position features of query points by anchor points around 3D surfaces. With AnchorUDF, we establish a unified shape learning framework to reconstruct 3D garment from a single image. As shown in Fig. 1, our method can handle non-closed garment surfaces, capture multiple topologies of different garment categories while retaining fine-scale geometric details.

Specifically, for a 3D query point, like PIFu [32], we first compute its pixel-aligned local image features. However, instead of using an absolute depth value to encode the 3D position of the query point, we leverage a set of anchor points representing 3D shape profile to enrich its position feature to make the distance function better sense the garment topologies. To obtain a small number of anchor points which can adequately cover the surface, we cluster points sampled from the surface with k-means and use clustering centers as targets to learn a regression network on top of the backbone to predict anchor points. Note that compared to the full shape, these few anchor points (typically a few hundred) are easier to estimate. A 3D convolutional

---

*Corresponding author.

Figure 1. Single-view garment reconstruction using our method, which handles non-closed garment surfaces, captures multiple topologies of different garment categories and retains fine-scale geometric details.

network is further employed to compute 3D feature tensors from anchor points and a 3D position feature vector for the query point can be extracted via trilinear interpolation with its 3D coordinates, which encodes the relative position relationship between query and anchor points and provides stronger 3D space context information compared to the absolute depth value. Different from signed distance functions [29], we need to compute the gradient field of unsigned ones at inference to project the query point onto the surface along the negative gradient direction. Thus, in order to obtain more accurate estimation for the projection direction, we explicitly constrain the spatial gradient of our AnchorUDF during training to align its direction with the ground-truth direction to the surface.

Our contributions can be summarized as follows: 1) We propose a unified shape learning framework for single-view garment reconstruction; 2) We introduce AnchorUDF, a learnable unsigned distance function with anchored 3D position features; 3) We propose to learn the unsigned distance function with gradient direction alignment to more accurately project query points at inference.

2. Related Work

Single-view garment/clothed body reconstruction. Current work can generally be divided into template-based and template-free methods. For the former one, parametric 3D models [6, 25, 30] are used to provide strong priors for constraining the solution space of shape estimation. For better geometric detail representation, a high-frequency displacement is usually computed on the basis of the mesh model [15, 4, 38, 3, 41, 40, 22]. DeepWrinkles [24] jointly represents global shape deformation and surface details by adding fine clothing wrinkles onto normal maps of a coarse garment mesh. Tex2shape [5] regards shape regression as an aligned image-to-image translation problem and estimates detailed normal and vector displacement maps from a partial texture, which can be applied to a body model to add detail and clothing. MGN [11] predicts PCA coefficients of garment parametric models and a displacement field on top of PCA for clothing details. However, these template-based methods usually are limited by topologies of pre-defined garment template meshes and are hard to handle out-of-scope deformations.

Some template-free methods which do not use parametric models have been proposed to directly regress 3D shapes for modeling more complex garment topologies [34, 2, 32]. Bodynet [34] infers volumetric body shape from a single image by an end-to-end trainable network with intermediate supervision of pose and body part segmentation. Deep-Human [39] introduces an image-guided volume-to-volume translation CNN by taking a semantic body volume as an additional input. These methods based on voxel representations require high memory and often fail to obtain fine-scale shape details. Based on implicit function representations, PIFu [32] proposes a memory-efficient deep learning framework for clothed body reconstruction, which locally aligns 2D image pixels with the global context of 3D objects to retain more shape details. More recently, PIFuHD [33] formulates a multi-level architecture based on PIFu, where a coarse level focuses on holistic reasoning and a fine level estimates highly detailed geometry. Geo-PIFu [20] extends PIFu to learn latent voxel features using a structure-aware 3D U-Net for less shape distortion and sharper surface details. There also exist some methods [10, 21] that combine implicit functions and parametric models to obtain both detailed and controllable 3D body reconstructions.

Differences from related reconstruction methods. PIFuHD [33] uses larger input image resolution and extra normal maps for higher-fidelity reconstruction. Our method can also incorporate the HD module to further improve the reconstruction performance as demonstrated in our experi-
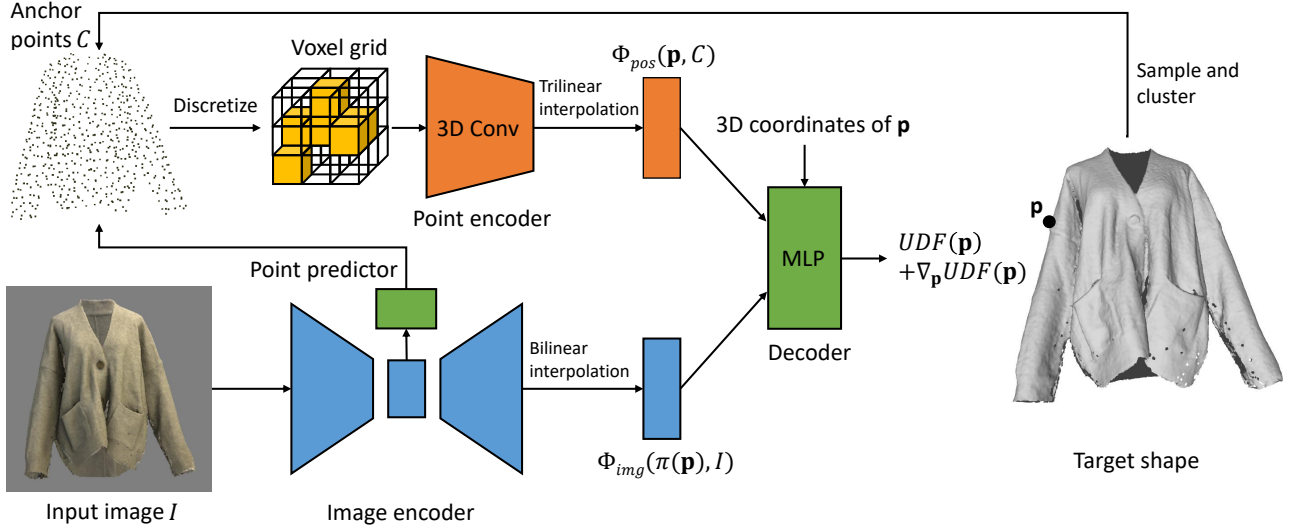
Figure 2. Framework of our proposed method. AnchorUDF predicts the unsigned distance field (UDF) of surface by using both local image features $\Phi_{img}(\pi(\mathbf{p}), I)$ and 3D position features $\Phi_{pos}(\mathbf{p}, \mathcal{C})$ based on a set of anchor points. During training, we explicitly constrain the spatial gradient of AnchorUDF to align its direction with the ground-truth direction to the surface.

ments. ARCH [21] requires a body mesh model to obtain body landmarks, which is hard to extend to garments which have multiple topologies. Geo-PIFu [20] uses an extra 3D backbone network to regress dense 3D volumes, which is computationally intensive for both training and testing. In contrast, our method does not rely on external parametric models, and only adds a lightweight head net on the top of backbone for anchor point prediction and a small 3D convolutional network for point feature extraction. Besides, the most important point is that the aforementioned methods are only able to generate closed surfaces and thus cannot handle open boundaries of garments.

Implicit surface representation. Compared to voxel [37, 36, 14, 17], point [16, 1] and mesh [23, 31] representations, implicit functions allow to represent 3D surfaces at infinite resolution without excessive memory cost [26, 29, 27, 7, 13]. OccNet [26] proposes to encode the 3D surface as the decision boundary of a deep neural network classifier. DeepSDF [29] introduces a learned continuous signed distance function for a class of shapes, which implicitly represents a shape's boundary as the zero level-set of the learned function. IF-Nets [12] learn implicit functions by extracting and classifying deep features extracted from a 3D grid of multi-scale features at a continuous query point. SAL [7] defines a family of loss functions for sign agnostic learning with raw geometric data. IGR [18] encourages the neural network to vanish on input points and has a unit norm gradient to learn smooth and natural implicit surfaces. To enable implicit functions to model open surfaces, NDF [13] proposes to predict unsigned distance fields (UDFs) to represent open surfaces for point cloud completion. However, it is nontrivial to learn UDFs for image based garment reconstruction due to lack of 3D space information. SALD [8]

includes derivatives during sign agnostic learning, which leads to a lower sample complexity and better fitting. Different from its motivation and loss form, we aim to optimize the gradient direction of UDF to obtain more accurate point projection direction at inference.

3. Method

Given a single image, we aim to recover the 3D garment shape with open surfaces and fine-scale surface details. In this section, we propose Anchored Unsigned Distance Function (AnchorUDF) which predicts the unsigned distance field (UDF) of surface with anchored 3D position features. To project query points onto the surface along more accurate direction at inference, we also explicitly align the spatial gradient direction of AnchorUDF with the ground-truth direction to the surface. The framework of our proposed method is illustrated in Fig. 2.

3.1. Unsigned Distance Fields

To represent shapes with open surfaces, we adopt unsigned distance fields (UDFs) [13] which assign a non-negative scalar value s to a spatial point $\mathbf{p}$:

$$UDF(\mathbf{p}) = s : \mathbf{p} \in \mathbb{R}^3, s \in \mathbb{R}_0^+, \qquad (1)$$

where s represents the unsigned distance from $\mathbf{p}$ to the closest surface and the shape surface is implicitly represented by the zero level-set $UDF(.) = 0$. In contrast to signed distance fields (SDFs) [29] or occupancies [26] which divide the 3D space into inside and outside the surface, UDFs allow to naturally represent open surfaces. We can project $\mathbf{p}$ onto the surface by moving $\mathbf{p}$ along the negative gradient

direction of UDF:

$$\mathbf{q} := \mathbf{p} - UDF(\mathbf{p}) \cdot \nabla_{\mathbf{p}} UDF(\mathbf{p}). \qquad (2)$$

Dense point clouds are able to be easily computed from the implicit surfaces using fast gradient evaluation for UDFs [13] and naive classical algorithms for meshing [9] can be used to generate the corresponding meshes.

3.2. Anchored Unsigned Distance Functions

It is worth noting that regressing a continuous distance field is much more difficult than classifying a binary occupancy value. Moreover, different from SDFs, UDFs need gradient direction computation at inference. Besides a point itself, the prediction accuracy for the neighborhood of the point is also required to ensure the correct gradient direction. Thus, discriminative features are more critical for learning a good UDF.

To this end, our AnchorUDF employs both local image features and 3D position features of query points to predict UDF values. The key idea is to leverage a set of anchor points located around 3D surfaces to compute relative position features for query points, which makes the distance function better fit the garment topologies.

Given an input image I, a fully convolutional image encoder is first used to compute a feature map of I. For a 3D query point $\mathbf{p}$, we project it onto its corresponding position on the image plane to compute its local image features from the feature map by bilinear interpolation at the projected pixel coordinates: $\Phi_{img}(\pi(\mathbf{p}), I)$, where π represents the (weak) perspective camera projection.

In order to compute more discriminative 3D position features for the query point $\mathbf{p}$, a set of anchor points $\mathcal{C}$ located around the surface is introduced to encode the 3D position of $\mathbf{p}$. We then discretize the anchor points $\mathcal{C}$ to a voxel grid and feed the grid into a point encoder consisting of a series of 3D convolutional layers to produce a 3D feature tensor of $\mathcal{C}$. By applying trilinear interpolation to the feature tensor according to the 3D coordinates of $\mathbf{p}$, we extract the feature vector $\Phi_{pos}(\mathbf{p}, \mathcal{C})$ as the position features of the query point. Compared to using an absolute depth value to provide the 3D position information [32], our position features reflect relative position relationship between query and anchor points and provides stronger 3D space context to make UDF better capture multiple topologies.

To obtain a set of anchor points which can adequately cover the surface only using a small number of points, we cluster points sampled from the ground-truth surface with k-means and use clustering centers $\tilde{\mathcal{C}}$ as targets to learn a regression network on top of the backbone to predict the anchor points. Here the anchor point prediction and the shape reconstruction are learned jointly, which can be seen as multi-task learning since they share the backbone. We argue that compared to the full shape, these few anchor points
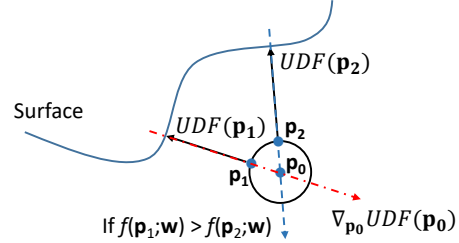

Figure 3. Motivation of gradient direction alignment. $\mathbf{p}_1$ and $\mathbf{p}_2$ are two points on the neighborhood of $\mathbf{p}_0$. When $UDF(\mathbf{p}_1)$ and $UDF(\mathbf{p}_2)$ are close, even if the distance loss is small, it is also possible that $f(\mathbf{p}_1; \mathbf{w}) > f(\mathbf{p}_2; \mathbf{w})$. In this case, the predicted gradient direction at $\mathbf{p}_0$ (the blue arrow) will be significantly different from the direction of $\nabla_{\mathbf{p}_0} UDF(\mathbf{p}_0)$ (the red arrow).

(typically a few hundred) are easier to estimate. Therefore, our method can also be regarded as a coarse-to-fine strategy, i.e., a shape profile is first predicted, then further refined to produce the detailed surface.

At last, we concatenate the local image features, the 3D position features and the 3D coordinates of the query point as the input of a decoder to predict UDF values and our AnchorUDF is formulated as:

$$f_I(\mathbf{p}, \mathcal{C}; \mathbf{w}) = f_{dec}([\Phi_{img}(\pi(\mathbf{p}), I), \Phi_{pos}(\mathbf{p}, \mathcal{C}), \mathbf{p}]), \quad (3)$$

where $\mathbf{w}$ denotes network parameters and the decoder f_{dec} is parameterized by multi-layer perceptrons (MLP) with ReLU in its last layer.

3.3. Learning with Gradient Direction Alignment

To learn our AnchorUDF, for an input image I, we generate training examples $\mathcal{P}_I$ by sampling points near its corresponding surface and computing their ground-truth UDF values $UDF(\mathbf{p})$. Then, we minimize the L1 loss between the predicted and ground-truth UDF values on $\mathcal{P}_I$ by updating parameters $\mathbf{w}$ of $f_I(\mathbf{p}, \mathcal{C}; \mathbf{w})$:

$$L_{UDF} = \sum_{\mathbf{p} \in \mathcal{P}_I} |\min(f_I(\mathbf{p}, \mathcal{C}; \mathbf{w}), \delta) - \min(UDF(\mathbf{p}), \delta)|.$$
$$(4)$$

Similar to [13, 29], a small value δ is used to clamp the maximal regressed distance to concentrate the model capacity on details in the vicinity of the surface.

In order to learn the anchor point predictor, we consider a loss defined by the Chamfer distance [16] between the predicted anchor points $\mathcal{C}$ and the targets $\tilde{\mathcal{C}}$ because the set of anchor points is unordered:

$$L_{AP} = \sum_{\mathbf{c} \in \mathcal{C}} \min_{\tilde{\mathbf{c}} \in \tilde{\mathcal{C}}} ||\mathbf{c} - \tilde{\mathbf{c}}||_2^2 + \sum_{\tilde{\mathbf{c}} \in \tilde{\mathcal{C}}} \min_{\mathbf{c} \in \mathcal{C}} ||\mathbf{c} - \tilde{\mathbf{c}}||_2^2. \quad (5)$$

The standard back-propagation through AnchorUDF can be used to compute the spatial gradients of the learned distance field to project points by Eq. (2).
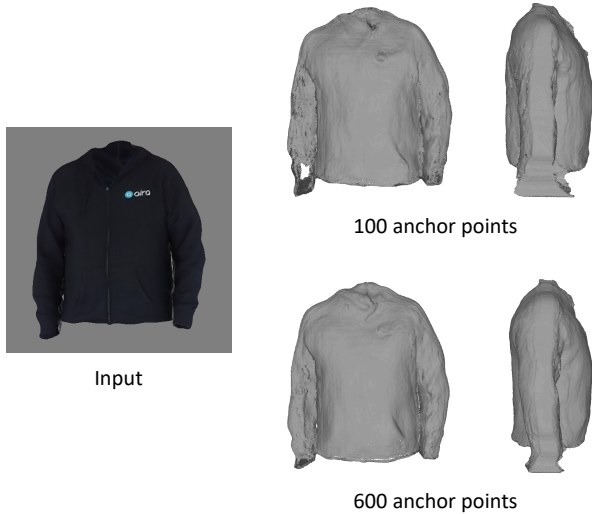
100 anchor points

Input

600 anchor points

Figure 4. Visual comparison of using different numbers of anchor points. More anchor points can better project query points to the surfaces

However, only optimizing the point-wise distance loss (4) cannot guarantee a good estimation for the true gradient directions of UDF. As illustrated in Fig. 3 (for the sake of brevity, here I and $\mathcal{C}$ in $f_I(\mathbf{p}, \mathcal{C}; \mathbf{w})$ are omitted), consider two points $\mathbf{p}_1$ and $\mathbf{p}_2$ on the neighborhood of a point $\mathbf{p}_0$, when the ground truth UDF values of $\mathbf{p}_1$ and $\mathbf{p}_2$ are close, because the distance loss does not penalize the relationship between points, even if the loss is small, it is also possible that $f(\mathbf{p}_1; \mathbf{w}) > f(\mathbf{p}_2; \mathbf{w})$. In this case, the predicted gradient direction at $\mathbf{p}_0$ will be significantly different from the ground truth one. Therefore, during training we explicitly constrain the spatial gradient of AnchorUDF to align its direction with the true gradient direction by the following loss:

$$L_{GDA} = \sum_{\mathbf{p} \in \mathcal{P}} 1 - \cos(\nabla_{\mathbf{p}} f(\mathbf{p}; \mathbf{w}), \nabla_{\mathbf{p}} UDF(\mathbf{p})). \quad (6)$$

In practice, the direction of $\nabla_{\mathbf{p}} UDF(\mathbf{p})$ can be computed by $(\mathbf{p} - \mathbf{q})/||\mathbf{p} - \mathbf{q}||_2$, where $\mathbf{q}$ is the point closest to $\mathbf{p}$ on the surface.

Finally, the overall objective function is

$$L = L_{UDF} + \lambda_1 L_{AP} + \lambda_2 L_{GDA}, \quad (7)$$

where λ_1 and λ_2 are loss weights.

4. Experiments

We evaluate the proposed AnchorUDF on two 3D garment datasets, i.e, MGN [11] and Deep Fashion3D [40]. We adopt the Chamfer distance and the average point-to-surface Euclidean distance (P2S) to measure the quality of shape reconstruction. We also compare AnchorUDF against
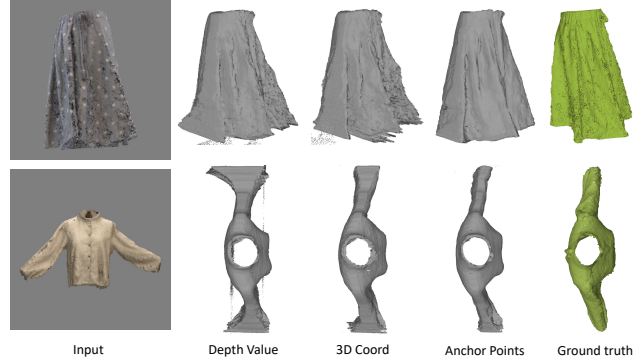

Figure 5. Visual comparison of using different 3D position features. Our anchor point based position features can better recover surfaces, especially along the depth direction of input images

Table 1. Chamfer and P2S errors ($\times 10^{-3}$) of using different numbers of anchor points (AP) on MGN dataset.

Num. AP	100	300	600	900
Chamfer	0.731	0.716	**0.696**	0.758
P2S	0.987	0.967	**0.899**	1.014

Table 2. Chamfer and P2S errors ($\times 10^{-3}$) using different 3D position features on MGN and Deep Fashion3D datasets.

Methods	MGN		Deep Fashion3D	
	Chamfer	P2S	Chamfer	P2S
Depth Value	1.063	1.974	1.411	2.440
3D Coord	0.757	1.060	1.062	1.748
Anchor Points	**0.696**	**0.899**	**0.712**	**0.932**

other single-view 3D reconstruction methods with different shape representations [35, 32, 22, 33].

Datasets. MGN [11] contains 5 garment categories and 154 textured garments models. 134 garments models are randomly selected as the training set and the remaining 20 models form the test set. Following PIFu, we render images with 360 degrees in yaw axis for each garment model and obtain 48,240 images for training. Deep Fashion3D [40] is a large-scale collection of 3D garment models with diverse shapes and poses. It consists of 2075 garment models covering 10 different categories and 598 instances. We use 1880 garment models for training and 195 models for testing, where the training and test sets have disjoint instances. Because the dataset does not release the corresponded multi-view real images, we only use rendered images as the training inputs. For each model, we render images by sampling 18 front viewpoints, resulting in 33,840 training images. Similar to [40], here we focus on front-view reconstruction.

Implementation Details. We adopt a stacked hourglass [28] network with 5 stacks as our backbone to encode images. For anchor point prediction, we add a 4-layer network on top of the backbone, which consists of 3 con-

Figure 6. Visual comparison of with and without gradient direction alignment. By aligning the spatial gradient direction of the distance field, we can obtain sharper geometric details for the visible region and less artifacts for the self-occluded region.

Table 3. Chamfer and P2S errors ($\times 10^{-3}$) with and without gradient direction alignment on MGN and Deep Fashion3D datasets.

Methods	MGN		Deep Fashion3D	
	Chamfer	P2S	Chamfer	P2S
w/o L_{GDA}	0.696	0.899	0.712	0.932
w/ L_{GDA}	**0.635**	**0.762**	**0.621**	**0.839**

volutional (Conv) layers and 1 fully connected (FC) layer. To extract point features, we use a 5-layer 3D full Conv network with the $32 \times 32 \times 32$ input grid resolution. The decoder is a 6-layer FC network with skip connections. Following [13], we sample 5,000 training query points for each input image by Gaussian sampling with mixed variances in the vicinity of the ground truth surface. The maximal regressed distance is set to 0.2. The weights of L_{AP} and L_{GDA} are set to 1.0 and 0.02, respectively. We compute dense point clouds from the learned implicit surfaces using the dense point cloud extraction algorithm introduced in [13]. Please refer to the supplemental materials for more details about training and inference.

4.1. Ablation Study

We validate the effectiveness of main components in our proposed AnchorUDF by both qualitative and quantitative evaluations. We first investigate the impact of the number of anchor points on MGN dataset. Table 1 lists Chamfer and P2S errors obtained by using 100, 300, 600 and 900 anchor points. It can be seen that using more anchor points achieves better results because stronger space context can be referenced by query points. However, when the point number exceeds a certain value, e.g., 900, the performance degrades. The reason is that too many points to regress increases the difficulty of learning the point prediction net-
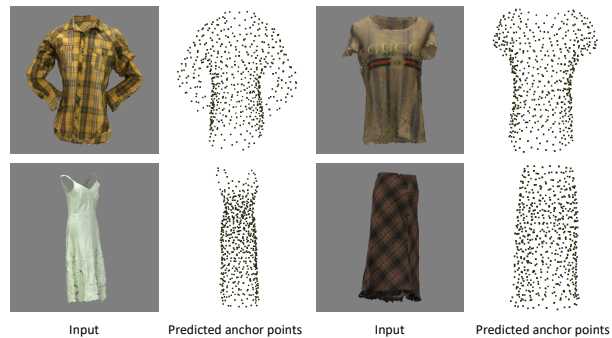

Figure 7. Visualization of anchor points predicted by our method.

work and reduces its stability when testing, which further affects the subsequent shape reconstruction. Fig. 4 shows that compared to 100 anchor points, using 600 points produces finer shapes, especially for the side view.

We further assess the importance of the proposed anchored 3D position features by applying different 3D position features to our reconstruction framework, including the depth value, 3D point coordinates and our anchor point based position features. As reported in Table 2, our position feature obtains the best results on both MGN and Deep Fashion3D. The depth value and 3D coordinates only provide absolute and local position information of query points, which is not enough to predict UDF with good neighborhood consistency. In contrast, anchor points located around the surface provide the information about shape profile which enables relative position computation between query points and the surface for more global and discriminative position features. Fig. 5 illustrates some qualitative results. As one can see, the depth value fails to project some points which are far away from the surface, and 3D point coordinates perform a little better but still cannot obtain accurate surfaces at edges of the shapes.

To evaluate the influence of the proposed gradient direction alignment, Table 3 reports the results with and without L_{GDA} in Eq. (6). It can be observed that optimizing L_{GDA} further reduces Chamfer and P2S errors. Visual comparison is shown in Fig. 6. By explicitly aligning spatial gradient direction of the distance field during training, we can obtain sharper geometric details for the visible region and less artifacts for the self-occluded region.

We also visualize the anchor points predicted by our model to see if these points actually indicate shape profiles. As illustrated in Fig. 7, the predicted anchor points are evenly distributed around the input garments for different garment topologies and thus are able to provide holistic 3D shape information for query points.

4.2. Comparison with Related Methods

We compare our method against related single-view 3D reconstruction methods with different shape representa-
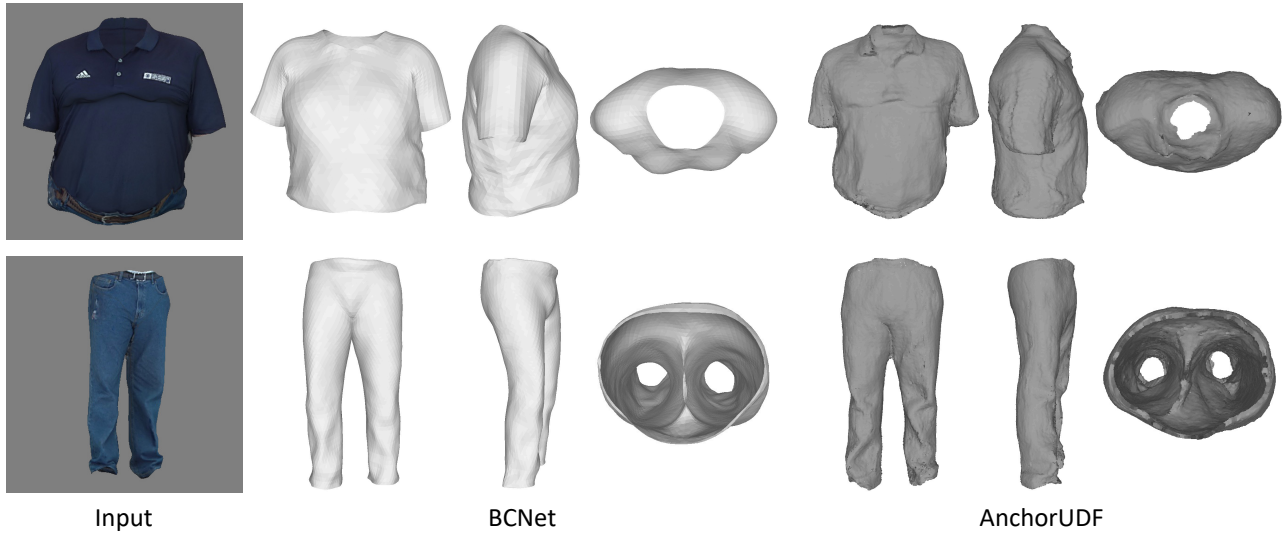
Figure 8. Visual comparison with BCNet [22] on MGN dataset. Although BCNet can recover garment shapes with open surfaces, it tends to produce overly smooth surfaces and loses lots of geometric details.
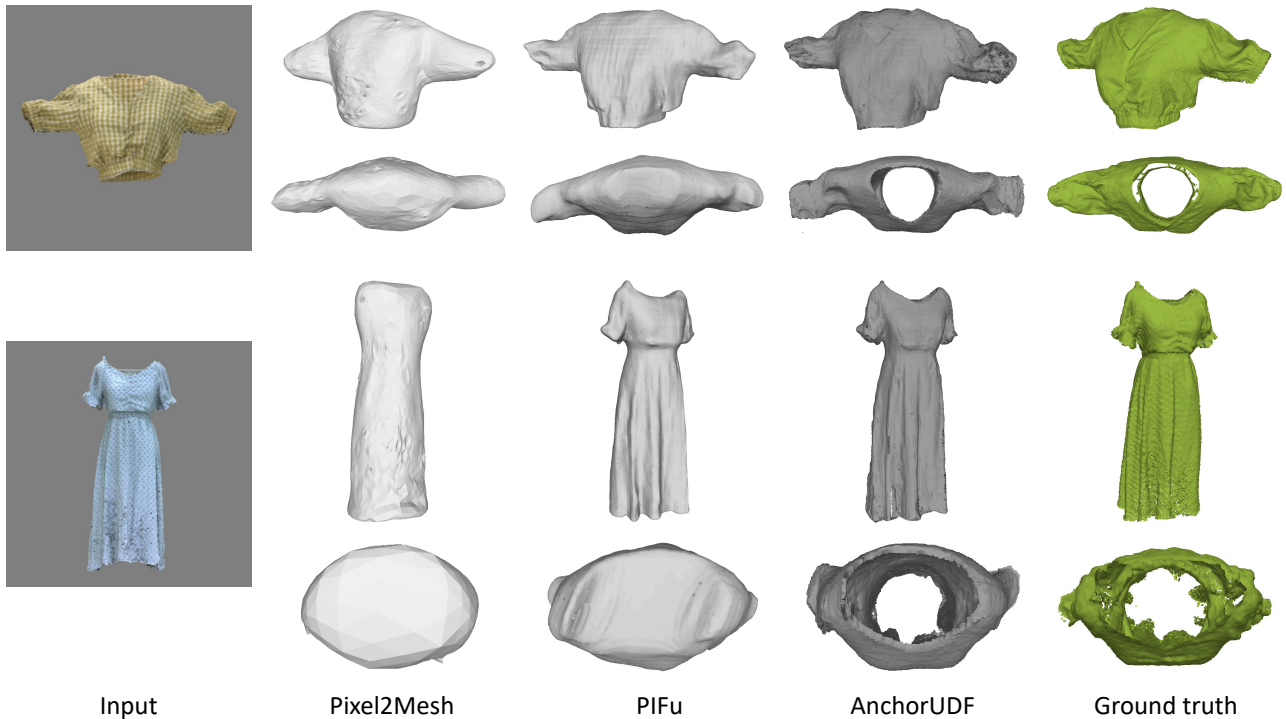

Figure 9. Visual comparison of different single-view reconstruction methods on Deep Fashion3D dataset. Neither Pixel2Mesh [35] nor PIFu [32] can handle garment reconstruction with open surfaces.

tions, including BCNet [22], Pixel2Mesh [35], PIFu [32] and PIFuHD [33]. As listed in Table 4, our AnchorUDF obtains the lowest reconstruction errors on both MGN and Deep Fashion3D datasets. Fig. 8 qualitatively compares our method with BCNet [22] which represents garment shapes using template meshes on MGN dataset. Although such template-based method can recover garment shapes with open surfaces, it tends to produce overly smooth surfaces

and loses lots of geometric details. Fig. 9 presents visual comparison on Deep Fshion3D dataset. Pixel2Mesh [35], a mesh-based method whose shape is initialized from an ellipsoid, only recovers coarse shapes and cannot handle large deformations of garments. PIFu [32] using implicit function representation is able to generate detailed surfaces. However, it can only reconstruct closed surfaces, thus has difficulty handling garment reconstruction with multiple open
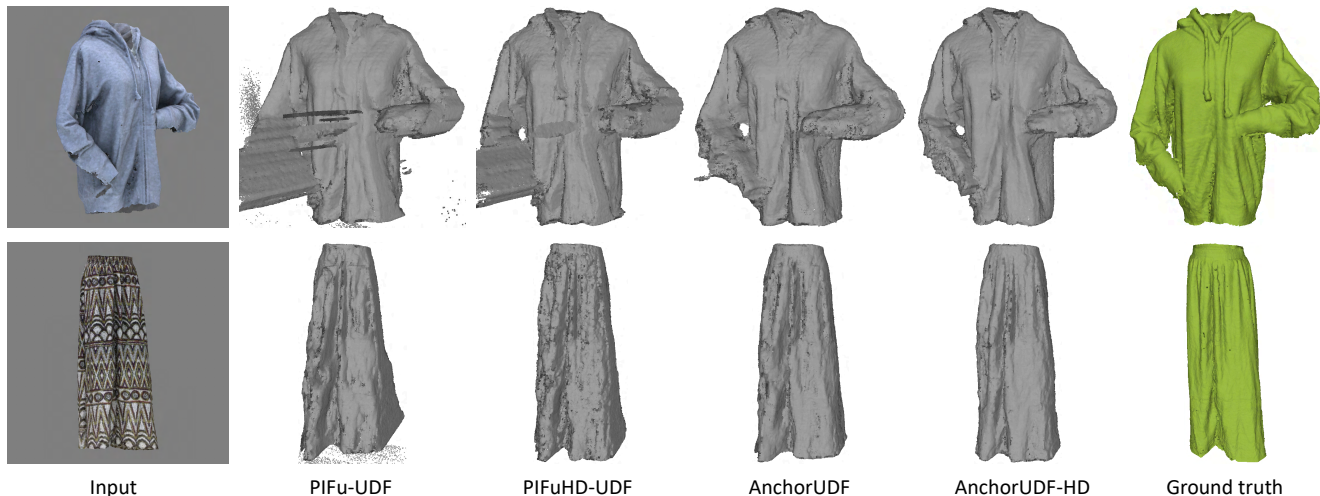
Figure 10. Visual comparison of methods extended with the HD or UDF module on Deep Fashion3D dataset. Our method can reconstruct more accurate shapes and richer details by adding the HD module to take high-resolution images as input.

Table 4. Chamfer and P2S errors ($\times 10^{-3}$) obtained by different single-view reconstruction methods on MGN and Deep Fashion3D datasets.

Methods	MGN	
	Chamfer	P2S
BCNet [22]	4.053	4.512
AnchorUDF	**0.635**	**0.762**
Methods	Deep Fashion3D	
	Chamfer	P2S
Pixel2Mesh [35]	4.266	5.330
PIFu [32]	1.368	1.670
AnchorUDF	**0.621**	**0.839**

Table 5. Chamfer and P2S errors ($\times 10^{-3}$) obtained by methods extended with the HD or UDF module on Deep Fashion3D dataset.

Methods	Chamfer	P2S
PIFu [32]-UDF	1.411	2.440
PIFuHD [33]-UDF	0.969	1.624
AnchorUDF	0.621	0.839
AnchorUDF-HD	**0.574**	**0.737**

boundaries. In contrast, our method can not only retain fine-scale geometric details but also faithfully capture open garment surfaces.

We further incorporate the HD module introduced in PI-FuHD [33] into our AnchorUDF (AnchorUDF-HD) to evaluate the scalability of our method on high-resolution input images (1024×1024). To make a fairer comparison, we try to directly use UDF as the implicit functions in PIFu (PIFu-UDF) and PIFuHD (PIFuHD-UDF) so that they can also represent open surfaces. Note that here we do not use extra normal maps as input because our main purpose is to verify if our method can be further improved under the HD framework. Table 5 shows quantitative comparison on Deep Fashion3D. We can see that by taking high-resolution images as input, AnchorUDF-HD effectively reduces the reconstruction errors, especially P2S, indicating more accurate shape details are recovered. PIFuHD-UDF performs better than PIFu-UDF but still worse than our AnchorUDF, which shows that direct combination with UDF cannot work well and further confirms the necessity of the proposed components in AnchorUDF. Qualitative evaluation is shown in Fig. 10. We can see that AnchorUDF-HD reconstructs more accurate shapes and richer details. Note that although PIFuHD-UDF uses 3D embedding provided by the coarse level as the input of distance function instead of the depth value, it sill cannot project points appropriately onto the surface. More qualitative results can be found in the supplemental materials.

5. Conclusion

We present Anchored Unsigned Distance Function (AnchorUDF) for single-view 3D garment reconstruction. For each query point, AnchorUDF not only extracts its pixel-aligned image features, but also computes its 3D position features based on a set of anchor points located around 3D surface to make the distance function better fit diverse garment shapes. Furthermore, we explicitly align the spatial gradient direction of AnchorUDF with the ground-truth direction to the surface during training to obtain more accurate projection directions for query points at inference. Experiments show that AnchorUDF achieves the state-of-the-art single-view reconstruction performance and is able to be scaled to high-resolution inputs.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018. 3

[2] Antonio Agudo and Francesc Moreno-Noguer. A scalable, efficient, and accurate solution to non-rigid structure from motion. *Computer Vision and Image Understanding (CVIU)*, 2018. 2

[3] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[4] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2018. 2

[5] 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 International Conference on Computer Vision (ICCV)*, 2019. 2

[6] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH*. 2005. 2

[7] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[8] Matan Atzmon and Yaron Lipman. Sald: Sign agnostic learning with derivatives. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 3

[9] Fausto Bernardini, Joshua Mittleman, Holly Rushmeier, Claudio Silva, and Gabriel Taubin. The ball-pivoting algorithm for surface reconstruction. *IEEE transactions on visualization and computer graphics*, 1999. 4

[10] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[11] 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 International Conference on Computer Vision*, 2019. 1, 2, 5

[12] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[13] Julian Chibane, Aymen Mir, and Gerard Pons-Moll. Neural unsigned distance fields for implicit function learning. In *Advances in neural information processing systems (NeurIPS)*, 2020. 3, 4, 6

[14] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1, 3

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

[16] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017. 1, 3, 4

[17] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2017. 3

[18] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. 3

[19] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[20] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Advances in neural information processing systems (NeurIPS)*, 2020. 2, 3

[21] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[22] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 5, 7, 8

[23] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[24] Zorah Lahner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 2015. 1, 2

[26] 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 (CVPR)*, 2019. 1, 3

[27] Patrick Mullen, Fernando De Goes, Mathieu Desbrun, David Cohen-Steiner, and Pierre Alliez. Signing the unsigned: Robust surface reconstruction from raw pointsets. In *Computer Graphics Forum*, 2010. 3

[28] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 5

[29] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 4

[30] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[31] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[32] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 4, 5, 7, 8

[33] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 5, 7, 8

[34] 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)*, 2018. 2

[35] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 5, 7, 8

[36] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in neural information processing systems (NeurIPS)*, 2016. 3

[37] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015. 3

[38] Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, and Yebin Liu. Simulcap: Single-view human performance capture with cloth simulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

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

[40] 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 *Proceedings of the European conference on computer vision (ECCV)*, 2020. 1, 2, 5

[41] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2