

3D Virtual Garment Modeling from RGB Images

Yi Xu*

OPPO US Research Center

Shanglin Yang[†]

JD.COM American Technologies Corporation

Wei Sun[‡]

North Carolina State University

Li Tan[§]

JD.COM

Kefeng Li^{||}

JD.COM

Hui Zhou^{||}

JD.COM American Technologies Corporation

ABSTRACT

We present a novel approach that constructs 3D virtual garment models from photos. Unlike previous methods that require photos of a garment on a human model or a mannequin, our approach can work with various states of the garment: on a model, on a mannequin, or on a flat surface. To construct a complete 3D virtual model, our approach only requires two images as input, one front view and one back view. We first apply a multi-task learning network called JFNet that jointly predicts fashion landmarks and parses a garment image into semantic parts. The predicted landmarks are used for estimating sizing information of the garment. Then, a template garment mesh is deformed based on the sizing information to generate the final 3D model. The semantic parts are utilized for extracting color textures from input images. The results of our approach can be used in various Virtual Reality and Mixed Reality applications.

Index Terms: Computing methodologies—Computer graphics—Graphics systems and interfaces; Computing methodologies—Artificial intelligence—Computer vision—Computer vision problems;

1 INTRODUCTION

Building 3D models of fashion items has many applications in Virtual Reality, Mixed Reality, and Computed-Aided Design (CAD) for apparel industry. A lot of commercial efforts have been put into this field. For example, there are a few CAD software systems that are created for 3D garment design, but most of them focus on creating 3D garment models based on 2D sewing patterns, such as Mavelous-Designer and Optitex. Recently, a few e-commerce platforms have begun to use 3D virtual garments to enhance online shopping experiences. However, large variation, short fashion product life cycle, and high modeling costs make it difficult to use virtual garments on a regular basis. This necessitates a simple yet effective approach for 3D garment modeling.

There have been a lot of research for creating 3D virtual garment models. Some use specialized multi-camera setups to capture 4D evolving shape of the garments [7, 26]. These setups are complicated; therefore limiting their usage. Other methods take 2D sewing patterns [6] or 2D sketches [29] as input and build 3D models that can be easily manufactured. Although these methods use 2D images as input, they still rely on the careful and lengthy design of expert users. Another group of methods deform/reshape 3D template meshes to design garments that best fit 3D digital human models [23]. This can be an overkill in certain applications where



Figure 1: Two product photo sets (left) on an e-commerce site and 3D textured models (right) computed using two photos from each input set.

an accurate design is not needed. Recently, there have been some methods that create 3D garment models from a single image or a pair of images [10, 17, 40, 45]. All of these methods assume the garment is worn by a human model or a mannequin; therefore, do not provide the convenience of working with readily available photos.

We propose a method that can construct 3D virtual garment models from photos that are available on the web, especially on e-commerce sites. Fig. 1 shows two examples. Each photo set displays several different views of a piece of garment on a fashion model, on a mannequin, or flattened on a support surface. To generate a 3D virtual model, a user needs to specify one front and one back image of the garment. The generated 3D model is up to a scale, but can have absolute scale if user specifies a real world measurement (e.g., sleeve length in meters).

We train a multi-task learning network, called JFNet, to predict fashion landmarks and segment a garment image into semantic parts (i.e., left sleeve, front piece, etc.). Based on the landmark predictions, we estimate sizing information of the garment and deform a template mesh to match the estimated measurements. We then deform the semantic parts onto a 2D reference texture to lift textures. It is worth noting that our method is capable of using a single image as input if front-back symmetry is assumed for a garment. Our contributions are as follows:

- We present a complete and easy-to-use approach that generates a 3D textured garment model using product photo set. T-shirt and pants are modeled in this paper; however, our approach can be extended to other garment types.

*e-mail: yi.xu@oppo.com, currently with OPPO US Research Center.

The work was done when Yi Xu was with JD.

[†]e-mail:shanglin.yang@jd.com

[‡]e-mail:wsun12@ncsu.edu, the work was when Wei Sun was with JD.

[§]e-mail:tanli5@jd.com

^{||}e-mail:likefeng@jd.com

^{||}e-mail: hui.zhou@jd.com

- We propose a multi-task learning framework that predicts fashion landmarks and segments garment image into semantic parts.
- We present algorithms for size estimation and texture extraction from garment images.

2 RELATED WORK

In this section, we discuss related work in garment modeling, joint human body and garment shape estimation, semantic parsing of fashion images, and image-based virtual try-on.

2.1 Garment Modeling and Capturing

Garment modeling methods can be classified into the following three categories: geometric approaches, image-based 3D reconstruction, and image-based template reshaping.

2.1.1 Geometric Approaches

Methods in this category typically have roots from the CAD community. Wang et al. [35] automated the Made-to-Measure (MtM) process by fitting 3D feature templates of garments onto different body shapes. Meng et al. [23] proposed a method that preserves the shape of user-defined features on the apparel products during the automatic MtM process.

Other methods use 2D sketches or patterns as input. For example, Decaudin et al. [12] fitted garment panels to contours and seam-lines that are sketched around a virtual mannequin. These panels are then approximated with developable surfaces for garment manufacturing. Robson et al. [29] created 3D garments that are suitable for virtual environments from simple user sketches using context-aware sketch interpretation. Berthouzoz et al. [6] proposed an approach that parses existing sewing patterns and converts them into 3D models. Wang et al. [36] presented a system that is capable of estimating garment and body shape parameters interactively using a learning approach. All of these methods rely on certain level of tailoring expertise from users.

2.1.2 Image-based 3D Reconstruction

Some approaches aimed to create 3D models directly from input images and/or videos of a garment. Early work by White et al. [38] used a custom set of color markers printed on the cloth surface to recover 3D mesh of dynamic cloth with consistent connectivity. Markerless approaches were also developed by using multi-camera setup [7], multi-view 3D scans with active stereo [26], or depth cameras [9]. These methods require specialized hardware and do not work with existing garment photos.

2.1.3 Shape Parameter Estimation

Our approach is most similar to methods that utilize parametric models of human and/or garments. Zhou et al. [45] took a single image of a human wearing a garment as input. Their approach first estimates human pose and shape from images using parameter reshaping. Then, a semi-automatic approach is used to create an initial 3D mesh for the garment. Finally, shape-from-shading is used to recover details. Their method requires user input for pose estimation and garment outline labeling, assumes the garment is front-back symmetric, and does not extract textures from the input image.

Jeong et al. [17] fitted parameterized pattern drafts to input images by analyzing silhouettes. However, their method requires input images of a mannequin both with and without garment from the same viewpoint. Yang et al. [40] used semi-automatic processing to extract semantic information from a single image of a model wearing the garment and used optimization with a physics-inspired objective function to estimate garment parameters. Compared to this

method, our method provides a more advanced joint learning model for semantic parsing.

The DeepGarment framework proposed by Danžřek et al. [10] learns a mapping from garment images to 3D model using Convolutional Neural Networks (CNN). More specifically, the learned network can predict displacements of vertices from a template mesh. However, garment texture is not learned.

2.2 Joint Human Body and Garment Shape Estimation

There have been a lot of efforts that address the challenging problem of joint human body and garment shape estimation.

Alldieck et al. [3] reconstructed detailed shape and texture of clothed human by transforming a large amount of dynamic human silhouettes from a single RGB sequence to a common reference frame. Later, the same authors introduced a learning approach that only requires a few RGB frames as input [1]. Natsume et al. [24] reconstructed a complete and textured 3D model of a clothed person using just one image. In their work, deep visual hull algorithm is used to predict 3D shape from silhouettes and a Generative Adversarial Network (GAN) is used to infer the appearance of the back of the human subject. Habermann et al. [14] presented a system for real time tracking of human performance, but relied on a personalized and textured 3D model that was captured during a pre-processing step. These work do not separate underlying body shape from garment geometry.

Using RGBD camera as input device, body shape and garment shape can be separated. For example, Zhang et al. [43] reconstructed naked human shape under clothing. Yu et al. [41] used a double layer representation to reconstruct geometry of both body and clothing. Physics based cloth simulation can also be incorporated into the framework to better track human performance [42].

2.3 Fashion Semantic Parsing

In this section, we review related work in fashion landmark prediction, semantic segmentation, and multi-task learning.

2.3.1 Fashion Landmark Prediction

Fashion landmark prediction is a structured prediction problem for detecting functional key points, such as corners of cuff, collar, etc. Despite it being a relatively new topic [21], it has roots in a related problem-human pose estimation. Early work on human pose estimation used pictorial structures to model spatial correlation between human body parts [4]. Such method only works well when all body parts are visible, so that the structure can be modeled by graphical models. Later on, hierarchical models were used to model part relationships at multiple scales [33]. Spatial relationship can also be learned implicitly using a sequential prediction framework, such as Pose Machines [27]. CNNs can also be integrated into Pose Machines to jointly learn image features and spatial context features [37].

Different from human pose, fashion landmark detection predicts functional key points of fashion items. Liu et al. proposed a Deep Fashion Alignment (DFA) [21] framework that cascades CNNs in three stages similar to DeepPose [34]. To achieve scale invariance and remove background clutter, DFA assumes that bounding boxes are known during training and testing; thus limiting its usage. This constraint was later removed in Deep LAndmark Network (DLAN) [39]. It is worth noting that the landmarks defined in these approaches cannot be used for texture extraction. For example, a mid-point on the cuff is a landmark defined in their work. In our work, two corners of the cuff are predicted and they carry critical information for texture extraction.

2.3.2 Semantic Segmentation

Semantic segmentation assigns semantic labels to each pixel. CNNs have been successfully applied to this task. Long et al. pro-

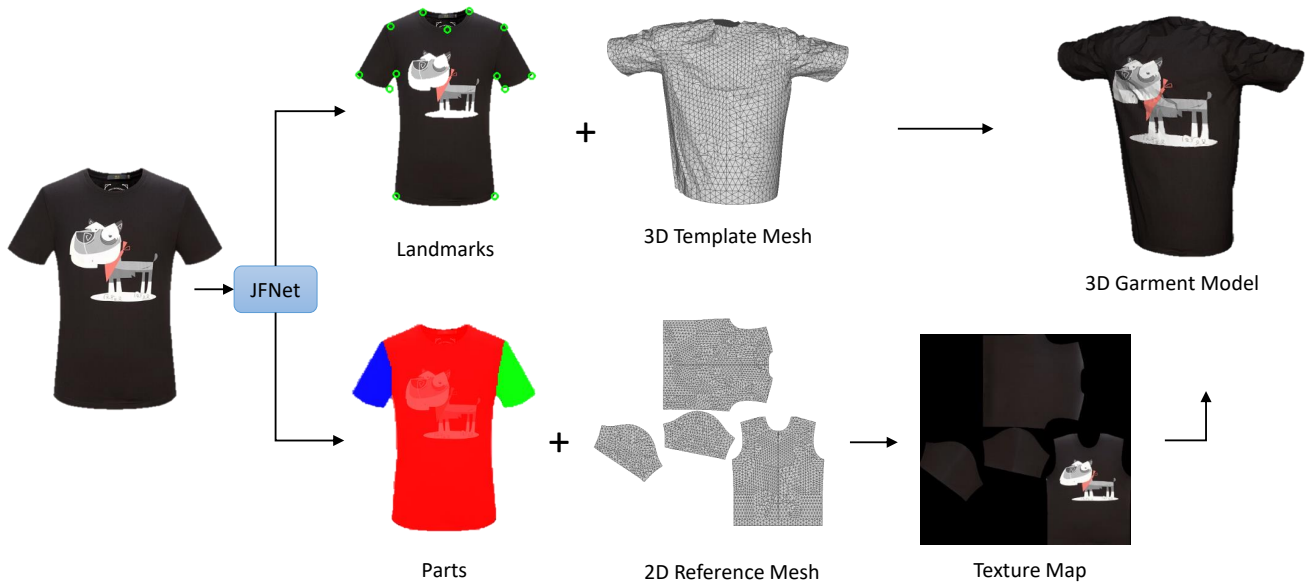


Figure 2: System Overview. For each input image, we jointly predict landmark locations and segment the garment into semantic parts using the proposed JFNet. The predicted landmarks are used to guide the deformation of a 3D template mesh. The segmented parts are used to extract garment textures. Finally, 3D textured garment model is produced.

posed Fully Convolutional Networks (FCNs) for semantic segmentation [22], which achieved significant improvements over methods relied on hand-crafted features. Built upon FCNs, Encoder-Decoder architectures have shown great success [5, 30]. Such an architecture typically has an encoder that reduces feature map and a decoder that maps the encoded information back to input resolution. Spatial Pyramid Pooling (SPP) can also be applied at several scales to leverage multi-scale information [44]. DeepLabV3+ [8] combines the benefits of both SPP and Encoder-Decoder architecture to achieve state-of-the-art result. Our part segmentation sub-network is based on DeepLabV3+ architecture. Similar to our work, Alldieck et al. [2] also used human semantic part segmentation to extract detailed textures from RGB sequences.

2.3.3 Multi-task Learning

Multi-task learning (MTL) has been used successfully for many applications due to the inductive bias it achieves when training a model to perform multiple tasks. Recently, it has been applied to several computer vision tasks. Kokkinos introduced UberNet [18] that can jointly handle multiple computer vision tasks, ranging from semantic segmentation, human parts, to object detection. Ranjan et al. proposed HyperFace [28] for simultaneously detecting faces, localizing landmarks, estimating head pose, and identifying gender. Perhaps the most similar work to ours is the work of JPPNet [20]. It is a joint human parsing and pose estimation network, while our work uses MTL for garment image analysis. Another MTL work on human parsing from the same group is [13], where semantic part segmentation and instance-aware edge detection are jointly learned.

2.4 Image-based Virtual Try-on

As an alternative to 3D modeling, image-based virtual try-on has also been explored. Neverova et al. [25] used a two-stream network where a data-driven predicted image and a surface-based warped image are combined and the whole network is learned end-to-end to generate a new pose of a person. Lassner et al. [19] used only image information to predict images of new people in different clothing

items. VITON [15] on the other hand transfers the image of a new garment onto a photo of a person.

3 OUR APPROACH

In this section, we explain our approaches on garment image parsing, 3D model creation, and texture extraction. Fig. 2 shows an overview of our approach.

3.1 Data Annotation

To train JFNet, we built a dataset with both fashion landmarks and pixel-level segmentation annotations. We collected 3,000 images of tops (including T-shirts) and another 3,000 images of pants from the web. For each type of garment, a set of landmarks are defined based on fashion design. 13 landmarks are defined for tops including center and corners of neckline, corners of both cuffs, end points on hemline, and armpits. 7 landmarks are defined for pants including end points of waistband, crotch, and end points of the bottom.

For part segmentation, we defined a set of labels and asked the annotators to provide pixel-level labeling. For tops, we used 5 labels including left-sleeve, right-sleeve, collar, torso, and hat. For pants, we used 2 labels including left-part and right-part. Some labeling examples are shown in Fig. 3.

3.2 Garment Image Parsing

Our joint garment parsing network JFNet built upon Convolutional Pose Machines (CPMs) [37] for landmark prediction and DeepLabV3+ [8] for semantic segmentation.

The network architecture of JFNet is illustrated in Fig. 4. We use ResNet-101 [16] as our backbone network to extract low-level features. Then we use two branching networks to obtain landmark prediction and part segmentation. Finally, we use a refinement network to refine the prediction results.

3.2.1 Landmark Prediction

For landmark prediction (bottom half of Fig. 4), we use a learning network with T -stages similar to that of [37]. At first stage, we extract second stage outputs of ResNet-101 (Res-2) followed by a 3×3

convolutional layer as low level features from the input image. Then, we use two 1x1 convolutional layers to predict landmark heatmap at the first stage. At each of the subsequent stages, we concatenate the landmark heatmap predicted from the previous stage with shared low-level features from Res-2. Then we use five convolutional layers followed by two 1x1 convolutional layers to predict the heatmap at the current stage. The architecture repeats this process for T stages, where the size of receptive field increases with each stage. This is crucial for learning long-range relationships between fashion landmarks. The heatmap at each stage is compared against labeled ground truth and calculated towards total training loss.

3.2.2 Garment Part Segmentation

For semantic garment part segmentation (top half of Fig. 4), we followed the encoder architecture of DeepLabV3+ [8]. Atrous Spatial Pyramid Pool (ASPP) module, which can learn context information at multiple scales effectively, is applied after the last stage output of ResNet-101, followed by one 1x1 convolutional layer and up-sampling.

3.2.3 Refinement

To refine landmark prediction and part segmentation, and to promote each other, we concatenate the landmark prediction result from the T -th stage of the landmark sub-network, the part segmentation result from the segmentation sub-network, and the shared low-level features together. We then apply a 3x3 convolutional layer for landmark prediction and part segmentation respectively. The sum of loss from both branches is used for jointly training the network end-to-end.

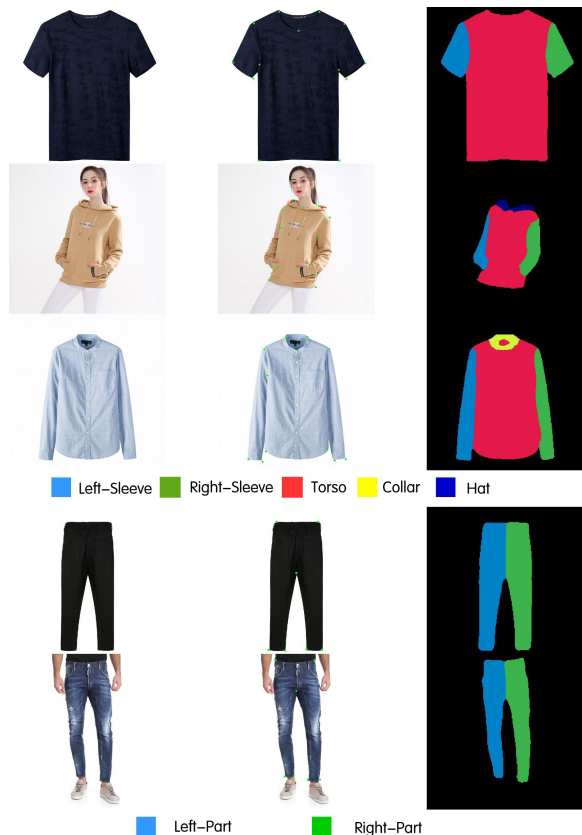


Figure 3: Annotation Examples. Top and bottom shows landmark and part labeling for tops (including T-shirt) and pants respectively.

3.2.4 Training Details

We load ResNet-101 parameters that are pre-trained on ImageNet classification task. During training, random crop and random rotation between -10 and 10 degrees are applied for data augmentation and the final input image size is resized to 256x256. We adopt SGD optimizer with 0.9 as momentum. Learning rate is initially set as 0.001 and “poly” decay [44] is set to 10^{-6} in 100 total training epochs.

3.3 3D Model Construction

Our approach uses fashion landmarks to estimate the sizing information and to guide the deformation of a template mesh. Textures are extracted from input images and mapped onto the 3D garment model. In this section, we first discuss the garment templates used in our system. Then, we discuss our 3D modeling and texturing approaches.

3.3.1 Garment Templates

We use 3D garment models from Berkeley Garment Libraries [11] as templates. For each garment type, a coarse base mesh and a finer isotropically refined mesh are provided by the library. We use the refined mesh in world-space configuration as our base model. In addition, the texture coordinates of the refined mesh store the material coordinates that refer to a planar reference mesh. We use this 2D reference mesh for texture extraction. Currently, our system supports two garment types: T-shirt and pants as shown in Fig. 5.

3.3.2 3D Model Deformation

To create 3D garment models that conform to the sizing information from the input images, we apply Free-Form Deformation (FFD) [32] to deform a garment template. We chose FFD because it can be applied to 3D models locally while maintaining derivative continuity with adjacent regions of the model. For two view data (front and back), FFD is a plausible solution. When there are multi-view images, videos, or 4D scans of garments, other mesh fitting techniques can be used to generate more accurate results.

For each garment template, we impose a grid of control points P_{ijk} ($0 \leq i < l$, $0 \leq j < m$, $0 \leq k < n$) on a lattice. The deformation of the template is achieved by moving each control point P_{ijk} from its original position. Control points are carefully chosen to facilitate deformation of individual parts so that a variety of garment shapes can be modeled. For T-shirt, as shown in Fig. 6 (a, b), we use $l = 4$, $m = 2$, $n = 4$. For pants, as shown in Fig. 6 (c, d), we use control points with $l = 3$, $m = 2$, $n = 3$.

If metric scale of the resulting 3D model is desired, we ask the user to specify a measurement l in world space (e.g., sleeve length). Otherwise, a default value is assigned to l . Based on the ratio between image space sleeve length to l , we can convert any image space distance to world space distance.

FFD control points do not directly corresponded to image landmarks. Instead, we compute 2D distances between garment landmarks and use them to compute 3D distances between control points. Tab. 1 shows how to calculate control point distances for the T-shirt

Table 1: Control Points Distances from Landmarks for T-shirt

Distance	How to calculate
$D(P_{0jk}, P_{1jk})$	left sleeve length * $\cos(\alpha)$
$D(P_{1jk}, P_{2jk})$	chest width (armpit_left to armpit_right)
$D(P_{2jk}, P_{3jk})$	right sleeve length * $\cos(\beta)$
$D(P_{ij0}, P_{ij1})$	distance from armpit to hemline
$D(P_{ij1}, P_{ij2})$	distance from armpit to shoulder
$D(P_{ij0}, P_{ij3})$	distance from neck to hemline
$D(P_{i0k}, P_{i1k})$	$D(P_{ij1}, P_{ij2}) * S$
S	$D(P_{i0k}, P_{i1k}) / D(P_{ij1}, P_{ij2})$, un-displaced.

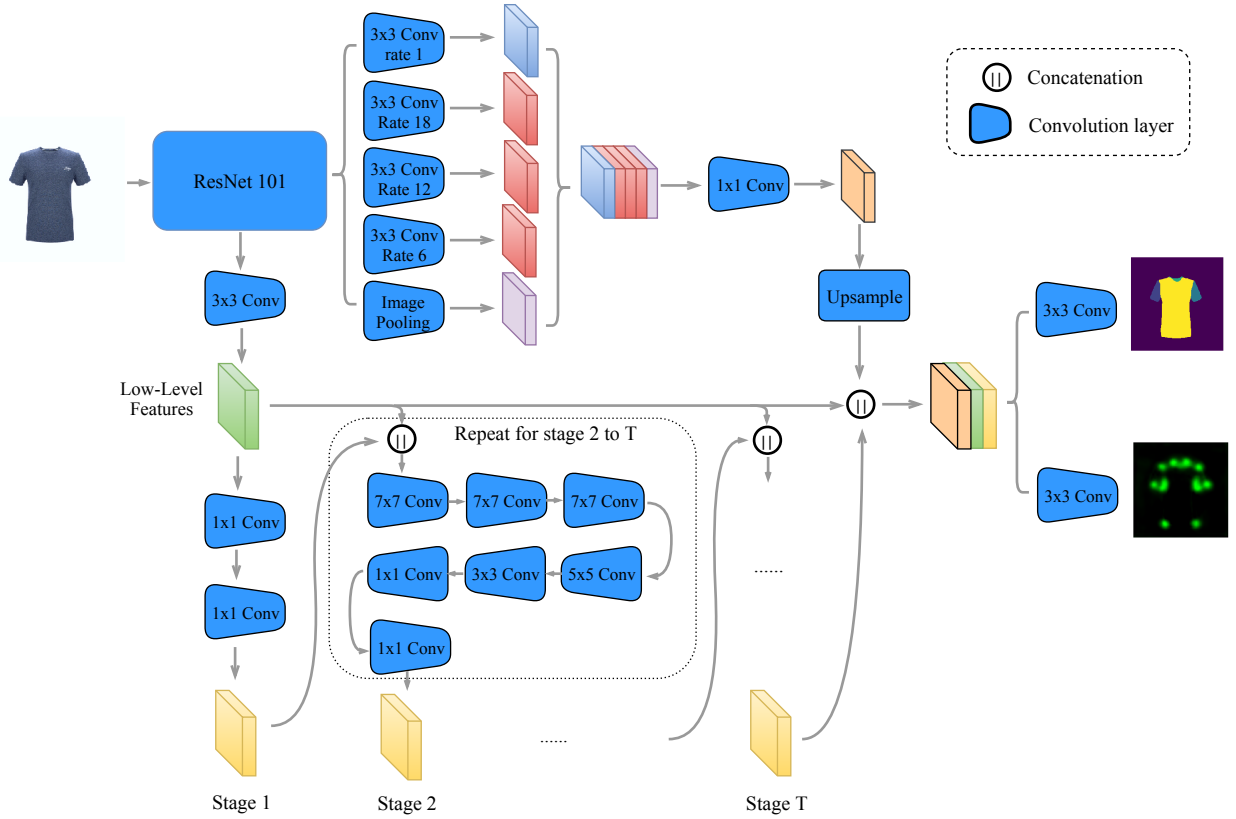


Figure 4: JFNet. Our proposed multi-task learning model use ResNet-101 as backbone network to extract shared low level features. For landmark prediction (bottom half), we apply T -stage CNNs. Each stage refines the prediction iteratively. For garment part segmentation, Atrous Spatial Pyramid Pool (ASPP) is applied on the ResNet output and followed by 1x1 convolution and up-sampling. At the last stage of the network, results from two branches are concatenated together for joint learning.

type. Constants α and β are the angle between horizontal direction and left sleeve and the angle between horizontal direction and right sleeve respectively. They are measured from the template T-shirt mesh. The distances are then used to compute new locations of control points for template mesh deformation.

Since the T-shirt template resembles the shape of a T-shirt on a mannequin, using photos of T-shirts on mannequins achieves most accurate results. On such images, the distance between two armpits corresponds to the chest width of the mannequin. When a T-shirt lays on a flat surface, the distance between two armpits corresponds to half perimeter of the chest. In this case, we fit an ellipse to the horizontal section of the chest. We then compute the width of the horizontal section as the major axis of the ellipse using the perimeter measurement. Images of fashion models are not suitable for garment size estimation due to self-occlusion, wrinkles, etc. Tab. 2 shows the calculation of control points for the pants.

Table 2: Computing Control Points Distances for Pants

Control Points	How to calculate
$D(P_{0jk}, P_{1jk})$	un-displaced distance * S^*
$D(P_{1jk}, P_{2jk})$	un-displaced distance * S^*
$D(P_{ij0}, P_{ij1})$	distance from crotch to bottom
$D(P_{ij1}, P_{ij2})$	distance from crotch to waist line
$D(P_{i0k}, P_{i1k})$	un-displaced distance * S^*

* S is ratio between new waist girth to template waist girth.

3.4 Texture Extraction

The texture coordinates in the 3D mesh refer to the vertices in the planar 2D reference mesh. This allows us to perform 3D texture mapping by mapping input images onto the 2D reference mesh as a surrogate. The different pieces in the reference mesh correspond to different garment segmentation parts. This is the reason semantic segmentation is performed during garment image analysis. Texture mapping becomes an image deformation problem where the source is a garment part (e.g., left sleeve) and the target is its corresponding piece on the reference mesh.

On the reference mesh, we manually label the landmarks (Fig. 7 (b) red circles). This only needs to be done once for each garment type. In this way, we establish feature correspondence between predicted landmarks on the source image and manually-labeled landmarks on the target image. However, using a sparse set of control points leads to large local deformation, especially around contours. To mitigate this, we map each landmark point onto the contour of the part by finding the closest point on the part contour. Then between each pair of adjacent landmarks, we sample N additional points uniformly along the contour. We do this for both input garment image and reference mesh (green circles in Fig. 7). The corresponding points are then used by Moving Least Squares (MLS) method with similarity deformation [31] to deform textures from the input image to the reference mesh. Alternatively, a Thin Plate Spline (TPS) based approach similar to that used in VITON [15] can also be used for image warping.

Before image deformation, each garment segment is eroded

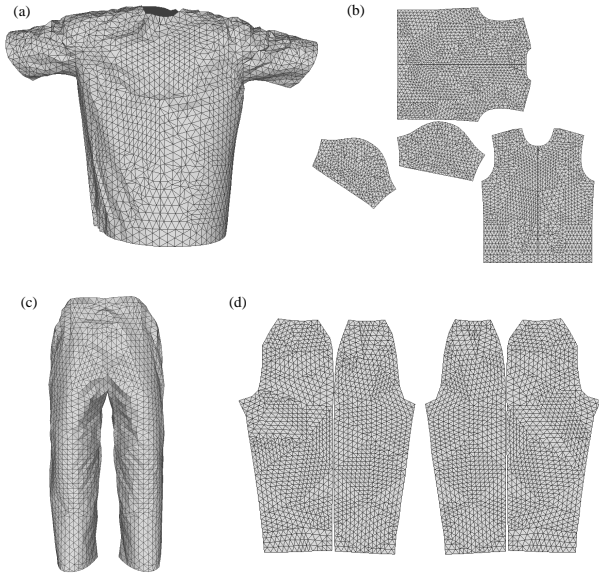


Figure 5: Our approach uses garment templates for modeling and texturing. (a) The template mesh for T-shirt, whose texture coordinates match the vertex coordinates of the (b) reference mesh. (c) Template mesh for pants, and the corresponding (d) reference mesh.

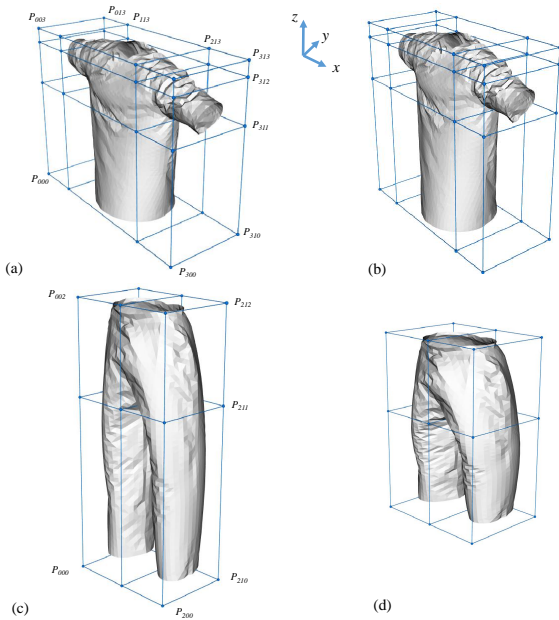


Figure 6: Template Deformation. (a) The original template for T-shirt with control grid. (b) Deformed template that captures a different shape. (c) The original template for pants. (d) Deformed template.

slightly to accommodate for segmentation artifacts. Then, color texture is extrapolated from the garment to surrounding area to remove background color after deformation. Fig. 7 shows the process of deforming the front segment of a T-shirt to the desired location on its 2D reference mesh. Fig. 8 shows that for the right leg of pants. Note that to better illustrate the idea, we use a small value of $N = 10$

in Fig.7 and 8. In our experiments, we found that denser control point set (e.g. $N = 50$) works better.

In our current implementation, the back piece around the neck/collar is often included in the front piece segmentation result. To handle this, we cut out the back piece automatically. JFNet predicts the front middle point of the neck as a landmark. We then correct the front piece segmentation by tracing the edge from two shoulder points to the middle neck point.

4 EXPERIMENTS

In this section, we show quantitative experimental results for JFNet. We also show results on 3D modeling.

4.1 Evaluation of JFNet

Our model requires both landmark and segmentation annotations, thus we cannot compare our results directly with other SOTAs by training our model on public dataset. Nevertheless, we have trained CPM and DeepLabV3+ on our dataset and compare them with JFNet.

We trained JFNet for tops and pants separately. For each model, 2,000 images are used for training and 500 images for validation. Evaluation is performed on the remaining 500 images. We used the standard intersection over union (IoU) criterion and mean IoU ($mIoU$) accuracy for segmentation evaluation and normalized error (NE) metric [21] for landmark prediction evaluation. NE refers to the distance between predicted landmarks and ground truth locations in the normalized coordinate space (*i.e.*, normalized with respect to the width of the image), and it is a commonly used evaluation metric.

Tab. 3 shows performances of different methods. For both tops and pants, JFNet achieves better performance on *both* landmark prediction and garment part segmentation. Our landmark prediction on tops greatly outperforms CPM (0.031 vs. 0.075). This shows that constraints and guidance from segmentation task have helped landmark prediction. Landmark prediction performance on pants also improves, but not as much because landmarks of pants are less complex than those of tops. Part segmentation is a more complex task. Thus, it is reasonable that our model does not boost the segmentation task as much. Nevertheless, JFNet still improves upon DeepLabV3+.

It is worth noting that the purpose of the proposed model is to handle multiple tasks simultaneously with performance improvement compared to individual tasks. Thus, our method focuses on information sharing and multi-task training while other SOTAs focus on network structure and training for each individual task. In the

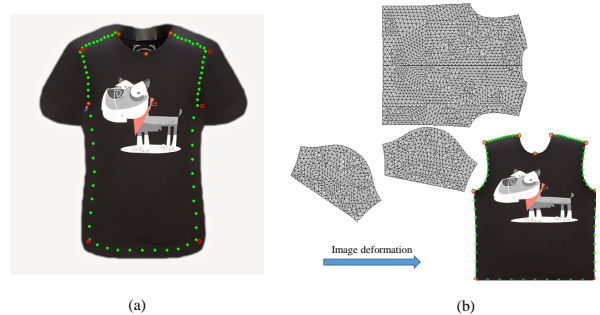


Figure 7: Texture Extraction for T-Shirt. (a) The extrapolated T-shirt image with control points computed along the contour of the front segment. (b) The front segment is deformed to the desired location on the 2D reference mesh.



Figure 8: Texture Extraction for Pants. (a) The extrapolated pants image with control points. (b) The image segment is deformed to the desired location on the 2D reference mesh.

future, we can also incorporate other SOTA networks into our joint learning model.

Table 3: Landmark Prediction and Garment Segmentation Performance Comparison

Methods	Tops		Pants	
	NE	mIOU	NE	mIOU
<i>CPM</i> [37]	0.075	—	0.034	—
<i>Deeplabv3+</i> [8]	—	0.721	—	0.964
<i>JFNet</i>	0.031	0.725	0.022	0.968

4.2 3D Modeling Results

We applied our 3D garment modeling algorithm on various input images and the results are in Fig. 9. Our approach utilizes the sizing information estimated from fashion landmarks to model different styles of garments (e.g., different length of legs or different fits of T-shirt). For example, the 3rd T-shirt is slightly longer, the 2nd T-shirt is slight wider, and the 1st T-shirt has narrower sleeves. These correspond to the characteristics of the input garment images. Our approach can also extract textures from garment images and map them on to different parts of the constructed 3D model.

To quantitatively evaluate our 3D modeling is expensive. This involves capturing 2D images of various garments and scanning them into 3D models. An alternative is to use synthetic data with ground truth to evaluate accuracy of size estimation and 3D reconstruction. We leave these for future work. Nevertheless, 3D modeling results of our approach are visually plausible for applications where accuracy requirement is not strict.

5 CONCLUSION

We present a complete system that takes photos of a garment as input and creates a 3D textured virtual model. We propose a multi-task network called JFNet to predict fashion landmarks and segment the garment into parts. The landmark prediction results are used to guide template-based deformation. The semantic part segmentation results are used for texture extraction. We show that our system can create 3D virtual models for T-shirt and pants effectively.

6 LIMITATION

One limitation is due to the representation power of the templates. Because our model is deformed from a template, the shape of the template limits the range of garments we can model. For example, our pants template is a regular fit. Modeling slim or skinny pants

will be impractical. Our approach recovers shape, but not the pose of the garment. To learn the 3D pose of garments, more data and annotations are required.

Another limitation is that we only use two photos (front and back view) for texture extraction. This leads to excessive local deformation when source and target contours are very different (see stickers on the jeans in Fig. 9 last row).

The photo sets for testing our 3D modeling approach are from online shopping sites. Two occlusion-free images can always be selected from each set. In general, occlusion can pose a problem for texture extraction. However, missing textures can be mitigated using image in-painting. Missing landmarks can be mitigated using symmetry-based landmark completion.

Finally, our system only supports T-shirt and pants now and we only address a simplified version of the garment modeling problem, which usually involves wrinkles, folds and pleats.

7 FUTURE WORK

Currently, 2D proportions from the photos are transferred to the 3D model. In the future, We want to use a garment modeling approach that uses sewing patterns [17]. We can fit the shape of each individual 2D sewing pattern using image part segmentation. Then, these 2D patterns can be assembled in 3D space as in commercial garment design process. In this way, we can better transfer the shapes from 2D images to 3D models.

We also want to investigate if more than two images can be used together to texture a 3D model [2]. The distorted textures along the silhouettes of front and back view can be filled in by a side view photo.

For applications that require accurate 3D information, we would like to perform quantitatively evaluation of our 3D modeling algorithm.

Finally, by incorporating more garment templates, more garment types can be supported. Since we only need to create a template once for each type/fit, the overhead is small if used in large scales. There are certain garments that are not suitable for our approach (e.g., fancy dresses with customized design). A possible approach is to use a hybrid system where template-based deformation generates a base model and 3D details can be added via other methods. Part segmentation in its current state is not suitable for open jackets. It would be interesting to see if semantic segmentation model with more data and annotation can distinguish between back side and front side.

ACKNOWLEDGMENTS

The authors wish to thank the reviewers for their insightful comments and suggestions.

REFERENCES

- [1] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. 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.
- [2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pp. 98–109, Sep. 2018. doi: 10.1109/3DV.2018.00022
- [3] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8387–8397, June 2018. doi: 10.1109/CVPR.2018.00875
- [4] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1014–1021, June 2009. doi: 10.1109/CVPR.2009.5206754
- [5] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation.



Figure 9: 3D Modeling Results. On each row we show front image and its landmark prediction and part segmentation, followed by back image and its landmark and part segmentation results. The final two columns show 3D textured models for two view points.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12):2481–2495, Dec 2017. doi: 10.1109/TPAMI.2016.2644615

- [6] F. Berthouzoz, A. Garg, D. M. Kaufman, E. Grinspun, and M. Agrawala. Parsing sewing patterns into 3d garments. *ACM Trans. Graph.*, 32(4):85:1–85:12, July 2013. doi: 10.1145/2461912.2461975
- [7] D. Bradley, T. Popa, A. Sheffer, W. Heidrich, and T. Boubekeur. Markerless garment capture. *ACM Trans. Graph.*, 27(3):99:1–99:9, Aug. 2008. doi: 10.1145/1360612.1360698
- [8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- [9] X. Chen, B. Zhou, F. Lu, L. Wang, L. Bi, and P. Tan. Garment modeling with a depth camera. *ACM Trans. Graph.*, 34(6):203:1–203:12, Oct. 2015. doi: 10.1145/2816795.2818059
- [10] R. Danžřek, E. Dibra, C. Öztireli, R. Ziegler, and M. Gross. Deepgarment: 3d garment shape estimation from a single image. *Comput. Graph. Forum*, 36(2):269–280, May 2017. doi: 10.1111/cgf.13125
- [11] J. M. de Joya, R. Narain, J. F. O’Brien, A. Samii, and V. Zordan. Berkeley Garment Library, 2012. Available at <http://graphics.berkeley.edu/garment/>

berkeley.edu/resources/GarmentLibrary/index.html.

- [12] P. Decaudin, D. Julius, J. Wither, L. Boissieux, A. Sheffer, and M.-P. Cani. Virtual garments: A fully geometric approach for clothing design. *Computer Graphics Forum*, 25(3):625–634. doi: 10.1111/j.1467-8659.2006.00982.x
- [13] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin. Instance-level human parsing via part grouping network. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- [14] M. Habermann, W. Xu, M. Zollhöfer, G. Pons-Moll, and C. Theobalt. LiveCap: Real-time human performance capture from monocular video. *ACM Trans. Graph.*, 38(2):14:1–14:17, Mar. 2019. doi: 10.1145/3311970
- [15] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. Viton: An image-based virtual try-on network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7543–7552, June 2018. doi: 10.1109/CVPR.2018.00787
- [16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016. doi: 10.1109/CVPR.2016.90
- [17] M.-H. Jeong, D.-H. Han, and H.-S. Ko. Garment capture from a photograph. *Computer Animation and Virtual Worlds*, 26(3-4):291–300. doi: 10.1002/cav.1653
- [18] I. Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5454–5463, July 2017. doi: 10.1109/CVPR.2017.579
- [19] C. Lassner, G. Pons-Moll, and P. V. Gehler. A generative model of people in clothing. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [20] X. Liang, K. Gong, X. Shen, and L. Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, p. 1, 2018. doi: 10.1109/TPAMI.2018.2820063
- [21] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang. Fashion landmark detection in the wild. In *European Conference on Computer Vision (ECCV)*, October 2016.
- [22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, June 2015. doi: 10.1109/CVPR.2015.7298965
- [23] Y. Meng, C. C. L. Wang, and X. Jin. Flexible shape control for automatic resizing of apparel products. *Comput. Aided Des.*, 44(1):68–76, Jan. 2012. doi: 10.1016/j.cad.2010.11.008
- [24] R. Natsume, S. Saito, W. C. Zeng Huang, C. Ma, H. Li, and S. Morishima. SiCloPe: Silhouette-based clothed people. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.
- [25] N. Neverova, R. Alp Guler, and I. Kokkinos. Dense pose transfer. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- [26] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Trans. Graph.*, 36(4):73:1–73:15, July 2017. doi: 10.1145/3072959.3073711
- [27] V. Ramakrishna, D. Munoz, M. Hebert, J. Andrew Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds., *Computer Vision – ECCV 2014*, pp. 33–47. Springer International Publishing, Cham, 2014.
- [28] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018. doi: 10.1109/TPAMI.2017.2781233
- [29] C. Robson, R. Maharik, A. Sheffer, and N. Carr. Context-aware garment modeling from sketches. *Comput. Graph.*, 35(3):604–613, June 2011. doi: 10.1016/j.cag.2011.03.002
- [30] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds., *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241. Springer International Publishing, Cham, 2015.
- [31] S. Schaefer, T. McPhail, and J. Warren. Image deformation using moving least squares. *ACM Trans. Graph.*, 25(3):533–540, July 2006. doi: 10.1145/1141911.1141920
- [32] T. W. Sederberg and S. R. Parry. Free-form deformation of solid geometric models. *SIGGRAPH Comput. Graph.*, 20(4):151–160, Aug. 1986. doi: 10.1145/15886.15903
- [33] Y. Tian, C. L. Zitnick, and S. G. Narasimhan. Exploring the spatial hierarchy of mixture models for human pose estimation. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, eds., *Computer Vision – ECCV 2012*, pp. 256–269. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
- [34] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660, June 2014. doi: 10.1109/CVPR.2014.214
- [35] C. C. L. Wang, Y. Wang, and M. M. F. Yuen. Design automation for customized apparel products. *Comput. Aided Des.*, 37(7):675–691, June 2005. doi: 10.1016/j.cad.2004.08.007
- [36] T. Y. Wang, D. Ceylan, J. Popovic, and N. J. Mitra. Learning a shared shape space for multimodal garment design. *ACM Trans. Graph.*, 37(6):1:1–1:14, 2018. doi: 10.1145/3272127.3275074
- [37] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4732, June 2016. doi: 10.1109/CVPR.2016.511
- [38] R. White, K. Crane, and D. A. Forsyth. Capturing and animating occluded cloth. *ACM Trans. Graph.*, 26(3), July 2007. doi: 10.1145/1276377.1276420
- [39] S. Yan, Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In *Proceedings of the 2017 ACM on Multimedia Conference, MM ’17*, pp. 172–180. ACM, New York, NY, USA, 2017. doi: 10.1145/3123266.3123276
- [40] S. Yang, T. Ambert, Z. Pan, K. Wang, L. Yu, T. L. Berg, and M. C. Lin. Physics-inspired garment recovery from a single-view image. *ACM Trans. Graph.*, 2018.
- [41] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu. DoubleFusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [42] T. Yu, Z. Zheng, Y. Zhong, J. Zhao, Q. Dai, G. Pons-Moll, and Y. Liu. SimulCap: Single-view human performance capture with cloth simulation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [43] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5484–5493, July 2017. doi: 10.1109/CVPR.2017.582
- [44] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230–6239, July 2017. doi: 10.1109/CVPR.2017.660
- [45] B. Zhou, X. Chen, Q. Fu, K. Guo, and P. Tan. Garment modeling from a single image. *Computer Graphics Forum*, 32(7):85–91. doi: 10.1111/cgf.12215