NANYANG TECHNOLOGICAL UNIVERSITY

GARMENT MESH RECONSTRUCTION AND ANIMATION

JU CHUANG

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University in partial fulfilment of the requirement for the degree of Master of Engineering

2022

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

08/09/2022

. . . . . .

Date

. . . . . .

Ju Chuang

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

08/09/2022

. . . . . .

Date

. . . . . .

Loy Chen Change

Authorship Attribution Statement

This thesis **does not** contain any materials from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

08/09/2022

. . . . . .

Date

. . . . . .

Ju Chuang

Acknowledgements

I would like to thank my supervisor Associate Professor Chen Change Loy, my co-supervisor Doctor Dai Bo and my senior Zhang Junzhe for their continuous support, encouragement and motivation. I could not have finished the project without their insightful and knowledgeable guidance. I would also like to thank my friends for their support.

Table of Contents

List of Figures

List of Tables

Abstract

With great potential in online garment shopping and game animation, recent years research in garment virtual try-on becomes more and more popular. Garment virtual try-on can be divided into three parts, namely 3D garment mesh reconstruction, animation and texture transfer. Most of current research only focus on one of the sub-fields and are difficult to be combined. There is still no end-to-end pipeline. In this report, we propose a network to reconstruct the 3D garment mesh that is easily to be combined with animation model and make contribution to the completeness of the end-to-end pipeline.

Our network can properly predict 3D garment meshes according to various garment styles. Unlike some of the existing approaches that require complex inputs, our model works with 2D garment images or masks that are easily accessible. Unsupervised learning is implemented during training, making the data collection much easier as no labelling is required. We have successfully combined our model with existing 3D animation model. Now the network can directly work from 2D images instead of 3D garment information.

Chapter 1

Introduction

Recent years, 3D garment model reconstruction and animation become popular research fields due to great business potential in virtual try-on for online garments shopping, 3D content production, entertainment, video game industries, virtual reality, augmented reality and game animation. Garment virtual try-on applies deep learning techniques to wear garments in a virtual environment and then animates the human body together with the constructed 3D garments. The process mainly involves the sub-fields of garment mesh reconstruction [3, 8, 14, 15, 17, 19, 21, 34, 50 - 61], wrinkle prediction [2, 7, 9, 10, 11, 42, 44 - 48] and texture transfer [1, 3 - 6, 12, 20, 35, 36, 62 - 88].

Garment mesh reconstruction Previously, experienced designers are required to design garment models manually which can be laborious and time-consuming. So automatically generating garment models becomes an attractive research direction as it can improve the efficiency dramatically. However, automatically generating garment models faces many challenges. Due to the different topologies and various fashion apparels, it is difficult to design a unified generation pipeline. Moreover, the generated garment design is not easy to be re-targeted onto another body shape which limits the customization of the models. Some works tried to solve these problems by providing fixed topologies for cloths, such as pants or T-shirts [17], or through user-assisted inputs [51].

Researchers have tried to reconstruct 3D garment mesh from various inputs, such as sewing pattern images, multi-view garment images, garment masks, or even single 2D garment images. Most of previous studies can be divided into three main

categories: reconstruction based on sketch, 3D re-shaping or reconstruction based on image, and depth-based reconstruction.

In the past, it is one of the most popular ways to generate garment models with sketches. Various methods are explored in past few years, such as grid and geometric method [14, 17, 60], multi-view sketches method [61], front and back sketch images method [51] and so on. Plenty of efforts are paid to improve the reality of the generated garment model as well, for example, the context-aware method [15]. However, all those methods require domain knowledge on garment sketching, which is a critical limitation.

Figure 1.1: Garment sketches used in creating garment models [17]

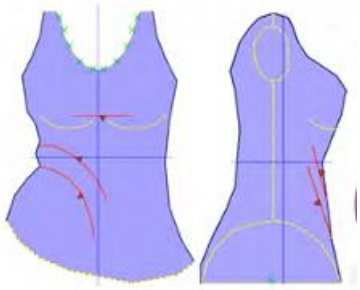


Figure 1.2: Multi-view garment sketches as input [61]

At the same time, garment reconstruction based on images are studied. Some of the works perform well but they may require complicated inputs, such as the multi-view garment images [19] and sewing pattern images [50], which are not easy to collect. Typically, such datasets need complicated setup, experts with relevant skills and high cost.

Figure 1.3: Acquisition setup for multi-view images [19]

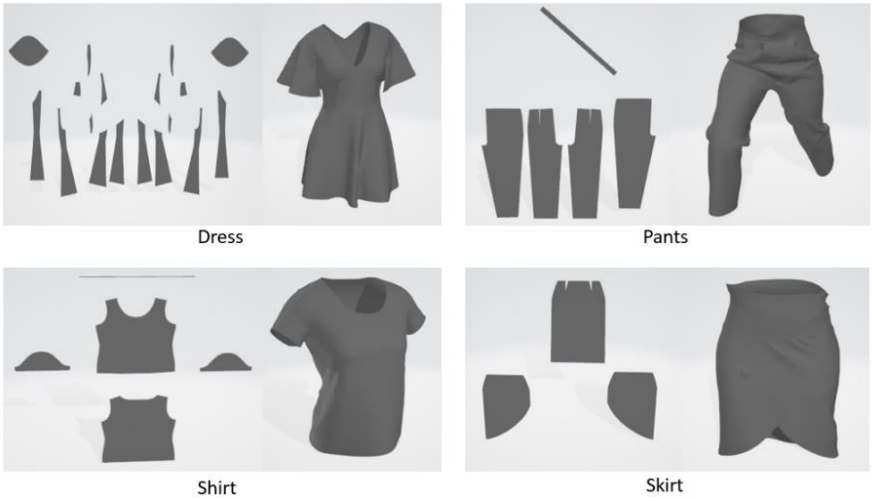


Figure 1.4: Garment sewing pattern samples [50]. 4 cases are shown in the image, namely dress, pant, shirt, and skirt

Recently, researchers have made significant progress in image-based 3D reconstruction with the benefits from learning shape representation through deep neural networks [52, 53, 54, 55, 56]. However, due to the diverse topologies, open surfaces and complex geometric details of garments, reconstructing 3D garment models with a single image is still a difficult task. So some of the methods use parametric models to overcome this issue by providing shape priors of garments [34, 57, 58], while some other works try to reconstruct the garment mesh on top of the SMPL body mesh [58]. Depth information is also useful in garment reconstruction [21], but few works are implemented due to the difficulties in data collection.

Wrinkle prediction Wrinkle prediction is to predict the wrinkles of garments based on different human body pose, shape and garment type. This is an important part for 3D garment animation. For clothed digital human animation including wrinkle prediction, physics-based simulation (PBS) [24-33, 49] is still the predominant approach. However, the classical PBS pipeline not only requires expert knowledge but also is very laborious and quite time consuming. To achieve desired results by PBS, it requires to edit the 2D garment shape with different patterns, place modified results on the digital character manually and finely tune the parameters. What is more, high quality PBS methods require tens of thousands or more vertices, which are expensive to compute, difficult to implement and control. And generally, it is not trivial to differentiate, which is fatal to recent deep learning techniques. In view of these difficulties, Pose Space Deformation (PSD) models [89, 90, 91, 92] and Linear Blend Skinning (LBS) [93, 94, 95] are explored to improve the efficiency and make it possible when high performance is required or computational resources are limited. But realism is highly compromised. In summary, there is a trade-off between realism and performance for classical computer graphics approaches.

Recent years, the deep learning techniques provide alternative and efficient ways for animating 3D garments. Much research has been conducted through deep learning to learn and predict the garment deformation based on various factors. Based on the exploration, three factors can affect the clothing deformation, namely body shape, pose and garment style (or garment geometry). Generally, they extract the features of these three factors from the input data and utilize deep learning techniques to train on one or more sets of these features to predict the displacement or deformation of the garment, for example, some methods [44, 42] predict deformations based on pose for a fixed shape, while some other methods [46, 47] predict deformations due to body pose and shape for a fixed garment style. However, most of the pioneer works do not model the deformations based on the overall effect of three factors jointly, even though some of them are intertwined. So some other works [45, 48, 47] try to improve the predictions by joint models of body shape and pose but they frequently generate results that cannot retain high frequency details even with a fixed garment style.

Therefore, some of the works focus on solving the over-smooth issue. The most impressive method is to decompose the deformations into high and low frequency components which are predicted separately and then joined together to give more details of the wrinkles [2].

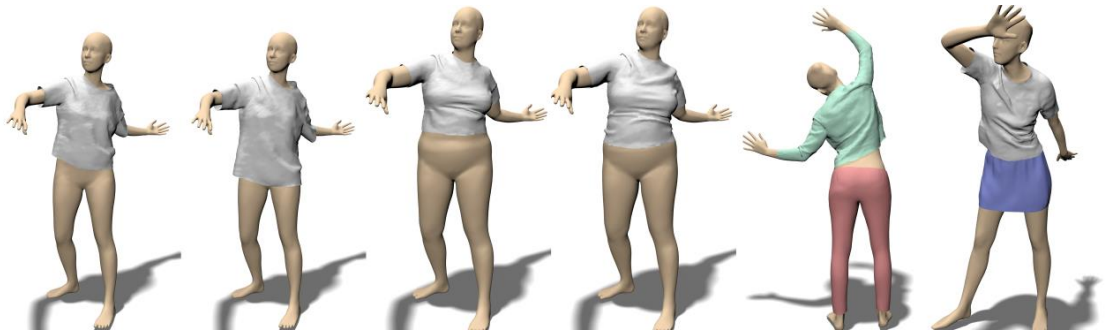


Figure 1.5: Examples of output of TailorNet model [2]. It solves the over-smooth issue and can retain the details of wrinkles

While the above methods focus on improving the reality of the predicted wrinkles, some other works focus on the generalization of the model for various garment styles, such as DeePSD [10] which describes the motion by utilizing the deep learning techniques to map the space of garment templates to the space of pose space deformation.

Figure 1.6: Qualitative samples of DeePSD [10]

Prior work [7] also tries to predict the garment wrinkles in 2D space instead of 3D space, which recasts the 3D cloth deformation information as an 2D RGB image, where the color of the RGB image represents displacement of the garment vertices. Then it learns cloth deformations in image space through CNNs.

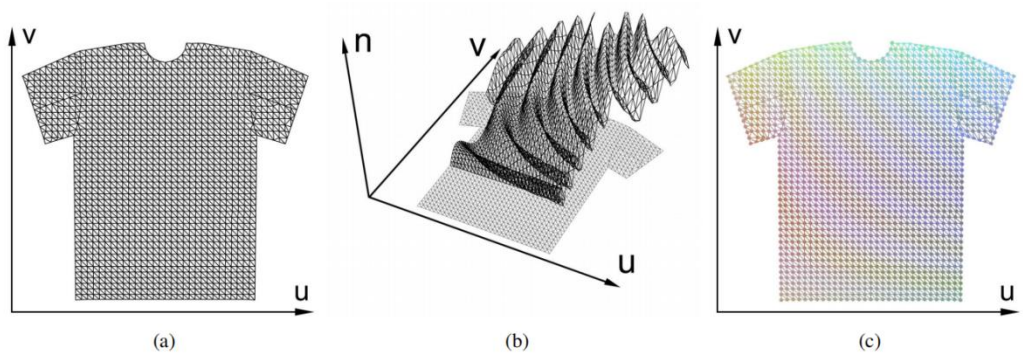


Figure 1.7: Data representation method for cloth deformation [7]. Left: with vertices' UV coordinates, garment mesh is represented in texture space. Middle: add displacement values to vertices. Right: the displacement is visualized by converting displacement values into RGB values at each vertex.

Texture transfer Texture transfer is to transfer the texture from 2D garment images onto 2D or 3D human or garment models. It is extremely helpful for realistic photo rendering of clothed humans, which is important in the various applications mentioned above. However, even with multi-view images, texture generation is still challenging. Although many techniques have been developed for combining the partial textures, such as blending [62, 63, 64], mosaicking [65,66,67], graph cuts [42], flow-based registration [3, 68, 69, 70, 71, 72], it is still difficult to reduce ghosting and stitching artifacts in synthesizing the partial textures generated from multi-views.

Current works can be divided into the following categories: image-based re-posing [68, 73, 74, 75, 76, 77, 78, 79], image-based virtual try-on [80, 81, 82, 83, 84, 85, 86], exemplar-based image translation [96, 97], fitting 3D SMPL human models to images [34, 35, 36] or mapping texture to 3D meshes of 3D scanning people or 3D garment templates [1, 4, 20, 87, 88].

Image-based person re-posing typically utilizes the deep learning techniques to learn the global human body structure and appearance details from images and then tries to produce novel poses of the same person with texture by synthesizing image pixels. Some of the methods use unsupervised settings while some of others use autoregressive models with ground truth. In the autoregressive models, the product of conditional distributions of pixels is computed in a pixel-by-pixel

manner to generate the joint distribution of pixels. The most popular network used for autoregressive models are the generative adversarial networks (GANs). It learns a generator and a discriminator at the same time. Samples are generated by the generator which will be discriminated from the real ones by the discriminator. With the adversarial loss instead of L1 loss, generally GANs can generate sharp images.

Figure 1.8: Examples for image-based re-posing method [74]. Left column is input, middle column is the ground truth and the right column is the output with texture for different poses

Image-based virtual try-on aims to modify the image of a garment item and then drape the modified one on a target person or directly transfer the garment textures of one person onto another person with various pose and shape in image space. Before extracting the features of the garments, it will estimate the human shape and pose from images. Then based on the predicted human pose and shape, it disentangles the clothing features from the body. It also needs to infer the surface appearance to recover the missing parts of the original image which are visible in the target image. With all these ready, the models start to transfer the clothing to a target person.

Figure 1.9: Examples for image-based virtual try-on [81]. For each row, first two images are the input while the last two images are the output, garment texture of which have been exchanged

With an exemplar image, the exemplar-based image translation methods [96, 97] can synthesize realistic images from the input in the form of segmentation mask, edge map or pose key-points. Typically, they will learn the alignment of semantically corresponding objects or features between input images and exemplar images first. Then, with the learned alignment, the network synthesizes images based on the exemplar images. The output has consistent style, liking the color and texture, with the semantically corresponding objects in the exemplar.

In view of the issue of local style 'wash away' in the ultimate image, the Cross-domain Correspondence Network [96] is proposed to address the problem. It can transform the images from different domains to an intermediate feature domain to establish reliable dense correspondence. Another Feature Transport Network [97] is designed to overcome the many-to-one matching issue by introducing optimal transport to match two sets of features as a whole.

Figure 1.10: Exemplar-based image synthesis [96]. The 1st row is the exemplar images and the second row is the synthesized images from the network

The above methods are all in image space which cannot be used for 3D garment or human models. As clothed human animation is performed in 3D space, so more and more research started to explore texture transfer in 3D space. The most popular methods are fitting 3D SMPL human models to images to extract texture or mapping texture to 3D meshes of 3D scanning people. For fitting 3D SMPL human models to images, it is difficult to deal with complex poses which can lead to bad texture quality easily. Even mapping texture to 3D meshes of 3D scanning people gives a better result, the 3D scanning is quite laborious, expensive and time consuming which limits the variety and availability of clothing textures. To break the barrier of lacking 3D scanning data, mapping from images of clothing to the surface of a 3D parametric garment template is proposed by pix2surf [1].

Even some works did well in one of the sub-fields, but there is still no paper proposing a full pipeline from end to end to animate clothed human with both human and garment textures from a single image input. And some of the works have a bad efficiency for training or testing which may take several days to train, while some of other works require complicated inputs which are not easily collected.

In this report, we mainly study in two aspects, namely, how to improve the completeness of the pipeline and how to improve the efficiency for both training and testing.

1.1 Research Objectives

For virtual garment try-on, most of current works are still focusing on separate sub-fields only, following are the challenges we try to address:

In view of current incomplete pipeline, in this report we will try to contribute to the completeness of the pipeline. And we will try to use data which are easily available to allow easy collection of datasets and give a highly efficient training process. This report will be based on the work of TailorNet [2], which provides a method to predict the garment wrinkles depending on various body shape, pose and garment type. It can only predict the wrinkles when the 3D garment models are given. It cannot start from the garment images through constructing the 3D garment models. So, we will focus on completing the pipeline by adapting the model with ours to enable the model to start from garment images instead of 3D garment models.

1.2 Research Scope

In view of the issue stated in previous part, we will focus on the contribution of the completeness of the pipeline by developing a network to generate 3D garment models from garment images to enable the pipeline to start from real images instead of 3D information. Existing 3D garment reconstruction models typically require complicated inputs and is difficult to be combined with 3D garment animation models. So most of the models of 3D garment animation or wrinkle prediction start from the 3D garment models instead of 2D real images. With the help of our model, they will be able to start with widely available 2D images. So, we mainly study how to generate 3D garment model parameters with garment images through various machine learning algorithms and make the predicted 3D garment parameters easily available to animation models.

1.3 Contributions

In this work, we propose a fast and effective model to generate 3D garment model parameters which can be easily fed to other animation models. The main contributions are summarized as below:

- We greatly simplify the input requirements for generating 3D garment models. Instead of complicated inputs, our model only needs normal garment images to generate the 3D garment model parameters
- Unsupervised techniques are applied which allow us to have various choices for data collection and no data labels are required
- The generated 3D garment model parameters are easily utilized by garment animation models. It helps to improve the completeness of the pipeline and make it able to start from images instead of 3D garment models

Chapter 2

Literature Review

In this chapter, we discuss the background information of the related works for 3D garment mesh reconstruction (Sec 2.1), garment wrinkles' prediction (Sec 2.2) and garment texture transfer (Sec 2.3).

2.1 Garment Mesh Reconstruction

Garment mesh reconstruction refers to reconstruct the 3D garment mesh from various inputs. Before machine learning became popular, many efforts have been put into developing software for 3D garment designs. But most of them take 2D sewing patterns as input for creating 3D garment models, such as Mavelous Designer and Optitex. The high modelling cost limits its application in real life. So, much research started to explore methods that can automatically generate 3D garment models. With the help of machine learning, automatically generating 3D garment models has been greatly promoted through various ways. The previous works can be divided into three main categories: reconstruction based on sketch, 3D re-shaping or reconstruction based on image, and depth-based reconstruction.

Geometric approaches generally have roots from the CAD community. Prior work [13] tries to fit the 3D features of clothing templates onto various body shapes to automatically process the Made-to-Measure, while some other works takes 2D patterns as input, for example, 2D sketches are utilized to generate garment models through grid and geometric methods [14, 17, 60]. Virtual-Garments [14] sketches the contours and seamlines around a virtual mannequin. Then it fits garment panels to sketches followed by approximating these panels with developable surfaces for manufacturing garment. A data-driven approach [17] is

proposed, which takes the users' sketches with desired fold patterns as input to estimate the parameters of body shape and garment, which is achieved through a shared shape space. Users can specify the desired characteristics with the shared shape space for multimodal input without seams. Running garment simulation is not required at design time. With a front or back outline of the garment, the overall shape of the cloth can be inferred [60]. The methodology of the inference is that it utilizes the distance between the 2D sketch silhouette and the human model to predict the distance between the 3D garment and human model. With the predicted distance information together with the border lines and silhouette of the sketch, it generates the garment surface.

Figure 2.1: Examples of sketch and generated garment pairs [60]. For each pair, left side is the sketch while the right side is the generated garment model.

To improve the reality of the generated garment model, a context-aware method [15] is proposed based on observing the key factors which may relate to the shape of garments. Other work [61] takes multi-view sketch as input. Other than the borders, seamlines and silhouettes, it allows user to draw the contours of internal folds as well. The silhouette edges can be identified and tied to silhouette strokes on the sketch which represents the discontinuous sets of non-planar curves on the 3D model. With front and back sketches of garment, realistic 3D garment models can be automatically generated [51]. Several feature points from the silhouette are selected to show the distribution characteristics which are used to classify the garment types. A critical limitation for these methods is that professional knowledge is required for garment sketching.

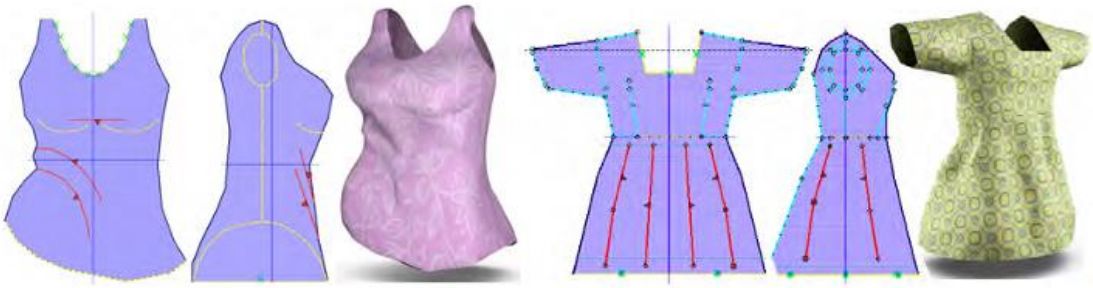


Figure 2.2: Multi-view sketches are utilized to generate 3D garment models [61]

For 3D reconstruction approaches based on images, 3D models are generated directly from garment images or videos. A multi-view stereo algorithm [19] is proposed to generate 3D garment models from multi-view garment images. The setup and data collection are quite time-consuming and expensive. Other work [50] utilizes sewing pattern images to predict 3D garment models. A universal method is proposed [50] with generative network. It is capable to handle different garment topologies as well as various fabric materials and sewing patterns. A set of colour markers are used [18] on the surface of cloth to recover dynamic 3D cloth mesh with consistent connectivity. A model [20] is created to extract clothing from a clothed person. With multi-view 3D scans as input, it can automatically segment each part of clothing and estimate the body pose and shape under the garment. Even some of the works give a nice performance, those approaches require specialized hardware or domain skills and they do not work with existing garment photos. The data collection process is time-consuming, tedious and expensive which make the research difficult through these approaches.

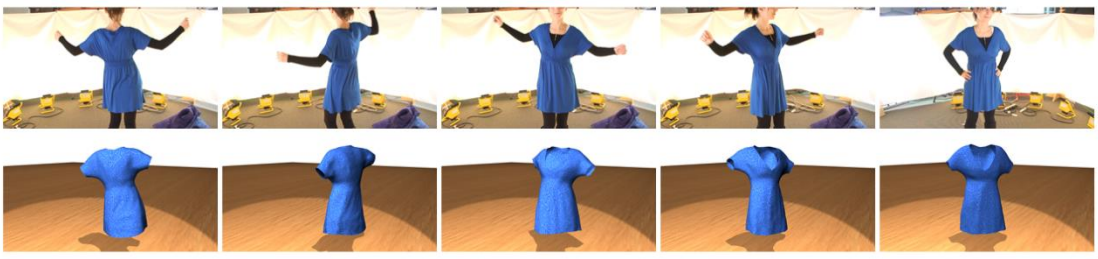


Figure 2.3: Capture 3D garment models from multi-view input images [19]

Figure 2.4: Segmentation results from 3D scans [20]

Recently, researchers have made significant progress in image-based 3D reconstruction with the benefits from learning shape representation through deep neural networks [52, 53, 54, 55, 56]. But due to complex geometric details, open surfaces and diverse topologies, it is still challenging to recover garment models in 3D space from a single image. So, some of the methods use parametric models to overcome this issue by providing shape priors of garments [34, 57, 58], while some other works try to reconstruct the garment mesh on top of the SMPL body mesh [58].

Image-based template reshaping is implemented by utilizing parametric human and garment models and estimating the shape parameters. The input of [22] is a single image of clothed human. With the input images, it estimates human shape and pose first. Then, A raw 3D mesh for the garment is created by a semi-automatic approach followed by recovering the details by using shape-from-shading. In this method, it requires user to input the outline labeling of garment and estimated pose parameters. Image pairs of clothed and unclothed mannequin from the same viewpoint are taken as input [23]. Then the drafted patterns are fitted to the input image by analyzing the outlines.

Depth information is also useful in garment reconstruction. With RGBD sequence of a garment, a network is proposed [21] to construct garment models based on garment RGBD sequence. It uses Kinect to scan the garment to get raw RGBD sequences based on which it builds a rough shape with KinectFusion. Then the

garment design attributes and components will be identified by the detectors and classifiers from the RGB images. After that, it stitches the chosen templates based on the components and designs. The stitched templates are fitted to the raw garment shape generated by KinectFusion.

Figure 2.5: With RGBD sequence as input, the network [21] detects and assembles the suitable components to produce a high-quality 3D garment model

2.2 Garment Wrinkles' Prediction

A main part of animating digital humans in clothing is the garment wrinkles' prediction. Physics-based simulation (PBS) and data-driven models are the main approaches used for clothing animation. For data driven models, they are generally learned from PBS generated data or real captures. The predominant approach for animating clothed digital humans is still the PBS. However, the classical PBS pipeline not only requires expert knowledge but also is very laborious and quite time consuming. To achieve desired results by PBS, it requires to edit the 2D garment shape with different patterns, place modified results on the digital character manually and finely tune the parameters. And high-quality animations are very computationally expensive and require simulating millions of triangles [24, 25, 26, 49]. So many works try to improve the efficiency by using low-resolution simulations to predict wrinkles [27, 28, 29, 30], or utilizing simpler models, such as mass-spring [33] and dynamics based on position [31, 32]. They sacrifice accuracy and physical correctness for higher speed.

Figure 2.6: Examples of physics-based simulation output [49]

In order to make animation easier, many efforts are put on works of learning efficient models from PBS generated data or real data. To achieve reality, many works try to extract real garments on human body from images [34, 3, 35, 36, 37], dynamic scans [20, 38] or RGBD [39, 40] and drape them to new body shapes. But this only re-animates a new body shape with same motion. Some other works try to learn models depending on poses from real captures [41, 42, 43]. But data collection for these models still is a challenging part which is quite time-consuming and high cost. Insufficient data always limits the quality of the model output. A good solution to this problem is to use the generated data by PBS. Some old models use the linear regression and nearest neighbour search to predict clothing deformation based on pose [44, 29, 30], or shape and pose [45].

Most of recent research try to utilize deep learning to learn how to animate the garment models by related factors or data representation. Based on the exploration, typically three factors can influence the cloth deformations, namely body shape, pose and garment style. A simple two-component model is proposed [41] to analyse the statistics of the garment layer. The model regresses any semantic parameter to the layer parameterization space. It reduces the layer information by PCA technique and use neural network to regress the generic parameters. A recurrent neural network [46] is used to train a learning-based model to regress garment wrinkles based on body shape and dynamics. For a given 3D body, Graph-NNs [47] is utilized to adjust the 3D garment template according to the body. By deep network technique, the features of garment are

extracted at different detail levels, such as point, patch and global features. These features are then fused to model the body-cloth interactions. Some other methods [44, 42] predict pose effect on garment deformation for a fixed shape. DeePSD [10] focuses on the generalization of the model for various of garment styles. It utilizes graph convolutional network and MPL by unsupervised learning to propose a methodology which is capable to work with cloth of multiple layers, various topologies and complexity. The method can generalize to completely unseen garments with complicated details as well. What is more, the model achieves real time performance and can be deployed on portable devices efficiently.

Figure 2.7: Output of DeePSD [10] with diverse topologies and multiple layers

TailorNet [2] utilizes deep learning techniques to predict wrinkles of clothing in 3D depending on body shape, pose and garment style. In the heart part of TailorNet, the deformation is decomposed into high and low frequency components, which are predicted separately. An MLP is used to estimate the low-frequency component based on the three factors, while the high frequency component is estimated based on shape-style models with specific pose. This strategy solves the over-smooth issue faced by previous works. It not only can retain the detailed deformations but also is much faster than the PBS method.

Instead of directly handling the animation in 3D space, the 3D task is converted to 2D task [7] by representing 3D garment deformation as 2D RGB images. Then the animation of a 3D cloth can be implemented through a sequence of the RGB images. With the RGB images, it learns the garment deformations by CNNs in image space.

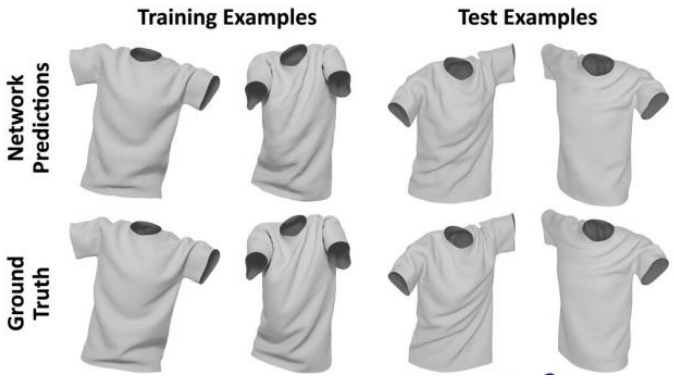


Figure 2.8: Output predicted in 2D space comparing with ground truth [7]

Deepcloth [8] can describe garments with various topologies and shapes by representing geometry of garment with the "UV-position map with mask". With the garment features learned from the UV space, it can edit and transit garment shape. PBNS [11] proposes an unsupervised method to learn realistic cloth deformations in pose space for clothed humans by using deep learning. With unsupervised methodology, it overcomes the issue of lacking data. It shows consistent animation results for garments and meaningful wrinkles and folds depending on poses, but the training efficiency is not high.

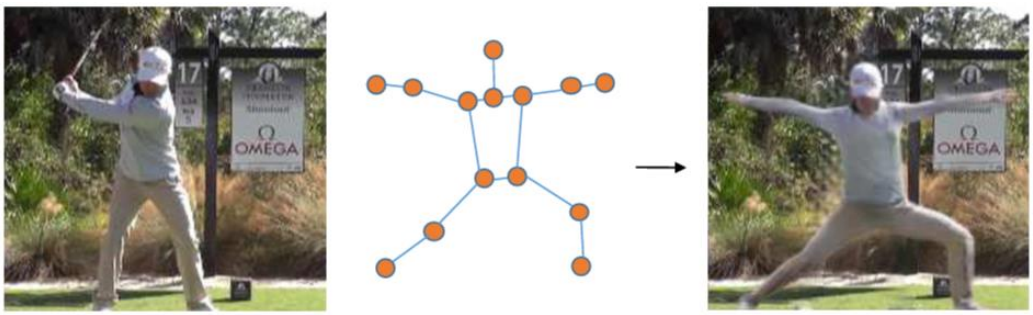
2.3 Garment Texture Transfer

In order for the generated garments to be more realistic, texture transfer is an important part. Texture transfer aims to transfer the texture from 2D garment images to 2D or 3D garment models. It is still a challenging task even there are many techniques are developed to assist the process. The most difficult parts lie in synthesizing the partial textures generated from multiple views and learn the correspondence between textures and garment models.

There are mainly five research lines for the past works, namely image-based person re-posing [68, 73, 74, 75, 76, 77, 78, 79], image-based virtual try-on [80, 81, 82, 83, 84, 85, 86], exemplar-based image translation [96, 97], fitting 3D SMPL human models to images [34, 35, 36] or mapping texture to 3D meshes of 3D scanning people or 3D garment templates [1, 4, 20, 87, 88].

Image-based person re-posing tries to learn human body structure and appearance details and then produce novel poses of the same person with texture by synthesizing image pixels. With the new Soft-Gated Warping Generative Adversarial Network, several challenging problems are resolved [68], such as different viewpoints, occlusions and significant changes in appearance, by introducing the spatial displacements and geometric variability. VariGANs [73] predicts the overall appearance of object by variational inference. Instead of a single pass, it generates the target image from coarse to fine. The coarse-to-fine strategy is implemented in other methods [74, 82] as well. The pose, foreground and background factors of the input image can be manipulated by learning a disentangled representation of these factors and samples new embedding features to generate textured human images, which provides more control in the process of generation [75].

A fully unsupervised strategy [76] is proposed by using generative adversarial learning to render the same human with new pose. It can not only synthesize the parts visible in the input image for novel views but also hallucinate those that are invisible. A Generative Adversarial Network [77] is proposed with deformable skip connections in its generator. The network can resolve the problem of pixel misalignment due to the differences among poses. Instead of common L1 and L2 losses, it utilizes the nearest-neighbour loss to improve the matching of details between the output and target images. An image is first split into body part and background layers [78]. Then the body appearance is adjusted according to new pose and the hole of the background layers is filled, followed by synthesizing the new appearance with the filled background to generate new image with target pose. A generative network [79] is built which consists of three parts, an appearance encoder, a semantic map encoder and an appearance generator. The appearance of image is extracted by the appearance encoder while the semantic map is generated through the semantic map encoder. Under the guide of the reference image and semantic map predicted, it synthesizes texture for the output images.



Source Image Target Pose Synthesized Image

Figure 2.9: Example for image-based person re-posing [78]. It takes source image and target pose in 2D skeleton to generate human image with texture for new pose

Image-based virtual try-on is to transfer garment texture from one person to another in image space. Swapnet [80] splits the process into two processes: separating the clothing features from the human body in the input image and synthesizing the clothing texture onto a novel human body. The training pairs are generated from single image through data augmentation by utilizing a weakly supervised approach. The appearance can be automatically transferred from person-to-person based on explicit representations of 3D parametric human model and synthesis of photographic image through deep translation network [81]. With a pair of source and target images as input, the texture of source image is transferred onto the target image without changing clothing segmentation layout and target shape [81]. Similar to [73, 74], Viton [82] also takes the coarse to fine strategy. It first synthesizes a coarse image of the target garment item and wears it on the human. Then it uses a refinement network to refine the initial blurry areas. Instead of computing the correspondences of interest points, the clothes are transformed to human body by a thin-plate spline transformation learned through a Geometric Matching Module [83]. The boundary artifacts are alleviated by a Try-On Module for warped clothes and the output looks more realistic through a composition mask which integrates the rendered image and warped clothes [83]. For a given person with desired pose, Vtnfp [84] first transforms the target garment to warped form which is compatible for the preferred pose. Then, it delineates clothing and body parts by a segmentation map, after which, it synthesizes the image by fusing the person image, warped clothing and segmentation map. It can better retain the features of clothing and body. With

a person image as input, a new image can be produced by synthesizing desired clothes with the input image [85]. The image can be adjusted according to preferred human poses [85]. From the target image, a human parsing map is synthesized which is utilized to fit the target pose and shape of clothes. The clothes appearance is then warped into the parsing map. The texture details of clothes are recovered by a refinement render and it uses the masks of multi-pose composition to remove some artifacts.

Figure 2.10: Examples for image-based virtual try-on [85]: with a reference human, target cloth and pose, it transfers the garment texture to the human under the desired pose

The exemplar-based image translation methods [96, 97] focus on synthesizing realistic images from the input in a distinct domain with an exemplar image. The output has consistent style, liking the color and texture, with the semantically corresponding objects in the exemplar.

Although much research has been conducted in image space, recently more and more research start to focus on texture transfer in 3D space. Most of the works try to fit the 3D SMPL human models to images or mapping texture to 3D meshes of 3D scanning people or 3D garment templates. Multi-garment [34] takes a few frames of a video as input to predict the clothing draped over the SMPL model. It can reconstruct the garments individually and transfer the geometry and texture

of the predicted garments to a novel human model. Octopus [35] improves the alignment between the input and output images by allowing the information to flow both forward and backward with bottom-up and top-down streams.

Source Images Target Re-dressed Target

Figure 2.11: Examples for fitting 3D SMPL human models to images [34]. With target body shape and pose, it transfers the texture from source image onto the target human model

The geometry and texture can be inferred by image-to-image translation in UV space [41]. With segmentation layout maps and partial textures from input image, the method can predict the complete segmentation, texture and a displacement map. The multi-mesh representation is introduced by Clothcap [20] to fit sequences of 3D scans for predicting high quality capture and segmentation. With the segmented cloth pieces, it renders them to produce plausible clothing layered appearance. It requires 4D scanned point cloud with texture as input. Double depth maps are employed in the non-parametric approach [88]. The first depth map is visible while the other one is “hidden”. With the 2D maps, higher resolution output can be generated with much lower dimension. An obvious disadvantage for these methods is that they require 3D scanning dataset, collection of which is very time-consuming. To solve the issue of lacking 3D scanning dataset, pix2surf [1] suggests building a dense mapping from garment images to the surface of 3D garment. The clothing images can be easily collected. It takes pairs of front and back garment images as input to train the model by aligning a parametric 3D garment model to these clothing images non-rigidly.

Figure 2.12: The model in pix2surf [1] learns the mapping between the images and the 3D parametric garment model. Given a garment image, it can transfer the texture onto the garment.

Chapter 3

Unsupervised Garment Mesh Reconstruction

3.1 Introduction

Most of recent research only focus on one part of the garment virtual try-on. Even some works have done a good job, but it is difficult to combine them together to create an end-to-end pipeline which starts from normal 2D garment images and ends with animating the 3D garment models according to different human pose and shape in virtual environment. In view of this, we try to improve the TailorNet model [2] by adding our garment mesh reconstruction network.

TailorNet proposes a pipeline to animate 3D garment models by predicting the wrinkles depending on different garment style, human pose and shape. The deformation is decomposed into high and low frequency components to improve the over-smooth issue. A MLP is used to predict the low-frequency component based on the three factors while the shape-style models of specific pose are utilized to predict the high-frequency component. K prototype shape-style pairs of high frequency deformation are computed separately and an RBF kernel is used to mix the K prototype pairs to generate the final deformations for high frequency. Then the predicted high and low frequency components are synthesized to generate the unposed garment output. After that, the standard skinning is used to pose the output to desired pose. The 3D garment models reserve the wrinkle details from PBS while it runs much faster than PBS. TailorNet generates more realistic wrinkles than prior works.

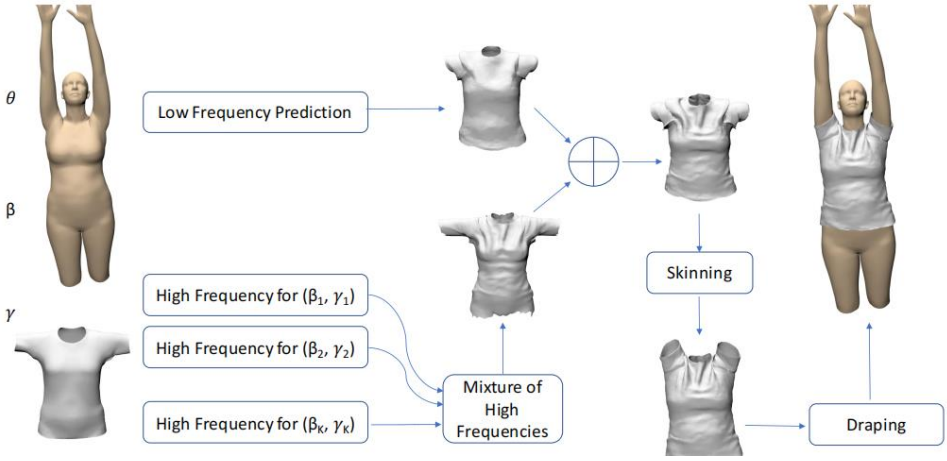


Figure 3.1: Overview of the TailorNet. With the input of garment style γ, body pose θ and human shape β, high and low frequency components of the garment deformations are predicted separately.

Even the network gives a nice result on predicting the wrinkles of garments, it is not fitted to real images. It can only be applied for 3D garment models. So users need to input the 3D garment style parameters before running the network. This may give troubles if the users are not familiar with 3D garment models. And the style parameters are difficult to be estimated directly by users. Therefore, we propose our network to solve this issue and mix it with TailorNet to create a new network which can automatically generate the 3D garment model style parameters from real images and enable the network to start from real images.

Most of the prior works for 3D garment model reconstruction require either ground truth or complex input data. The lack of dataset will be a big problem. Even the data can be collected by the researchers, such as the 3D scans and simulations, it generally takes a quite long time and the cost is quite high. To avoid the inconvenience faced by previous works, unsupervised learning has been implemented in our network. It is able to utilize the widely available online garment images to train the model.

In our network, it first uses multiple levels of convolutional layers and maxpooling layers to deal with the input garment images and followed by a MLP network. We have conducted extensive experiments to test the network on various types of garment images. The results show that the network is capable to predict the 3D

garment style parameters properly under various circumstances, such as long or short sleeve, long or short vertical length, wide or narrow horizontal width and so on. It also can work properly when the garment appears in different positions in the image.

3.2 Methodology

3.2.1 Problem Formulation

As our purpose is to improve the completeness for an end-to-end garment mesh animation from image inputs, we need to build a network to link the input image and the TailorNet model as well as modify the TailorNet model to assist the training process of the new network.

In order to link the input image to the TailorNet model, a garment mesh reconstruction network is proposed which takes garment 2D images or masks as input to infer the 3D garment model. Here we take only the front-view of the garment images which is easy to collect from various sources, such as online shopping websites, existing garment image datasets and so on. The output will be garment style parameters that can be directly fed to TailorNet model for animation purpose.

3.2.2 Overall Pipeline

Based on the analysis in problem formulation, the following overall pipeline is proposed. In our approach, it includes two parts. The first part converts input real garment images to 3D garment mesh style parameters, which can be directly utilized in animating of 3D garment models. In this part, unsupervised learning is implemented, which allows to train the model simply by the widely available online garment images. It is also able to give stable output when the location of garment is shifted in the image. The second part is to animate the 3D garment mesh based on body pose, shape and the garment style parameters predicted in the first part. The first part is newly added and the second part mainly is inherited

from the TailorNet model. With these two parts of networks, the model allows to animate the garment model in 3D space from 2D real images. More technical details will be further explained in the following sections.

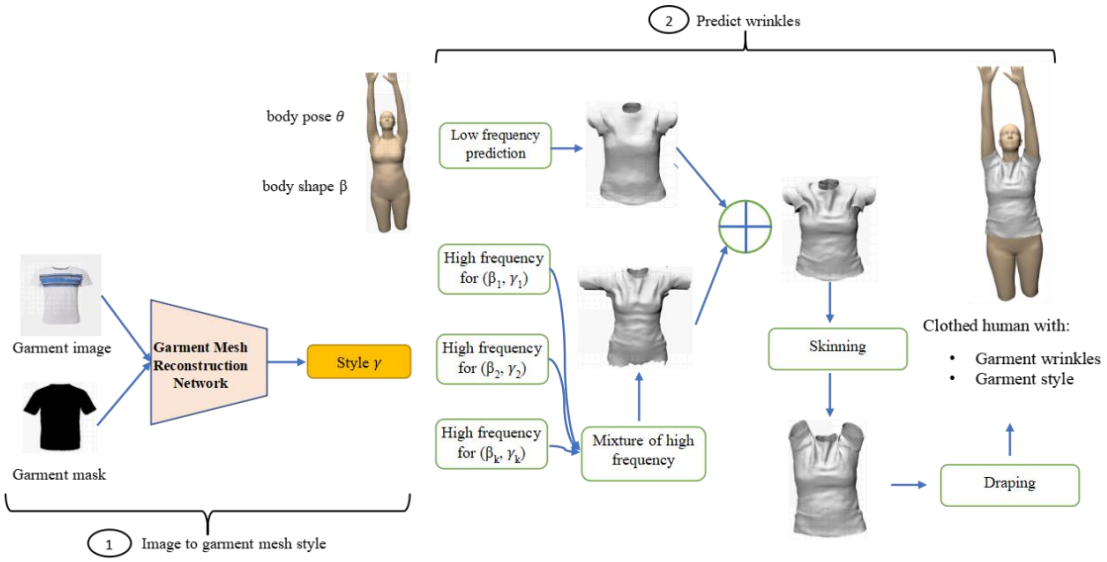


Figure 3.2: Overview of the network architecture. The network can be divided into two main parts. Part 1 (Newly designed) mainly focuses on predicting the garment style parameters from input garment images while part 2 (Inherited from TailorNet) focuses on animating the 3D garment model based on human pose, shape and the garment style predicted from part 1.

3.2.3 Details of Model Architecture

Part 1 focuses on predicting the style parameter of the 3D garment model from 2D image input. It is an unsupervised optimization network without referencing any 3D information of the input images by matching the silhouette between the garment image/mask and the projection of the generated 3D garment model. The working theory is shown in the following diagram:

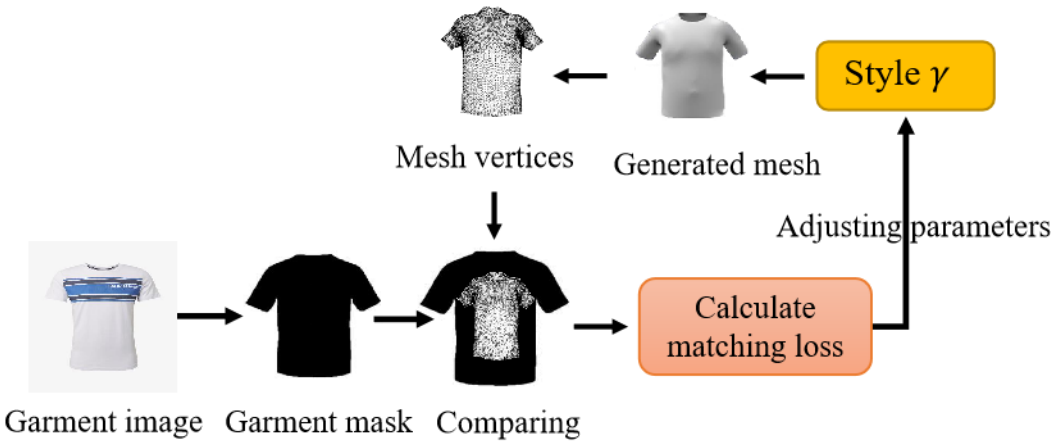


Figure 3.3: Illustration of working theory for part 1 of the network. It takes garment image or mask as input and the network is optimized by calculating the matching loss between the input image and projection of the generated 3D garment model

In order to achieve the above process, deep learning architecture is implemented. The core architecture of part 1 of the network starts with three convolution and maxpooling layers. This part is to do some initial treatment of the input images and reshape the input images to desired dimensions. It is followed by a MLP network to predict the 3D garment style parameters based on the output of the convolution and maxpooling layers. Then the predicted garment style parameters are fed to the garment mesh generator to generate the 3D garment mesh. With the generated 3D garment mesh, we project the vertices onto 2D images. Then the projected 2D vertices are compared with the input image to optimize the network until the network is stable. In this part, the input can either be 2D real garment images or their masks.

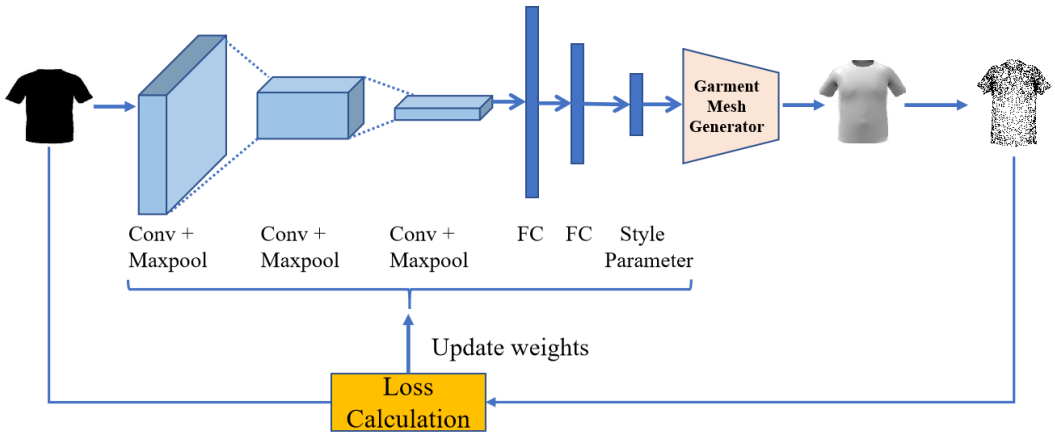


Figure 3.4: Overview of the core architecture of part 1 of the network

The architecture of the second part of the network is modified based on TailorNet model with minor changes to adapt to the whole network. The details of the network have been elaborated in part 3.1. The main difference is that the garment style parameters now is predicted instead of being given by users. Choosing this model as the second part is because it has the following outstanding performances: First, it predicts garment wrinkles based on three factors, namely body shape, body pose, garment styles, which is more reasonable compared with other models based on one or two of the three factors. Second, it can retain the garment wrinkles' details. It solves the over-smooth issue by decomposing the deformations into low and high frequency components and predicting the high and low frequency components separately. Lastly, both training and prediction efficiency are very high. Unlike other models, which may take several days for training, TailorNet only takes a reasonable time of several hours for training with excellent training results.

3.2.4 Loss and Training

From the architecture above, we know that the part 1 of network does not require any ground truth during training. Unsupervised learning is utilized in this step by calculating the matching loss between the input image/mask with the projected garment mesh vertices. Before we calculate the matching loss, we must align the position of the garment or its mask in the input image with the projected mesh vertices. Otherwise, the training process will be wrongly guided. After alignment of the position, the training loss will be the MSE matching loss of the pixel values

between the input garment or its mask image and the projected mesh vertices image pixel-wisely:

$$Training\ Loss = \frac{\sum_{i=1}^{n}(projected\_mesh\_vertices[i] - input\_image/mask[i])^2}{n}$$

The projected mesh vertices are in black colour which has a pixel value of 0 and its background colour is white which has a pixel value of 255. Similarly, the background colour of input images is also white. In the above equation, we compare each pixel pair of input image and the image of projected mesh vertices. The training loss is calculated by summing up the square of the differences first and then divided by the total number of pixels to get the mean square error (MSE). As illustrated in Figure 3.5, If the area of projected mesh vertices is much smaller than that of the garment or mask, many background white pixels from the projected mesh vertices image will be compared with the coloured pixels from garment image or black pixels from the mask image, which will make the training loss very large. So it will enlarge the generated mesh to reduce the loss. Similarly, if the area of projected mesh vertices is much larger than that of the garment or mask, many black vertices from the output image will be compared with the white background pixels from garment or mask image, which will make the training loss very large as well. So it will reduce the garment mesh size to decrease the training loss. Thus, the generated 3D garment mesh will be gradually fitted to the input garment image or mask during training.

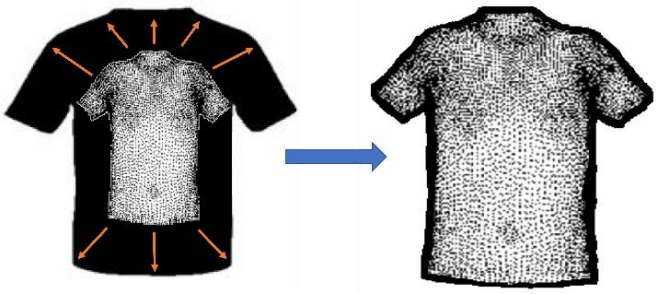


Figure 3.5: An illustration of the training process. To minimise the loss, the network will drive the outputted garment style parameters to generate garment mesh which overlaps with the input image/mask as much as possible.

3.3 Experiments

3.3.1 Datasets

Front-view is the only requirement for the images. As no other specific requirements for the input garment images, so the dataset can be sourced from anywhere and it does not need any 3D information about the images. We have collected 1150 RGB T-shirt images from various sources, such as online stores Zalando, Jack and Jones, open-source dataset Fashion Usage and so on. Some of the photos are downloaded from the websites while some of them are directly selected from existing open garment datasets. To train the model properly, we have collected T-shirt images with various styles, such as long/short sleeves, small and large sizes, wide or slim styles and so on. Out of the 1150 RGB images, 80% of them are used for training while 20% of them are used for testing. The dataset is split randomly.

Figure 3.6: 2D garment image dataset collected from various sources

Generally, due to actual garment images are taken from actual garments, they are lacking extreme cases, such as extreme long sleeves. To test the model with extreme cases, I have generated some 2D garment masks, some of which have extreme long sleeves, extreme short sleeves, extreme short vertical length and extreme long vertical length. There are totally 850 mask images generated,

varying evenly from short to long vertical length, short to long sleeve length and slim to fat. Similarly, the dataset is split into training and testing dataset randomly, where 80% of them are used for training and 20% of them are used for testing.

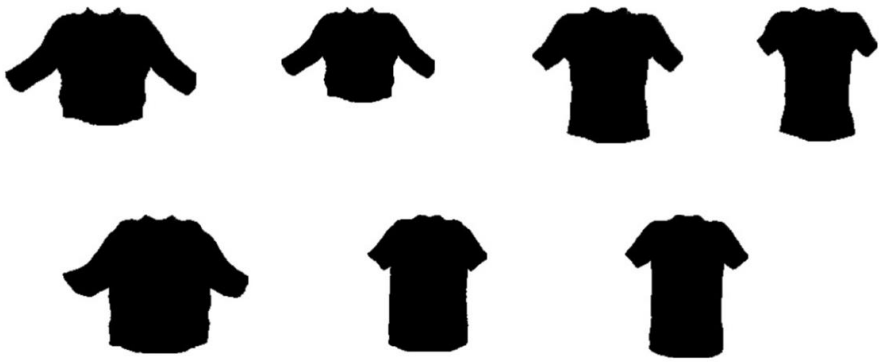


Figure 3.7: 2D garment masks generated with various and extreme garment styles

3.3.2 Implementation Details

As described above, our model is consisting of three layers of convolutional and maxpooling layers followed by an MLP with two hidden layers. Unsupervised learning is implemented in the training process. First the input images are relocated to the centre and followed by being reshaped to 256 x 256. Other reshaping sizes also can be applied depending on your own decision and requirement of training efficiency. The pairs of the setting of input and output channels of the three convolutional layers are (3,6), (6,12) and (12,12) with kernel size of 3. The size of maxpooling layer is (2,2). The input size of the MLP layers is 12288 and those of the two hidden layers are 1200 and 100. The output size is 4. The input images are shuffled before feeding to the model to smooth the training process. The optimizer used is the Adam with standard settings except the learning rate. During training, the learning rate is gradually reduced every two epochs to make the training process easily converge.

3.3.3 Experimental Results

In this section, we will show the testing results with various inputs, including the normal 3D garment RGB images as well as the garment masks with extreme garment styles. Then, we combine the part 1 3D garment mesh reconstruction and part 2 3D garment animation to show the overall output.

Test with 2D RGB garment images Some testing results are shown in Figure 3.8 with RGB garment images as input. From Figure 3.8, we can see that the network can work properly with different styles of garments. The vertical length of the generated 3D garment mesh is adjusted properly according to the input image. We can see obvious increase in vertical length of the generated 3D garment meshes with increasing of the vertical length in input images. Similarly, the output varies its width properly according to the width of input image as well. But for real 2D garment images, it is difficult to find some extreme examples, such as very long or short sleeves, very short vertical lengths and so on. So we have generated some garment masks with extreme designs as well as normal designs to test the model's adaptation ability which will be shown in the next subsection.

Figure 3.8: Overview of the test outputs of the network. The upper row is the input 2D RGB garment images which are split out from the dataset before training. The lower row is the corresponding projected 3D garment mesh vertices from the generated 3D garment meshes during testing.

In addition to the qualitative results, we have collected the quantitative results as well by measuring the different parts of the projected 3D garment mesh vertices from the generated 3D garment meshes as well as those parts of the input images to do comparison. The measurements are collected in pixel length. Three parts are measured, namely the sleeve length, vertical length and mid-width. The measurement method is shown in the following illustration:

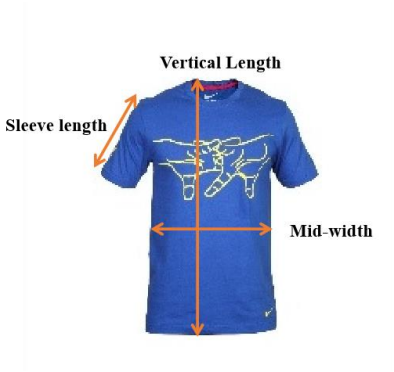


Figure 3.9: Illustration of the measurement method for garment

	Sleeve Length	Mid-width	Vertical Length
Variance	0.07%	0.22%	0.06%
Standard Deviation	2.72%	4.72%	2.41%
Average	4.76%	6.44%	4.20%
Median	4.49%	5.53%	4.07%
75th Percentile	6.52%	7.95%	5.88%
25th Percentile	2.44%	3.01%	2.26%

Table 3.1: Difference in percentage of the measurements of sleeve length, mid-width, and vertical length between input garment RGB images and the projected 3D garment mesh vertices of the outputs.

We have measured all 230 pairs of testing results. First, we measure the sleeve length, mid-width and vertical length for both testing input and output images separately. Then, we calculate the difference between them and convert the difference into percentage. After that, statistical quantities are calculated for analysis. The statistical data is summarized in the table above. From the data, it shows the average difference for sleeve length is 4.76%, while the median is 4.49%, which is close to the average. The 75^{th} and 25^{th} percentile are 6.52% and 2.44% respectively. The standard deviation is 2.72%, which shows that the variation is

not much. The data of vertical length is similar to those of sleeve length. The average difference of mid-width is slightly worse comparing to the other two, but it is still not bad with a value of 6.44%. The standard deviation also is slightly larger.

Test with generated 2D garment masks Examples of testing results with the generated garment masks as input are shown in Figure 3.10. From Figure 3.10, we can clearly see that the generated 3D garment meshes are quite well fitted to the input garment masks, even some of the garments have extreme designs. Example 1 provides a super long sleeve and short vertical length and example 3 has a super long sleeves and wide width. With these extreme cases, the network still can properly adjust the outputs to produce similar 3D garment models. The sleeves of the generated 3D garment models for example 1 and 3 is much longer than the rest. And from the slimmest example 4 to the widest example 3, the generated 3D garment models give an obvious increase in the width. With the increasing of vertical length from example 4 to example 6, the vertical length of the generated 3D garment models is significantly extended. From these results, it shows that our model is capable of dealing with various styles of garments as well as able to take garment RGB image or mask image as input.

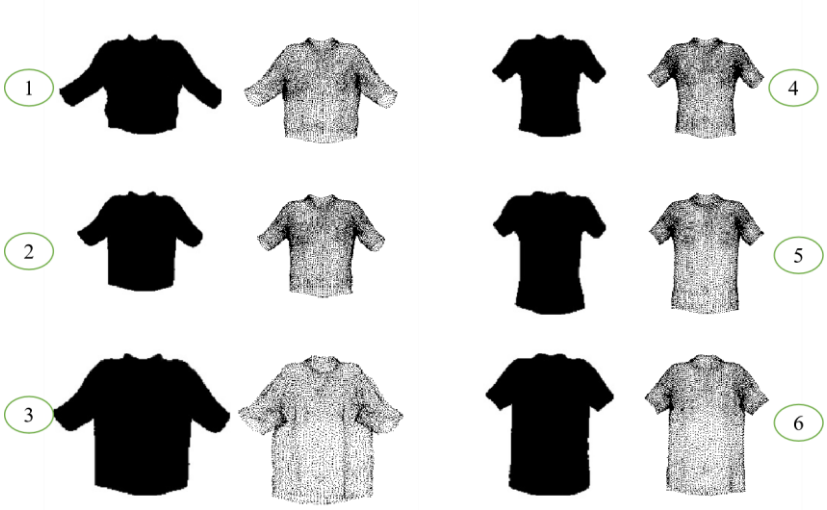


Figure 3.10: Overview of the test outputs of the network for the inputs of garment masks. For each pair, the left side is the input garment mask, and the right side is the corresponding test output of projected 3D garment mesh vertices from the generated 3D garment mesh.

	Sleeve Length	Mid-width	Vertical Length
Variance	0.15%	0.08%	0.03%
Standard Deviation	3.86%	2.89%	1.72%
Average	6.41%	4.61%	2.08%
Median	6.17%	4.44%	1.53%
75th Percentile	7.94%	6.71%	2.98%
25th Percentile	3.79%	2.11%	0.65%

Table 3.2: Difference in percentage of the measurements of sleeve length, mid-width, and vertical length between input garment mask images and the projected 3D garment mesh vertices of the outputs.

Similarly, we have measured all 170 pairs of testing results for the sleeve length, mid-width and vertical length for both testing input and output images separately. The difference between them is converted into percentage and the statistical quantities are summarized in Table 3.2. The average difference of sleeve length is higher than the other two, which may be due to the extreme long and short cases. But it is still reasonable with a value of 6.41%. And the 75th percentile is 7.94% that is not much different from the average. Data of mid-width shows better results comparing with the previous RGB images. Especially the vertical length gives the best result across all the testing results with an average difference of only 2.08%. The 75th percentile is only 2.98% that shows most of the cases with a difference under 3%, which is accurate.

Test with input images without aligning positions As mentioned above, we must align the position of the garment or its mask in the input image with the projected mesh vertices to avoid abnormality. Now let us explore what will happen without alignment and the same sets of garment and mask images in previous two sections are used for easy comparison.

The results are shown in Figure 3.11 and 3.12. Figure 3.11 is for RGB input images and Figure 3.12 is for mask input images. From the images we can see that the position of the garments and masks in input images have not been aligned with that of the projected mesh vertices. The position of some garments and masks are shifted to the right side, left side, upper part or lower part comparing to the position of the projected mesh vertices.

From the output, we can find that most of the projected mesh vertices images are abnormal. For example, the second input garment is smaller than the first one, but the output of second pair is larger than that of the first pair. Similarly, the outputs of the 3rd, 4th, 6th, 7th and 8th pairs become much smaller comparing with the results from the previous section. While the outputs of the 5th, 10th and 11th pairs become much larger than before. This is because the misalignment between the input garment or mask images and the projected mesh vertices misleads the comparison between the input images and the output image as well as the calculation of the loss.

Based on the analysis of the experiment results, it clearly indicates that we must align the positions first before both training and testing in order to get proper outputs from the network.

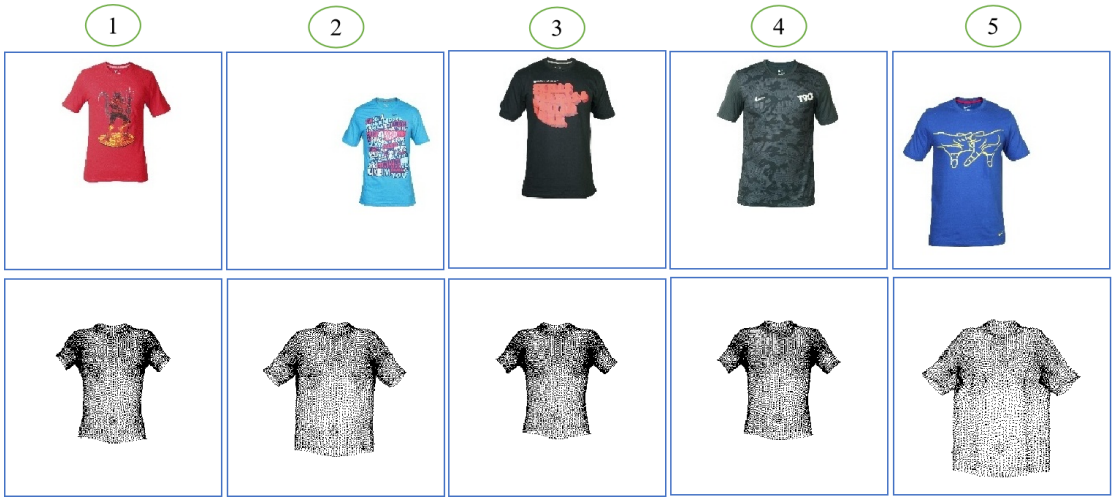


Figure 3.11: Overview of the outputs from the network for the RGB garment input images without alignment. The first row is the input RGB images and the second row is the outputs from network.

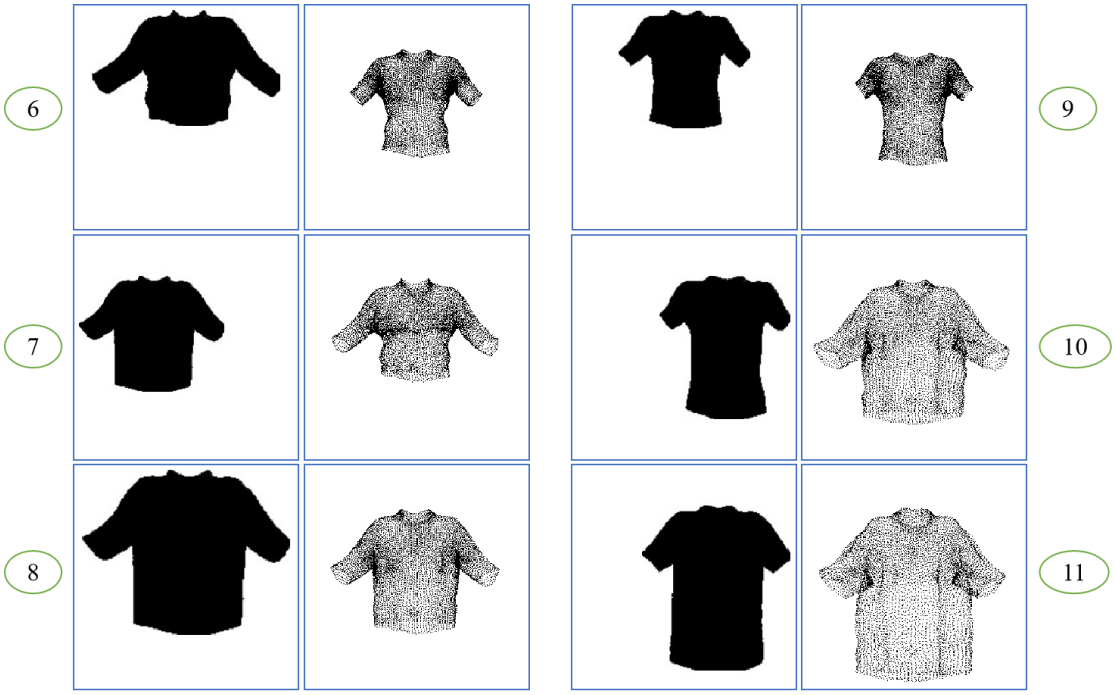


Figure 3.12: Overview of the outputs from the network for the mask input images without alignment. There are totally 6 pairs and for each pair, left side is the input mask images and the right side is the corresponding output from network

Overall output of the whole network Now let us combine the part 1 and part 2 of the network to take a look at the overall results. In Figure 3.13, we have selected some outputs as examples. Two real images and two generated garment masks are chosen. The selected garment images and masks have various styles: short vertical and long vertical length, short and long sleeves, slim and fat. For the first two real images, the second one has a larger size. The generated 3D garment mesh is larger as expected. When the generated 3D garment meshes are worn on 3D human models, it is also obvious that the second one is longer and looser than the first one. The fourth image has a much longer sleeve than others. Comparing the outputs of the fourth one with others, it clearly shows a longer sleeve than others for both the generated 3D garment mesh and the garment mesh worn on human body. What is more, the fourth one is the fattest. From the results of garment meshes draping on human body, they also properly show this difference. From the observations, the overall output of the whole network is good and as expected.

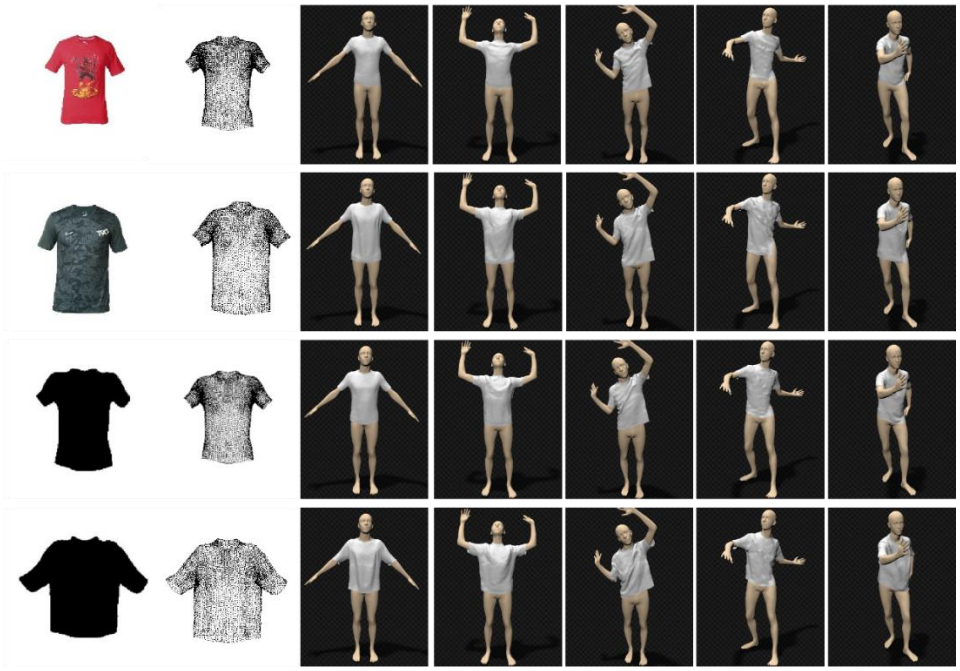


Figure 3.13: Overall test results of the whole network including 3D garment mesh generation and animation for some specific poses. The input of first two rows are real garment images while that of last two rows are generated garment masks. The first column is the input of the network, and the second column is the projected 3D garment mesh vertices. Column 3 to 7 are the animation results.

In Figure 3.14, we show a sequence of frames of animation results for selected examples. The first one selected is from actual images and the second one is from generated mask images. The first image has a longer vertical length than the second one while it has shorter sleeves. And the width of the first image is slimmer than the second one. From the animation results, we can clearly see these differences. The T-shirt mesh completely covers the hip for the first example while part of the hip is exposed for the second example. And the second example are obviously looser than the first one which is consistent with a larger width. The wrinkles predicted in the sequence of frames are quite realistic compared with the real ones in our daily life. The whole animation videos are available in the link Example 1 and Example 2 respectively.

Figure 3.14: Sequences of frames of animation results

Chapter 4

Conclusion and Future Work

4.1 Conclusion

This work has addressed two problems, namely unsupervised learning and 3D garment mesh generation. The output of the network is acceptable by garment animation models. The network allows any front-view garment image as input. It greatly contributes to the availability of datasets. The generated 3D garment meshes are properly adjusted according to different garment styles even the garment styles are extreme cases. They can be easily utilized by the garment animation network to be animated in 3D space. The training efficiency is high and it can finish the training in several hours with excellent training results. This work provides a significant step forward for the application of 3D garment animation which allows the network to start from real 2D garment images instead of 3D garment meshes. It makes the model easier to be used by people who do not have any knowledge about 3D modelling and users can simply run the model by providing a garment image as input.

4.2 Future Work

Two further works can be considered in the future:

Garment texture transfer In our work, we mainly focus on the 3D garment meshes' generation. Another important part is the garment texture transfer. It is a process that extracts texture from 2D images and then maps the extracted texture onto the 3D meshes. In the future, we can continue to work on mapping the garment texture onto the 3D garment meshes. Together with the wrinkles

predicted by the animation network, it will make the animated 3D garment models look more realistic.

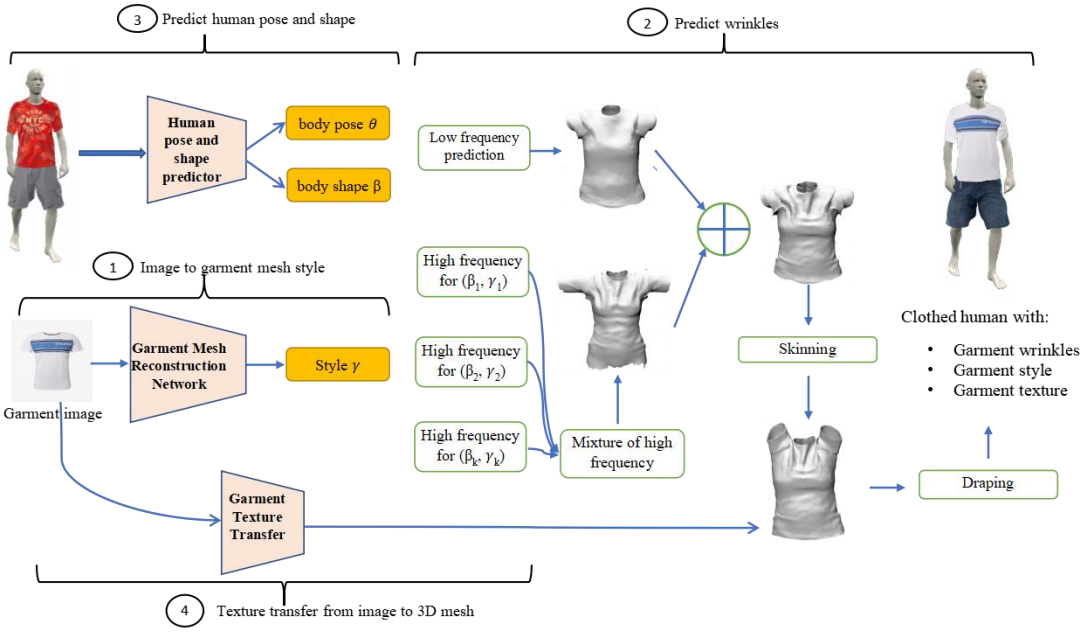


Figure 4.1: Possible network for the future work. Add two new parts 3 and 4. Part 3 is for predicting human pose and shape while part 4 is for texture transfer from image to 3D mesh.

Human pose and shape prediction In order to make the network be completely automatic, we can add the human pose and shape prediction part in the future. This will enable the network to predict the human pose and shape parameters directly from an image of human body. In the figure above, we have proposed a possible network with all of these parts which can be working on in the future. With all these parts, the network will be able to automatically predict all the necessary information for 3D garment animation from 2D input images. The output will be a clothed human with the garment texture and style same as the input garment image and the wrinkles predicted based on garment style, human pose and shape. This will be extremely useful in online shopping of clothes. It can immediately give consumers an overview of the dress simply with a photo of the garment and a selfie.

Reference

[1] A. Mir, T. Alldieck, and G. Pons-Moll. Learning to transfer texture from clothing images to 3D humans. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[2] C. Patel, Z. Liao, G. Pons-Moll. TailorNet: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[3] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2Shape: Detailed full human body geometry from a single image. In The IEEE International Conference on Computer Vision (ICCV), 2019.

[4] V. Lazova, E. Insafutdinov, and G. Pons-Moll. 360-degree textures of people in clothing from a single image. In International Conference on 3D Vision (3DV), 2019.

[5] A. Grigorev, A. Sevastopolsky, A. Vakhitov, and L. Victor. Coordinate-based texture inpainting for pose-guided image generation. In CVPR, 2019.

[6] T. Zhi, C. Lassner, T. Tung, C. Stoll, S. G Narasimhan, and M. Vo. TexMesh: Reconstructing detailed human texture and geometry from RBG-D video. In ECCV, 2020.

[7] N. Jin, Y. Zhu, Z. Geng, and R. Fedkiw. A pixel-based framework for data-driven clothing. arXiv preprint arXiv:1812.01677, 2018.

[8] Z. Su, T. Yu, Y. Wang, Y. Li, and Y. Liu. DeepCloth: Neural garment representation for shape and style editing. arXiv preprint arXiv:2011.14619, 2020.

[9] Y. Li, M. Habermann, B. Thomaszewski, S. Coros, T. Beeler, and C. Theobalt. Deep physics-aware inference of cloth deformation for monocular human performance capture. arXiv preprint arXiv:2011.12866, 2020.

[10] H. Bertiche, M. Madadi, and S. Escalera. DeePSD: Automatic deep skinning and pose space deformation for 3D garment animation. CoRR, abs/2009.02715, 2021.

[11] H. Bertiche, M. Madadi, and S. Escalera. PBNS: Physically Based Neural Simulator for Unsupervised Garment Pose Space Deformation. arXiv preprint arXiv:2012.11310, 2020.

[12] G. Pavlakos, N. Kolotouros, and K. Daniilidis. TexturePose: Supervising human mesh estimation with texture consistency. In ICCV, 2019.

[13] C. C. L. Wang, Y. Wang, and M. M. F. Yuen. Design automation for customized apparel products. Comput. Aided Des., 37(7):675–691, 2005. doi: 10.1016/j.cad.2004.08.007

[14] 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, 2006. doi: 10.1111/j.1467-8659.2006.00982.x

[15] C. Robson, R. Maharik, A. Sheffer, and N. Carr. Context-aware garment modeling from sketches. Comput. Graph., 35(3):604–613, 2011. doi: 10.1016/j.cag.2011.03.002

[16] 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, 2013. doi: 10.1145/2461912.2461975

[17] 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.

[18] R. White, K. Crane, and D. A. Forsyth. Capturing and animating occluded cloth. ACM Trans. Graph., 26(3), 2007. doi: 10.1145/1276377.1276420

[19] D. Bradley, T. Popa, A. Sheffer, W. Heidrich, and T. Boubekeur. Markerless garment capture. ACM Trans. Graph., 27(3):99:1–99:9, 2008. doi: 10.1145/1360612.1360698

[20] 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, 2017. doi: 10.1145/3072959.3073711

[21] X. Chen, B. Zhou, F. Lu, L. Wang, L. Bi, and P. Tan. Garment modelling with a depth camera. ACM Trans. Graph., 34(6):203:1–203:12, 2015. doi: 10.1145/2816795.2818059

[22] B. Zhou, X. Chen, Q. Fu, K. Guo, and P. Tan. Garment modelling from a single image. Computer Graphics Forum, 32(7):85–91, 2013. doi: 10.1111/cgf.12215

[23] 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, 2015. doi: 10.1002/cav.1653

[24] D. Terzopoulos, J. Platt, A. Barr, and K. Fleischer. Elastically deformable models. ACM Siggraph Computer Graphics, 21(4):205–214, 1987.

[25] A. Selle, J. Su, G. Irving, and R. Fedkiw. Robust high-resolution cloth using parallelism, history-based collisions, and accurate friction. IEEE Transactions on Visualization and Computer Graphics, 15(2):339–350, 2009.

[26] C. Jiang, T. Gast, and J. Teran. Anisotropic elastoplasticity for cloth, knit and hair frictional contact. ACM Transactions on Graphics (TOG), 36(4):152, 2017.

[27] R. Gillette, C. Peters, N. Vining, E. Edwards, and A. Sheffer. Real-time dynamic wrinkling of coarse animated cloth. In Proc. Symposium on Computer Animation, 2015.

[28] L. Kavan, D. Gerszewski, A. W. Bargteil, and P.-P. Sloan. Physics-inspired upsampling for cloth simulation in games. ACM Trans. Graph., 30(4):93:1–93:10, 2011.

[29] D. Kim, W. Koh, R. Narain, K. Fatahalian, A. Treuille, and J. F. O'Brien. Near-exhaustive precomputation of secondary cloth effects. ACM Transactions on Graphics, 32(4):87:1–7, 2013. Proceedings of ACM SIGGRAPH 2013, Anaheim.

[30] H. Wang, F. Hecht, R. Ramamoorthi, and J. F. O'Brien. Example-based wrinkle synthesis for clothing animation. ACM Transactions on Graphics, 29(4):107:1–8, 2010. Proceedings of ACM SIGGRAPH 2010, Los Angles, CA.

[31] M. Muller, B. Heidelberger, M. Hennix, and J. Ratcliff. Position based dynamics. Journal of Visual Communication and Image Representation, 18(2):109–118, 2007.

[32] M. Muller. Hierarchical position based dynamics. 2008.

[33] X. Provot. Deformation constraints in a mass-spring model to describe rigid cloth behaviour. In Graphics interface, pages 147–147. Canadian Information Processing Society, 1995.

[34] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multi-Garment Net: Learning to dress 3D people from images. In IEEE International Conference on Computer Vision (ICCV). IEEE, 2019.

[35] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[36] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3D people models. In IEEE Conf. on Computer Vision and Pattern Recognition, 2018.

[37] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt. LiveCap: Real-time human performance capture from monocular video. Transactions on Graphics (ToG), 2019.

[38] A. Neophytou and A. Hilton. A layered model of human body and garment deformation. In 2014 2nd International Conference on 3D Vision, volume 1, pages 171–178. IEEE, 2014.

[39] Y. Tao, Z. Zheng, Y. Zhong, J. Zhao, D. Quionhai, G. Pons-Moll, and Y. Liu. SimulCap: Single-view human performance capture with cloth simulation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[40] Y. Tao, Z. Zheng, K. Guo, J. Zhao, D. Quionhai, H. Li, G. Pons-Moll, and Y. Liu. DoubleFusion: Real-time capture of human performance with inner body shape from a depth sensor. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2018.

[41] J. Yang, J.-S. Franco, F. H. Wheeler, and S. Wuhrer. Analyzing clothing layer deformation statistics of 3D human motions. In European Conf. on Computer Vision, pages 237–253, 2018.

[42] Z. Lahner, D. Cremers, and T. Tung. DeepWrinkles: Accurate and realistic clothing modeling. In Proceedings of the European Conference on Computer Vision (ECCV), pages 667–684, 2018.

[43] Q. Ma, J. Yang, A. Ranjan, S. Pujades, G. Pons-Moll, S. Tang, and M. Black. Learning to dress 3D people in generative clothing. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[44] E. d. Aguiar, L. Sigal, A. Treuille, and J. K. Hodgins. Stable spaces for real-time clothing. ACM Trans. Graph., 29(4):106:1–106:9, 2010.

[45] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black. DRAPE: dressing any person. ACM Trans. Graph., 31(4):35:1–35:10, 2012.

[46] I. Santesteban, M. A. Otaduy, and D. Casas. Learning-based animation of clothing for virtual try-on. Comput. Graph. Forum, 38(2):355–366, 2019.

[47] E. Gundogdu, V. Constantin, A. Seifoddini, M. Dang, M. Salzmann, and P. Fua. GarNet: A two-stream network for fast and accurate 3D cloth draping. CoRR, abs/1811.10983, 2018.

[48] D. Casas, M. Volino, J. Collomosse, and A. Hilton. 4D video textures for interactive character appearance. Computer Graphics Forum, 33(2):371–380, 2014.

[49] R. Goldenthal, D. Harmon, R. Fattal, M. Bercovier, and E. Grinspun. Efficient simulation of inextensible cloth. ACM Transactions on Graphics (Proceedings of SIGGRAPH 2007), 26(3):to appear, 2007.

[50] Y. Shen, J. Liang, and M. C. Lin. GAN-based Garment Generation Using Sewing Pattern Images. ECCV, 2020.

[51] P. Huang, J. Yao, H. Zhao: Automatic realistic 3D garment generation based on two images. International Conference on Virtual Reality and Visualization (ICVRV), 2016.

[52] C. B Choy, D. Xu, J.Y. Gwak, K. Chen, and S. 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.

[53] H. Fan, H. Su, and L. 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.

[54] T. Groueix, M. Fisher, V. G. Kim, B. Russell, and M. 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.

[55] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3D reconstruction in function space. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[56] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang. Pixel2Mesh: Generating 3D mesh models from single RGB images. In Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[57] H. Zhu, Y. Cao, H. Jin, W. Chen, D. Du, Z. Wang, S. Cui, and X. 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.

[58] B. Jiang, J. Zhang, Y. Hong, J. Luo, L. Liu, and H. Bao. BCNet: Learning body and cloth shape from a single image. In Proceedings of the European Conference on Computer Vision (ECCV), 2020.

[59] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J Black. SMPL: A skinned multi-person linear model. ACM transactions on graphics (TOG), 2015.

[60] E. Turquin, M. Cani, J. Hughes: Sketching garments for virtual characters. In: 34. International Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2007, San Diego, California, USA, August 5-9, 2007, Courses. p. 28, 2007, https://doi.org/10.1145/1281500.1281539

[61] A. Jung, S. Hahmann, D. Rohmer, A. Begault, L. Boissieux, M. Cani: Sketching folds: Developable surfaces from non-planar silhouettes. ACM Trans. Graph. 34(5), 155:1–155:12, 2015, https://doi.org/10.1145/2749458

[62] J. Bednarik, P. Fua, and M. Salzmann. Learning to reconstruct texture-less deformable surfaces from a single view. In International Conf. on 3D Vision, pages 606–615, 2018.

[63] R. Danecek, E. Dibra, A. C. Oztireli, R. Ziegler, and M. Gross. DeepGarment: 3D garment shape estimation from a single image. In Computer Graphics Forum, volume 36, pages 269–280. Wiley Online Library, 2017.

[64] K. Park, K. Rematas, A. Farhadi, and S. M. Seitz. PhotoShape: Photorealistic materials for large-scale shape collections. ACM Trans. Graph., 37(6), 2018.

[65] D. Baraff and A. Witkin. Large steps in cloth simulation. In Proceedings of the 25th annual conference on Computer graphics and interactive techniques, pages 43–54. ACM, 1998.

[66] C. Lassner, G. Pons-Moll, and P. V. Gehler. A generative model of people in clothing. In Proceedings IEEE International Conference on Computer Vision (ICCV), Piscataway, NJ, USA, 2017.

[67] H. N Ng and R. L Grimsdale. Computer graphics techniques for modeling cloth. IEEE Computer Graphics and Applications, 16(5):28–41, 1996.

[68] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin. Soft-gated warping-GAN for pose-guided person image synthesis. In Advances in Neural Information Processing Systems, pages 474–484, 2018.

[69] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis & Machine Intelligence, (4):509–522, 2002.

[70] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rossl, and H-P Seidel. Laplacian surface editing. In Eurographics/ACM SIGGRAPH symposium on Geometry processing, pages 175–184. ACM, 2004.

[71] H. Dong, X. Liang, X. Shen, B. Wu, B.-C. Chen, and J. Yin. FW-GAN: Flow-navigated warping GAN for video virtual try-on. In The IEEE International Conference on Computer Vision (ICCV), 2019.

[72] S. Yang, Z. Pan, T. Amert, K. Wang, L. Yu, T. Berg, and M. C Lin. Physics-inspired garment recovery from a single-view image. ACM Transactions on Graphics (TOG), 37(5):170, 2018.

[73] B. Zhao, X. Wu, Z.-Q. Cheng, H. Liu, Z. Jie, and J. Feng. Multi-view image generation from a single view. In 2018 ACM Multimedia Conference on Multimedia Conference, pages 383–391. ACM, 2018.

[74] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. V. Gool. Pose guided person image generation. In Advances in Neural Information Processing Systems, pages 406–416, 2017.

[75] L. Ma, Q. Sun, S. Georgoulis, L. V. Gool, B. Schiele, and M. Fritz. Disentangled person image generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 99–108, 2018.

[76] A. Pumarola, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[77] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe. Deformable GANs for pose-based human image generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3408–3416, 2018.

[78] G. Balakrishnan, A. Zhao, A. V Dalca, F. Durand, and J. Guttag. Synthesizing images of humans in unseen poses. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8340–8348, 2018.

[79] S. Song, W. Zhang, J. Liu, and T. Mei. Unsupervised person image generation with semantic parsing transformation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2357–2366, 2019.

[80] A. Raj, P. Sangkloy, H. Chang, J. Hays, D. Ceylan, and J. Lu. SwapNet: Image based garment transfer. In European Conference on Computer Vision, pages 679–695. Springer, 2018.

[81] M. Zanfir, A.-I. Popa, A. Zanfir, and C. Sminchisescu. Human appearance transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5391–5399, 2018.

[82] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S Davis. VITON: An image-based virtual try-on network. In CVPR, 2018.

[83] B. Wang, H. Zheng, X. Liang, Y. Chen, and L. Lin. Toward characteristic-preserving image-based virtual try-on network. In Proceedings of the European Conference on Computer Vision (ECCV), pages 589–604, 2018.

[84] R. Yu, X. Wang, and X. Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In The IEEE International Conference on Computer Vision (ICCV), 2019.

[85] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, Z. Hu, and J. Yin. Towards multi-pose guided virtual try-on network. In The IEEE International Conference on Computer Vision (ICCV), 2019.

[86] N. Zheng, X. Song, Z. Chen, L. Hu, D. Cao, and L. Nie. Virtually trying on new clothing with arbitrary poses. In Proceedings of the 27th ACM International Conference on Multimedia, pages 266–274. ACM, 2019.

[87] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In IEEE Conf. on Computer Vision and Pattern Recognition, 2017.

[88] V. Gabeur, J.-S. Franco, X. Martin, C. Schmid, and G. Rogez. Moulding humans: Non-parametric 3D human shape estimation from single images. CoRR, abs/1908.00439, 2019.

[89] L. Kavan, S. Collins, J. Zara, C. O'Sullivan: Geometric skinning with approximate dual quaternion blending. ACM Transactions on Graphics (TOG) 27(4), 1–23, 2008.

[90] L. Kavan, J. Zara: Spherical blend skinning: a real-time deformation of articulated models. In: Proceedings of the 2005 symposium on Interactive 3D graphics and games. pp. 9–16, 2005.

[91] B.H. Le, Z. Deng: Smooth skinning decomposition with rigid bones. ACM Transactions on Graphics (TOG) 31(6), 1–10, 2012.

[92] N. Magnenat-thalmann, R. Laperrire, D. Thalmann, U.D. Montr´eal: Joint-dependent local deformations for hand animation and object grasping. In: In Proceedings on Graphics interface '88. pp. 26–33, 1988.

[93] B. Allen, B. Curless, Z. Popovic: Articulated body deformation from range scan data. ACM Transactions on Graphics (TOG) 21(3), 612–619, 2002.

[94] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, J. Davis: Scape: shape completion and animation of people. In: ACM SIGGRAPH 2005 Papers, pp. 408–416, 2005.

[95] J.P. Lewis, M. Cordner, N. Fong: Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. pp. 165–172, 2000.

[96] P. Zhang, B. Zhang, D. Chen, L. Yuan, F. Wen: Cross-domain correspondence learning for exemplar-based image translation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5143-5153, 2020.

[97] F. Zhan, Y. Yu, K. Cui, G. Zhang, S. Lu, J. Pan, C. Zhang, F. Ma, X. Xie, C. Miao: Unbalanced feature transport for exemplar-based image translation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15028-15038, 2021.