

Robust 3D Garment Digitization from Monocular 2D Images for 3D Virtual Try-On Systems

Sahib Majithia
Myntra Designs Pvt. Ltd.
Bangalore, India

sahib.majithia@myntra.com

Sandeep N. Parameswaran
Myntra Designs Pvt. Ltd.
Bangalore, India

sandeep.narayan@myntra.com

Sadbhavana Babar
Myntra Designs Pvt. Ltd.
Bangalore, India

sadbhavana.babar@myntra.com

Vikram Garg
Myntra Designs Pvt. Ltd.
Bangalore, India

vikram.garg@myntra.com

Astitva Srivastava
IIIT-Hyderabad,
India

astitva.srivastava@research.iiit.ac.in

Avinash Sharma
IIIT-Hyderabad,
India

asharma@iiit.ac.in



Figure 1: Textured template meshes of T-shirts & Trousers draped on a human avatar.

Abstract

In this paper, we develop a robust 3D garment digitization solution that can generalize well on real-world fashion catalog images with cloth texture occlusions and large body pose variations. We assumed fixed topology parametric template mesh models for known types of garments (e.g., T-shirts, Trousers) and perform mapping of high-quality texture from an input catalog image to UV map panels corresponding to the parametric mesh model of the garment. We achieve this by first predicting a sparse set of 2D landmarks on the boundary of the garments. Subsequently, we use these landmarks to perform Thin-Plate-Spline-based texture transfer on UV map panels. Subsequently, we employ a deep texture inpainting network to fill the large holes (due

to view variations & self-occlusions) in TPS output to generate consistent UV maps. Furthermore, to train the supervised deep networks for landmark prediction & texture inpainting tasks, we generated a large set of synthetic data with varying texture and lighting imaged from various views with the human present in a wide variety of poses. Additionally, we manually annotated a small set of fashion catalog images crawled from online fashion e-commerce platforms to finetune. We conduct thorough empirical evaluations and show impressive qualitative results of our proposed 3D garment texture solution on fashion catalog images. Such 3D garment digitization helps us solve the challenging task of enabling 3D Virtual Try-on.

1. Introduction

The fashion e-commerce market is a rapidly growing industry, especially in the post-pandemic era. It is expected to reach 670 billion dollars in 2021 [1]. One key challenge faced by this growing industry is the high return rate of around 72% [2]. A large part of these returns are due to customer apprehensions regarding garment ‘look’ and ‘fit’ [3, 4]. A realistic Virtual Try-on (VT) system would help alleviate the high return rate as it will help the customers visualize combinations of different apparel in different sizes on an avatar of their own body shapes. However, existing image-based 2D VT solutions [5, 6, 7, 8, 9, 10, 11, 12, 13, 14] have inherent limitations in providing a satisfactory user experience as they lack the ability to generate novel poses and view variations for the customer. Majority of 2D VT solutions, which can deal with the variations, use generative models [15, 16, 17, 18] and thus fails to reconstruct images with high-frequency texture details (visual content on clothing) apart from regularly failing on fidelity of exact cloth design features (e.g., neck or collar type). On the other hand, 3D VT solutions use 3D models for clothing and then render a draped version of the same on 3D human avatars (e.g., physics-based cloth deformation by Blender [19]) of varying body shapes. An important requirement for a 3D VT system is the availability of a large catalog of digitized 3D garment/cloth models, which requires an expensive multi-view capture/scan setup thus can’t scale up for large catalog sizes, especially in fast-fashion scenarios (owing to rapidly changing fashion trends [20, 21]).

Recent advancement in deep learning has enabled attempting this challenging task in generative frameworks. In this line of work, a few recent efforts in the literature [22, 23] attempt to automate 3D digitization of clothing from monocular/single catalog image by learning to transfer texture from catalog images to respective 2D texture maps. In particular, [22] learns a mapping from the input image pixels to the UV map, in the form of dense correspondences from garment image silhouettes pixels to the pixels present in the 2D UV map of a 3D garment surface. Nevertheless, such mapping is global in nature and does not guarantee to preserve high-frequency textures and tends to blur out fine-grained details like thin lines and curvatures in the cloth texture (as shown in their paper). Additionally, [22] uses pre-segmented ghost mannequin cloth images in the canonical (frontal) pose as target cloth image, whereas a large part of available real catalog images involves actual human models wearing clothing with hands in the different body poses. Another relevant work [23] attempted to retain high-frequency texture details using Thin-Plate-Spline (TPS) [24] based texture transfer on cloth UV map panels of template mesh. However, their method can not handle holes and missing parts in texture maps owing to self-occlusions

caused by articulated body poses and camera view variations. Their method also requires large-scale annotations of landmarks and segmentation on real data, which is expensive and time-consuming.

In this paper, we aim to overcome the limitations of existing methods to develop a robust 3D garment digitization solution that can generalize well on real-world catalog images. Such digitization will facilitate a 3D VT system to render a 360-degree view of garments with high-quality textures on realistic custom avatars. Our method assumes known types of garments (e.g., T-shirts and Trousers) and uses parametric templates with fixed topology as garment mesh for 3D representation. Thus, the critical task in this setup is mapping high-quality texture from a 2D catalog image to UV map panels of the parametric garment mesh. We achieve this by first predicting a sparse set of 2D landmarks on the boundary of the garment and segmentation map of the garment on the 2D image using JPPNet [25] architecture trained and finetuned on our synthetic and real dataset. Subsequently, these landmarks are used to perform TPS [24] based panel-wise texture transfer on the UV map. We further propose to employ a deep texture inpainting network (MADFNet [26]) to fill in the large holes / missing parts in the UV maps due to view variations and self-occlusions. Furthermore, a large amount of annotated data is required to train the supervised deep networks for landmark prediction and texture inpainting tasks. However, the majority of the existing fashion data sets (e.g., Deepfashion2 [27]) in this domain are focused on in-the-wild settings instead of the fashion catalog kind of setup. We empirically found that this caused degradation in performance for model trained on such datasets and tested on fashion catalog images. Thus, we generated a large set of synthetic data by varying texture and lighting (~3000 textures and 6 lighting variations) consisting of mainly T-shirts and trousers garments draped independently on virtual human avatars (~ 7000 human poses) imaged from various views. Additionally, we also manually annotated a small set (~ 1300) of fashion catalog images crawled from online fashion e-commerce platforms and finetuned and test deep networks trained on synthetic data. We intend to share this synthetic data with the larger academic community. We conducted thorough empirical evaluations and comparisons with SOTA methods and show impressive qualitative results of our proposed 3D garment texture solution on fashion catalog images.

2. Literature Review

2.1. 2D Techniques

Jetchev and Bergmann [28] proposed CAGAN, which first introduced the task of swapping fashion articles on human images. VITON [5] addressed the same problem by proposing a coarse-to-fine synthesis framework that in-

volves TPS transformation of clothes. Most existing virtual try-on methods tackle different aspects of VITON to synthesize perceptually convincing photo-realistic images. CP-VTON [29] adopted a geometric matching module to learn the parameters of TPS transformation, which improves the accuracy of deformation. VTNFP [30] and ACGPN [31] predicted the human-parsing maps of a person wearing the target clothes in advance to guide the try-on image synthesis. CP-VTON+ [13] improved upon CP-VTON by introducing a concrete loss function at the blending stage to preserve textures, but the authors of CP-VTON+ themselves pointed out that the 2D image-based approach has inherent limitations for coping with diversely posed target human cases. Therefore, the application is limited to simple clothing and standard posed target humans. For diverse cases, 3D reconstruction is more suitable.

2.2. 3D Techniques

3D human body reconstruction methods, which involve estimating a 3D clothed human body from a monocular RGB image, can be divided into two categories: Parametric and Non-parametric methods. Parametric methods [32, 33, 34] predict deformation from the SMPL [35] surface to model clothing, but they fail to model high-frequency textures in the clothing. [36] can retain high-frequency texture details but it can not deal with arbitrary topology of loose garments, such as skirts, long dresses etc. Non-parametric methods [37, 38, 39, 40, 41] can generate highly detailed clothed human mesh, but they also tend to produce blurry textures. Moreover, since these methods do not model clothes separately, extracting cloth part from the predicted mesh is not trivial. Additionally, the estimated geometry of the same cloth varies drastically with the pose and shape of the body.

Methods such as [42, 43, 44] predict how a synthetic 3D garment dynamically deforms over a given target human body pose and shape. These methods can deal with variations in shape, pose, and style, but they don't emphasize how to accurately transfer textures from real garment images onto the 3D synthetic garment.

Multiple works exist in the literature for creating 3D garment models. Conventional methods of generating 3D garment models use depth cameras [45, 46] or multi-view images [47] which require huge efforts for data capture. [48] can recover the underlying 3D mesh of the garment but it doesn't retain texture details, and moreover, it relies on initial human shape and pose estimation which can be unreliable. [49] can also recover 3D mesh which can be re-targeted to other body shapes, but it requires a large database of templates for the human body and garments to estimate parameters for template registration. MGN [50] and DeepFashion3D [51] attempt to reconstruct garment clothing by using real-world 3D scans. [52] uses sewing

patterns that are edited according to the SMPL body shape given as input. The edited sewing patterns are used to generate 3D garments to be draped on a specific body type. Although this method is robust and requires minimal input, it does not deal with the texture present in the clothing. [22] allows the transfer of textures to garments by learning to predict dense correspondences using a deep neural network. But, this method also requires images without occlusions and is not generic to every catalog process and does not produce textures at high resolutions.

CLOTH3D [53] uses synthetically generated data for the reconstruction of textured 3D meshes. But generalization of models trained on CLOTH3D to real-world images would require 3D scans. DeepGarment [54] claims to generalize on real-world samples by training on only synthetic data. But for their method, authors generate 100,000 samples of T-shirts and infer on images of 64x64 resolution. Also, they do not reconstruct textures along with 3D garment shape reconstruction.

JF-Net [23] automatically transfers textures of clothing images (front and back) to 3D garments of a fixed template by learning a mapping from images to the UV map of the garment. Although, [23] attempts to generate textures of 3D garments from RGB images, it restricts itself to images where the garment is completely visible. Pix2Surf [22] formulated their solution by using a silhouette mask of the garment. A general e-commerce catalog consists of various brands, which have their specific process for photo-shoot. This generates variations in human poses and occlusions of the garment, which [23] cannot handle. For [23] to work, a change in the catalog process is necessary.

3. Proposed 3D Garment Digitization Solution

Here, we introduce our novel solution for 3D garment digitization from generic e-commerce catalog images having body pose variations causing self-occlusions to garments. We assumed fixed topology parametric template mesh models for known types of garments (e.g., T-shirts) and perform mapping of high-quality texture from an input catalog image to UV map panels corresponding to the parametric mesh model of the garment. The architecture diagram is as shown in Figure 2. We achieve this by first predicting a sparse set of 2D landmarks/key points on the boundary of garments in the image. We adapt the existing JPPNet [25] architecture, which was originally proposed for human body joint prediction, for garment landmark prediction. Subsequently, these landmarks are used to perform TPS [24] based texture transfer on UV map panels. However, due to view variations and self-occlusions (as discussed before), these UV maps have large holes/missing parts. Additionally, clothing semantic segmentation maps predicted by JPPNet are used to remove pixels within the landmark boundary that are labeled as background or fash-

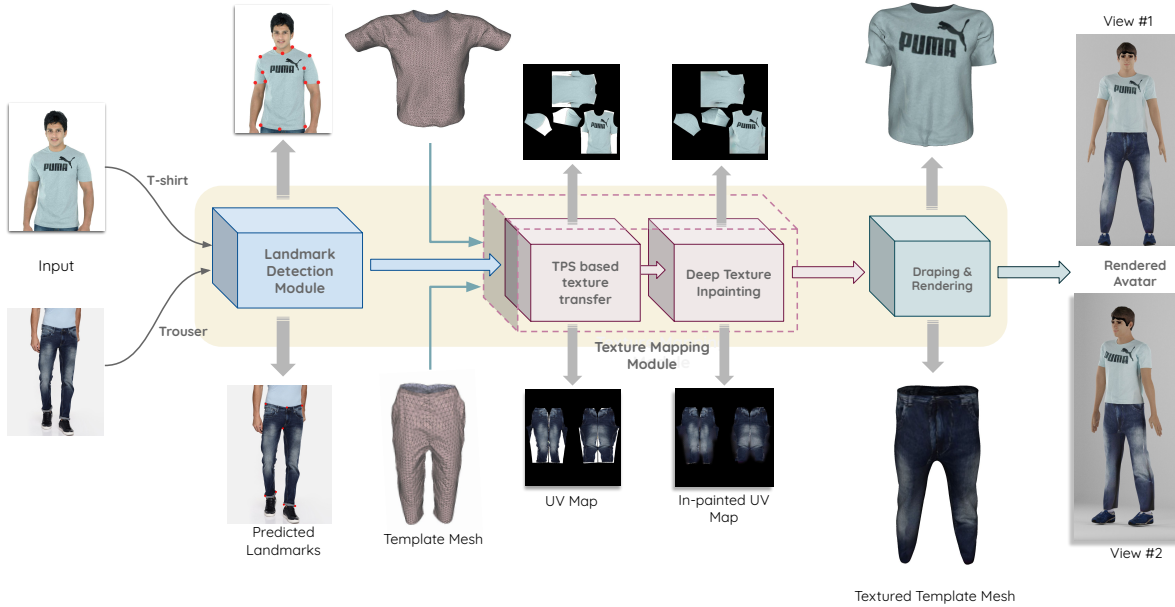


Figure 2: The proposed 3D garment digitization pipeline consists of 1) Landmark Detection Module which predicts a set of 2D landmarks and associated garment segmentation map, 2) Texture Mapping Module uses the predicted landmarks to transfer the texture onto the UV map of the template mesh using Thin-plate-Splines followed by inpainting on the self-occluded regions of the UV map using the deep network and 3) Draping & Rendering Module drapes the template mesh along with inpainted UV map on the target 3D human avatar. (Here we show results on T-shirts and Trousers simultaneously, but the whole process is carried out for each garment type separately. In the end, both garments are draped together.)

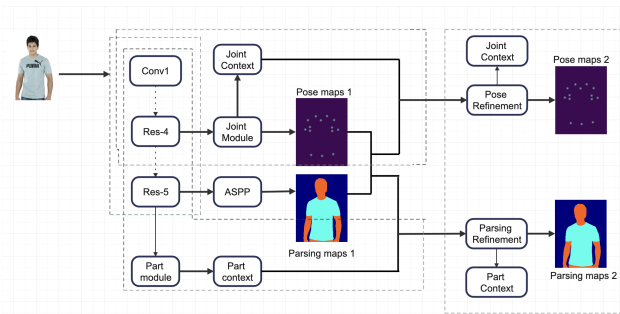


Figure 3: The architecture of JPPNet [25] adapted in our pipeline

ion articles occluding the clothing. Hence, we propose to employ a deep texture inpainting network (MADFNet [26]) to fill the large holes / missing parts to generate consistent UV maps. This network takes as input each panel in the UV along with the occluded regions and inpaints relevant texture using the information available from the non-occluded parts of the image.

3.1. Landmark Detection Module

We adapt the original JPPNet [25] architecture, as shown in Figure 3, to jointly predict landmarks and semantic maps. The landmarks for each garment are defined following JF-Net [23]. The number of landmarks varies with garment type (e.g. 14 for the T-shirts and 8 for the Trousers). Their

original JPPNet architecture was proposed for the human parsing and pose estimation, we adopt it for our task of clothing landmark prediction. We use clothing segmentation similar to the original paper. The architecture uses a common encoder and separate branches for both tasks, features of which are concatenated in the refinement stage to utilize the coherent representation of cloth parsing and landmark prediction. Similarly, in our multi-task learning setup, the parse maps produced by the network guide the predictions of landmarks as both tasks are spatially aligned. In addition to this, the network uses an iterative refinement scheme to aid individual tasks. Although we use both tasks for further stages, the texture mapping module is influenced more by pixel-level alignment of the landmarks in comparison to the segmentation maps. Thus, we focus our attention and evaluations on the landmark prediction task.

3.2. Texture Mapping Module

This module takes predicted landmarks as input and maps the relevant regions of the 2D image onto the UV map of the template garment mesh for texture transfer. Particularly, we use the inferred landmarks as the control points for the Thin-Plate-Spline [24] transformation. Each landmark in the 2D image is also mapped to a fixed point in UV map of the template garment mesh, which is mapped manually, once for each template based on the structure of panels that compose the template. The region of the 2D

image to be used for texture transfer is masked using the segmentation maps provided by the Landmark Detection Module. In comparison to the generative deep learning approaches that produce blurry textures, TPS enables us to transfer high-frequency texture details and provides a pixel-level accurate mask required for texture inpainting. The output of TPS yields partially filled UV maps with large missing parts/holes owing to self-occlusion and view-variations. In order to recover consistent UV maps, we employ automated texture inpainting network MADFNet [26]. It is an encoder-decoder-based architecture that comprises a Mask-Aware Dynamic Filtering (MADF) Module that aids in filling arbitrary missing regions. This network was originally proposed for in-the-wild images and trained on the Places dataset [55]. We propose to finetune the model (separately for each garment type) for the UV maps of the synthetically generated data. The training and test data for the inpainting network consist of multiple panels for each type of clothing.

Generation of back panels of T-shirts is severely impeded as the back camera is missing. Thus, we can either take into account the front panel and copy it to the back or take a uniform gradient patch from the front panel and perform texture copying to the back panel by replicating that patch on the entire back panel. However, if the back view image is readily available (as a part of the catalog), we can generate the back panel by using landmarks predicted on the back view of the T-shirt. In our experiments, we used the first strategy for filling back panel texture.

3.3. Draping and Rendering Module

Given the fixed template for the T-shirt and trousers alongside a human avatar, each garment is aligned with the avatar in T-pose independently. A cloth deformation is obtained for the garments by applying a cloth simulation [53, 42] to the template garment using Blender [19]. We apply a motion sequence to animate the human avatar and deform the garment corresponding to a target pose. The motion sequence has three keyframes: the first one where the avatar is in T-pose, an intermediate keyframe where the avatar remains in T-pose, and a final keyframe where the avatar is reposed to the target pose (we use an A-pose for the final 3D VT) using Linear Blend Skinning (LBS). A simple LBS skinning was used since the poses involved in the e-commerce cataloging process were largely devoid of extreme poses, and hence adequate realism was maintained. The human pose is kept constant for the initial set of frames to allow the cloth deformation to stabilize. The reposing from T-pose to the target pose is also done in multiple frames for enabling the cloth simulation to walk over multiple intermediate steps to obtain a realistic garment deformation. We individually apply the deformations to each template and perform a collision resolution to visualize both garments together on a human avatar.

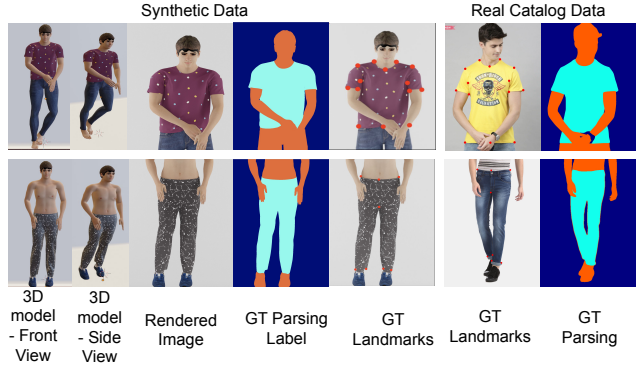


Figure 4: **Synthetic Data** and **Real Catalog Data** samples along with GT annotations.



Figure 5: Sample images from our synthetic dataset highlighting diverse textures and poses for (a) T-shirts and (b) Trousers

4. Experiments & Results

4.1. Our Synthetic Dataset

We generated synthetic data and rendered the images in a way that resembles an e-commerce catalog. We have taken template 3D garment meshes from the Berkley Garment Library [56] and human avatar meshes created using MakeHuman [57] that are rigged with a CMU MB armature. We utilized a T-shirt template and a trousers template for our use case. We included variations in the textures applied to the template meshes, poses of the human avatar, lighting of the scenes and minor variations in template size and camera coordinates. Over 3000 textures in the form of images were crawled from the internet to be applied to the template meshes. We also crawled over 7000 e-commerce catalog full shot images (i.e images that have the full human body visible) to generate a wide range of human poses. We further estimated the 3D human pose of each of these crawled images using [58]. The estimated skeleton information from the 3D pose was converted to joint angles and written in MoCap format [59]. The human avatar was further re-targeted to the estimated 3D pose by generating the rotation and translation parameters for each bone present

in the armature of the rigged skeleton with respect to the saved skeleton present in the form of motion data. The target poses were restricted to a near frontal pose by limiting the rotation matrix of the “Hips” and “Head” bones present in the CMU MB armature. The garments were draped as described in Sec.3.3 with the target poses obtained from estimated poses from the full-shot images (see Figure 4). The synthetic images were rendered using Blender CYCLES in 1080×1440 resolution. We also included 6 lighting conditions for rendering the scene by placing the multiple lights at different positions in the scene, mimicking an e-commerce cataloging activity. Few samples from our synthetic dataset are shown in Figure 5.

Each generated image was paired with segmentation maps and landmark annotations. These annotations were obtained similarly as in [60]. The T-shirt template mesh was associated with 14 landmark vertices and the trousers template mesh with 8 vertices. The corresponding 2D landmark annotations of these vertices in the rendered images were obtained by projecting the vertices onto the image using camera parameters. The segmentation labels were also obtained by labeling the garment in different rendering passes.



Figure 6: Landmarks and segmentation maps predicted by the landmark detection module on real catalog images.

4.2. Our Real-World Catalog Dataset

We manually annotated the landmarks on real-world catalog images. We crawled 1300 images for each type of garment from the web, out of which 1000 images were used for training and 300 for validation. Since the manually annotated landmarks are subjective in nature and prone to hu-



Figure 7: **Texture mapping module:** (a) Input image, (b) Texture transfer through TPS for the front panel and sleeves, & (c) Inpainted panels.

man mistakes, we also estimated the errors produced by the manual annotators. For this purpose, 200 images were sampled from the data and annotated by 5 different annotators and the estimated error values are compared with the prediction errors of the landmark detection module. The clothing segmentation labels are generated using [25].

4.3. Training Protocol

We train separate networks for T-shirts and Trousers clothing. For the landmark detection task, we adapted the original JPPNet architecture and train the network with 384×384 resolution images in two settings, namely, **scratch** and **finetune** with varying sizes of training data. The former refers to training the model on our real dataset starting from random initial weights. The latter refers to training with synthetic data first and further finetuning of the network parameters using images from our Real-World Catalog Dataset.

In regard to texture inpainting network, we finetune the original MADF-Net [26] on our synthetic data. Our dataset has 3000 images and 7300 masks for T-shirt-panels and 12000 images and 12000 masks for trousers panels. We have trained the network for both T-shirts and trousers panels for 20 epochs. Since, the back part of the sleeves is occluded in the frontal pose, for inpainting the sleeve panel we assumed that sleeves are symmetric.

4.4. Results

First, we discuss the results of our landmark detection module. We use normalized mean squared error (NMSE) [61] to evaluate our Landmark detection network and report quantitative results in Table 1. We compare **JPPNet(scratch)-1000** and **JPPNet(synthetic+finetune) - 1000** on the same test set of 300 images (1000 denotes number of **real images** used during training/finetuning). Since the authors of JFNet [23] did not provide code and dataset for evaluation, we report the results from their paper. However, the comparisons are made for the same set of land-



Figure 8: Qualitative comparison of our approach with Pix2Surf [22] and PIFu [37].

marks for comparison. The performance of our JPPNet (scratch)-1000 is comparable to JFNet* -2000 (uses 2000 images for training), although we use a significantly lesser number of real-world training images (almost 50% less). We further see that a network trained on synthetic data and finetuned on real-world data performs almost twice as good as JFNet in terms of NMSE. Another important observation to note is that the errors in our best performing model i.e., JPPNet (synthetic+finetune)-1000 are of the same order as that of human annotations, which have an NMSE of 0.006 primarily contributed by the difference in the exact position of the landmarks among different annotators. We also conducted an ablation study by varying the size of training data and reported NMSE values, as shown in Table 2. We trained with 250, 500 and 1000 real-world training images in scratch and finetuned settings, e.g., JPPNet (synthetic + finetune)-250 uses 5000 synthetic images for initial training and 250 real images for finetuning. We can observe that JPPNet (synthetic + finetune)-250, outperforms JPPNet(scratch) -1000 thereby establishing the fact that training on synthetic data significantly improves the performances while reducing the dependence on expensive manually annotated real-world data. Figure 6 shows the qualitative results on landmark detection and segmentation task on real-world test images. As we can observe, our model predicts accurate landmark locations and smooth and precise segmentation maps.

Further, we discuss the qualitative and quantitative results of our texture inpainting network. We use widely used image similarity metrics PSNR and SSIM (as used in [62]) for quantitative evaluation. Table 3 shows the quantitative evaluation of networks finetuned on our synthetic data in comparison to the original network (pretrained on the Places2 [55] dataset) inferred on our synthetic data. We can observe that finetuning on our synthetic data significantly

Method	T-shirts	Trousers
JFNet* -2000	0.031	0.022
JPPNet (scratch) -1000	0.029	0.026
JPPNet (synthetic+finetune) -1000	0.014	0.012

Table 1: Evaluation of the landmark detection module using normalized mean squared error (lower the better). * denotes results reported using JFNet evaluated on a different dataset.

Method variant	T-shirts	Trousers
scratch -250	0.192	0.212
scratch -500	0.095	0.102
scratch -1000	0.029	0.026
finetune -250	0.022	0.018
finetune -500	0.015	0.014
finetune -1000	0.014	0.012

Table 2: Performance of landmark detection module on different variants of the dataset used for training.

improves the reported performance on both the metrics on our test data (981 sample images for T-shirts and 1187 sample images for trousers). Figure 7 shows the output of TPS-based texture transfer followed by the output of the texture inpainting network. Figure 9 (a) and (b) shows the result of the inpainting network trained on our synthetic data. We also test our finetuned models on real data and portray the generalizability of our models trained on synthetic data as depicted in Figure 9 (c) and (d).

Since there is no ground-truth data for real world catalog images, and the garment mesh geometry varies across methods, we perform a qualitative comparison following [22]. We qualitatively compare our method with two existing approaches PIFu [37] and Pix2Surf [22] (which also predicts

Method	Clothing	PSNR	SSIM
MADF-Net pretrained	T-shirts	41.16	0.946
MADF-Net finetuned	T-shirts	42.41	0.977
MADF-Net pretrained	Trousers	46.22	0.985
MADF-Net finetuned	Trousers	46.39	0.99

Table 3: Quantitative results of texture inpainting on synthetic data (Higher the better for both PSNR and SSIM).

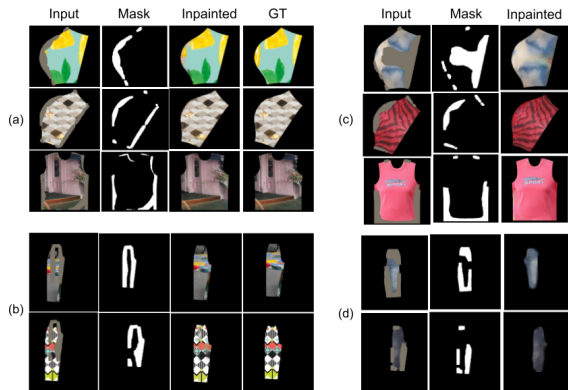


Figure 9: **Qualitative results of inpainting network:** on synthetic data (a) & (b) and on real-world data (c) & (d) (GT is not available for real data.)

the UV map for the garment meshes) in Figure 8. As can be seen clearly, PIFu [37] and Pix2Surf [22] failed to preserve the high-frequency texture details, whereas our method can preserve such fine texture details. Additionally, PIFu lacks the ability to model the geometry of the cloth separately. The bottom row in Figure 8 shows that unlike PIFu and Pix2Surf, our method can handle self-occlusions.

Please refer to our supplementary material for an extended set of results and 3D visualization of qualitative results in video format.

5. Discussion

We have shown the effectiveness of our method on both T-shirts and Trousers. However, it will be interesting to explore an extension of our approach on loose clothing scenarios (e.g., skirts and gowns) as well as relaxing the template mesh constraint for modeling free-form clothing and garment sizing.

Our landmark detection module can achieve good performance on the NMSE metric, where we observed that the error reported by our method is close to that of a human annotator. However, our method doesn't perform well on those real catalog images which are significantly far from the training data distribution of synthetic images, for e.g. if the region of interest is cropped out or severe self-occlusion is present, as shown in Figure 10 (a). The deep inpainting network, which is trained only on synthetic data, can



Figure 10: **Failure Cases:** of Landmark Detection Module (a), and Texture Mapping & Inpainting Module (b).

handle hand occlusions well and can also fill in the missing information along the border of individual panels in the UV map. Although a standard cataloging procedure tries to avoid occluding the regions of the garment with fine details, real world catalog images can have occlusions, along with extreme poses and shadows for enhanced realism. In such non-standard scenarios, the inpainting network fails to achieve the level of perfection needed. One such instance is shown in Figure 10 (b). The presence of wrinkles and folds in the real-world catalog images in comparison to our synthetic dataset aggravates the challenge in such scenarios. In regard to normalization for the garment sizes, since we have employed a fixed template garment mesh, the final digitized garment is robust to further modelling of geometrical deformations due to changes in human pose, body shape and garment sizes, that are essential for applications such as 3D VT. Our proposed TPS based texture transfer might yield slightly inaccurate texture mapping in regions with complicated geometrical structure or perspective distortions. Nevertheless, given the pose variations of fashion catalog images, such occurrences will be minimal.

6. Conclusion & Future Work

We propose a deep learning-based 3D garment digitization technique that is robust and can generalize well to real-world fashion catalog images. We developed a data generation pipeline that is capable of generating synthetic data with diverse poses and textures for any garment style. This synthetic data helped us train a landmark detection network and a texture inpainting network that is then finetuned with a relatively smaller number of real annotated images. We observed that our training strategy allows both of the networks to generalize well to unseen real images, indicating that our synthetic data is diverse enough to capture the characteristics of real-world images. The proposed pipeline is compact and scalable. The template-based digitization of the garment allows the transfer of the texture easily and avoids loss in texture quality. As part of future work, it will be interesting to explore domain adaptation techniques to further exploit learning from synthetic data.

References

- [1] The state of the ecommerce fashion industry: Statistics, trends & strategy. *Shopify : Industry Insights and Trends*, 2019.
- [2] BodyBlock.ai. Consumers to return half of online clothing purchases this holiday season. *Cision*, 2018.
- [3] Hannu Saarijärvi, Ulla-Maija Sutinen, and Lloyd C. Harris. Uncovering consumers’ returning behaviour: a study of fashion e-commerce. *The International Review of Retail, Distribution and Consumer Research*, 27(3):284–299, 2017.
- [4] Hyunwoo Hwangbo, Eun Hie Kim, So-Hyun Lee, and Young Jae Jang. Effects of 3d virtual “try-on” on online sales and customers’ purchasing experiences. *IEEE Access*, 8:189479–189489, 2020.
- [5] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7543–7552, 2018.
- [6] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [7] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10471–10480, 2019.
- [8] Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. Disentangled cycle consistency for highly-realistic virtual try-on. *CoRR*, abs/2103.09479, 2021.
- [9] Zhenyu Xie, Xujie Zhang, Fuwei Zhao, Haoye Dong, Michael C. Kampffmeyer, Haonan Yan, and Xiaodan Liang. Was-vton: Warping architecture search for virtual try-on network, 2021.
- [10] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. End-to-end learning of geometric deformations of feature maps for virtual try-on, 2019.
- [11] Zhonghua Wu, Guosheng Lin, Qingyi Tao, and Jianfei Cai. M2e-try on net: Fashion from model to everyone, 2019.
- [12] Bin Ren, Hao Tang, Fanyang Meng, Runwei Ding, Ling Shao, Philip H. S. Torr, and Nicu Sebe. Cloth interactive transformer for virtual try-on, 2021.
- [13] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
- [14] Surgan Jandial, Ayush Chopra, Kumar Ayush, Mayur Hemani, Balaji Krishnamurthy, and Abhijeet Halwai. Sievenet: A unified framework for robust image-based virtual try-on. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.
- [15] Shion Honda. VITON-GAN: virtual try-on image generator trained with adversarial loss. *CoRR*, abs/1911.07926, 2019.
- [16] Amir Hossein Raffiee and Michael Sollami. Garmentgan: Photo-realistic adversarial fashion transfer, 2020.
- [17] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing, 2021.
- [18] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating+preserving image content, 2020.
- [19] Blender - a 3d modelling and rendering package. <http://www.blender.org>.
- [20] Ann FairhurstAnn Fairhurst Vertica Bhardwaj. Fast fashion: Response to changes in the fashion industry. *The International Review of Retail Distribution and Consumer Research Distribution and Consumer Research(1):165-173*, 2010.
- [21] Roberta Pezzetti Patrizia Gazzola, Enrica Pavione and Daniele Grech. Trends in the fashion industry. the perception of sustainability and circular economy: A gender/generation quantitative approach. *Sustainability 12(7):2809*, 2020.
- [22] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. Learning to transfer texture from clothing images to 3d humans. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 7021–7032. IEEE, 2020.
- [23] Yi Xu, Shanglin Yang, Wei Sun, Li Tan, Kefeng Li, and Hui Zhou. 3d virtual garment modeling from rgb images. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 37–45, 2019.
- [24] Alexander C. Berg and Jitendra Malik. *Shape Matching and Object Recognition*, pages 483–507. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.
- [25] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
- [26] Manyu Zhu, Dongliang He, Xin Li, Chao Li, Fu Li, Xiao Liu, Errui Ding, and Zhaoxiang Zhang. Image inpainting by end-to-end cascaded refinement with mask awareness. *IEEE Transactions on Image Processing*, 30:4855–4866, 2021.
- [27] Yuying Ge, Ruimao Zhang, Lingyun Wu, Xiaogang Wang, Xiaou Tang, and Ping Luo. A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. *CVPR*, 2019.
- [28] Nikolay Jetchev and Urs Bergmann. The conditional analogy gan: Swapping fashion articles on people images, 2017.
- [29] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network, 2018.
- [30] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10510–10519, 2019.

- [31] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [32] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation, 2019.
- [33] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3d people in generative clothing, 2020.
- [34] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. *CoRR*, abs/2103.06871, 2021.
- [35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.
- [36] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. *CoRR*, abs/1908.07117, 2019.
- [37] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019.
- [38] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [39] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans, 2020.
- [40] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction, 2020.
- [41] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus A. Magnor. Tex2shape: Detailed full human body geometry from a single image. *CoRR*, abs/1904.08645, 2019.
- [42] Chaitanya Patel, Zhouyingcheng Liao, and Gerard 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, jun 2020.
- [43] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. Garnet: A two-stream network for fast and accurate 3d cloth draping. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019.
- [44] Igor Santesteban, Nils Thuerey, Miguel A. Otaduy, and Dan Casas. Self-supervised collision handling via generative 3d garment models for virtual try-on, 2021.
- [45] Masahiro Sekine, Kaoru Sugita, Frank Perbet, Björn Stenger, and Masashi Nishiyama. Virtual fitting by single-shot body shape estimation. In *Int. Conf. on 3D Body Scanning Technologies*, pages 406–413. Citeseer, 2014.
- [46] Zorah Laehner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling, 2018.
- [47] Derek Bradley, Tiberiu Popa, Alla Sheffer, Wolfgang Heidrich, and Tamy Boubekeur. Markerless garment capture. In *ACM SIGGRAPH 2008 papers*, pages 1–9. 2008.
- [48] Bin Zhou, Xiaowu Chen, Qiang Fu, Kan Guo, and Ping Tan. Garment modeling from a single image. *Computer Graphics Forum*, 32, 10 2013.
- [49] Shan Yang, Zherong Pan, Tanya Amert, Ke Wang, Licheng Yu, Tamara Berg, and Ming Lin. Physics-inspired garment recovery from a single-view image. *ACM Transactions on Graphics*, 37:1–14, 11 2018.
- [50] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019.
- [51] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *European Conference on Computer Vision*, pages 512–530. Springer, 2020.
- [52] Yu Shen, Junbang Liang, and Ming C Lin. Gan-based garment generation using sewing pattern images. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 225–247. Springer, 2020.
- [53] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: Clothed 3d humans. In *European Conference on Computer Vision*, pages 344–359. Springer, 2020.
- [54] Radek Daneczek, Endri Dibra, A. Cengiz Öztireli, Remo Ziegler, and Markus Gross. Deepgarment : 3d garment shape estimation from a single image. *Computer Graphics Forum (Proc. Eurographics)*, 36(2), 2017.
- [55] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.
- [56] J de Joya, R Narain, J O’Brien, A Samii, and V Zordan. Berkeley garment library. <http://graphics.berkeley.edu/resources/GarmentLibrary/index.html>, 2012.
- [57] J de Joya, R Narain, J O’Brien, A Samii, and V Zordan. Makehuman: Open source tool for making 3d characters. <http://www.makehumancommunity.org/>.
- [58] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6173–6183, 2020.

- [59] Cmu graphics lab motion capture database. <http://mocap.cs.cmu.edu>.
- [60] João Cartucho, Samyakh Tukra, Yunpeng Li, Daniel S. Elson, and Stamatia Giannarou. Visionblender: a tool to efficiently generate computer vision datasets for robotic surgery. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–8, 2020.
- [61] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild, 2016.
- [62] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369. IEEE, 2010.