From Model to Person Virtual Try-On of Clothes

A thesis submitted in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in
Computer Science

by

Debapriya Roy

Under the supervision of

Prof. Bhabatosh Chanda

Electronics and Communication Sciences Unit
Indian Statistical Institute

February, 2022

To my father, my mother and my little sister.
They supported me in every possible way to learn, and to pursue a life of my choice.

Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Bhabatosh Chanda. I am immensely grateful to him not only for accepting me as his student but also for teaching me the very basic steps of independently conducting quality research. His immense knowledge and excellent grasp on computer vision and image processing helped me understand the topic better. Without his invaluable insights and advice, this thesis could have never been a reality.

I am greatly indebted to Dr. Diganta Mukherjee for motivating me to pursue a research career. I feel fortunate enough to be blessed by his guidance and encouragement throughout my evolution from the beginning of my research career. I would also like to thank Dr. Swagatam Das whose way of teaching extremely inspired me and helped me to get a grip on some of the important topics related to machine learning and deep learning. I am grateful to Prof. Dipti Prasad Mukherjee and the rest of the dignified professors of the Indian Statistical Institute, Kolkata for offering me their helpful suggestions whenever I needed. I like to express my gratitude to the staff of the Electronics and Communication Sciences Unit, the Dean's, and the Director's offices for their active support and Indian Statistical Institute for funding my research and Electronics and Communication Sciences Unit for providing the required infrastructure.

I am thankful to Dr. Ranjan Mondal who introduced to me the research areas of computer vision and motivated me to learn many new topics. My endless gratitude goes to Dr. Sanchayan Santra who constantly helped me during my work with some great discussions which finally helped to reshape my knowledge. I would also take the opportunity to thank Dr. Sankha Subhra Mullick and Dr. Kaustav Nag for their helpful discussions that enhanced my knowledge and greatly helped me to enhance my style of scientific article writing.

I express my sincere gratitude to my father and my mother, whose principles in life, rationality made me a better person and motivated me to pursue a quality life. I would thank my little sister and my husband for their endless love, support. I also appreciate the support of my friends and my extended family members. I would also like to thank my fellow research scholars of ECSU department for helping me whenever I needed.

Abstract

The advent of e-commerce has given us easy access to purchasing products regardless of location and time of day. However, a study shows 30% of online shoppers deliberately over-purchase and subsequently return unwanted items, while 19% admit to ordering multiple versions of the same item so that they could make up their minds when products are delivered and are tried out [1]. To drive smoother customer experiences the marketers are taking an interest to invest in technologies like virtual try-on (VTON). VTON is an Augmented Reality (AR) application that allows a user to try a product before actually buying it. With this technology, product returns may be reduced and, in turn, so is the transport expenses. In addition, VTON may also increase customer loyalty, attract new customers, provide near to in-store shopping experiences to the customers, and many more. With some of the world-famous fashion companies like Michael Kors, Warby Parker, M·A·C Cosmetics, Diamond Hedge, Gucci, Nike, and Lenskart, VTON has now made its way in clothing, shoes, jewelry, Makeup, and even in furniture industry [2]. In this work we explore VTON approaches in the clothing domain.

Considering the cost intensiveness of 3D model-based VTON approaches, the researchers have focused on the image-based VTON system, where the primary inputs are a reference clothing image and the image of the person who wants to try the clothing. In general, most of the existing approaches are data-based learning systems and use clean clothing images as reference clothing. However, in the major data sources, e.g., e-commerce websites or social media, the displayed images of clothing articles, in general, are in the form of a model or a general user wearing them. This work takes an attempt to solve this problem considering the reference clothing in the form of a human wearing it. This is usually denoted as the model-to-person (M2P) try-on problem. In general, the existing VTON methods follow a two-staged approach. In the first stage, the clothes are aligned according to the target person's body shape and pose, and, in the second stage, the warped clothes are combined with the target person's image. A warping approach widely explored by the existing works is to employ a geometric matching network which is a learnable equivalent of traditional feature matching methods. This is used to learn and match features from the source clothing and the target person images to predict the parameters of thin plate spline

[1] https://home.barclaycard/press-releases/

[2] https://www.divante.com/blog/examples-of-ar-powered-virtual-try-ons-in-the-fashion-industry

(TPS) transform, which is primarily a warping function. The predicted transform is then used to align the source clothing in the body shape and pose of the target person. However, the human body undergoes very restricted deformation because of the unique organization of its bones and muscles, and without explicit consideration of human structure related constraints the network falls short in computing accurate warps especially in case of complex textures of the clothing. Our first work takes an attempt towards this direction where we employ the correspondences between the structural key points of humans and clothes between the model and the target person to compute the parameters of the TPS transform. Explicit consideration of the structural constraints in the form of landmarks aids us in computing more accurate target warps which result in a better FID (Fréchet Inception Distance) score over the state-of-the-art. However, employing correspondences of all human and clothing landmarks fail to give satisfactory result in case of folded or cross-armed postures. In other words, this method is restricted to only simple human poses. Moreover, annotations are costly, and, this work requires both human and fashion landmark annotation. Thus instead of explicitly specified features i.e., landmarks, inspired by the capability of deep neural networks in learning task-specific features, in the next work we intend to explore the correspondences of a rich body shape and pose representations, i.e., densepose between the model and the person. We employ a geometric matching network to learn and correlate the features learned from the denseposes of the two and thereby predict the parameters of TPS. Though TPS is a well-explored transformation for clothes warping, this has some inherent limitations which limit its applicability in such problem domain where parts of the object can make a significant movement. For instance, human arms can move in a variety of ways, but the second-order difference constraint of TPS restricts the bending of the axis in the target grid. Hence, it falls short in modeling the cases when the source or the target person is posing with his arms folded or bent. We address this issue using a hand-crafted feature-based warping technique that exploits the human landmarks as well as the human limb correspondences to compute the target warp. In addition, to refine the fit of the computed warp and to synthesize the final try-on output we propose two learnable deep neural networks. Extensive experiments show the potential of our approach over the state-of-the-art.

Contents

List of Figures

List of Tables

List of Notations

M	Image of the model wearing the reference clothing.
P	Image of the person willing to try the model's clothing.
P'	Image of the person wearing the model's clothing i.e., the try-on result.
M_{dp}	Densepose representations of the model.
P_{dp}	Densepose representations of the person.
$\hat{P}_{dp}$	Predicted densepose representations of the person.
c	reference clothing of the model.
c'	target warp of the reference clothing of the model.
c'_m	mask of the predicted warp.
$\bar{c'_m}$	the complement of c'_m.
$c_{torso}, c_{lsleeve}, c_{rsleeve}$	semantic parts (torso, left-sleeve, right-sleeve) of c.
$c'_{torso}, c'_{lsleeve}, c'_{rsleeve}$	semantic parts (torso, left-sleeve, right-sleeve) of c'.
P_{hlm}	Human landmarks of the person.
M_{hlm}	Human landmarks of the person.
α_i^P	i^{th} human landmark of the person a 2-tuple.
α_i^M	i^{th} human landmark of the model.
β_j^M	j^{th} fashion landmark of the model's clothing.
β_j^P	j^{th} fashion landmark of the target warp.
n_h	the total number of upper body landmarks of a human used in this work which is in total 9.
n_f	the number of fashion landmarks of a clothing used in this work which in total is 6.
$P_{hlm} = \{\alpha_1^P, \alpha_2^P, \ldots, \alpha_{n_h}^P\}$	the multiset of human landmarks of the person.
$M_{hlm} = \{\alpha_1^M, \alpha_2^M, \ldots, \alpha_{n_h}^M\}$	the multiset of human landmarks of the model.
$c_{flm} = \{\beta_1^c, \beta_2^c, \ldots, \beta_{n_f}^c\}$	the multiset of fashion landmarks of the reference clothing c of the model.
$c'_{flm} = \{\beta_1^{c'}, \beta_2^{c'}, \ldots, \beta_{n_f}^{c'}\}$	the multiset of fashion landmarks of the target warp c' of the model's clothing c.
$\mathscr{R} = \{\alpha_1^M, \alpha_2^M, \ldots, \alpha_{n_h}^M, \beta_1^c, \beta_2^c, \ldots, \beta_{n_f}^c\}$	combined mulltiset of M_{hlm} and c_{flm}.
$\mathscr{T} = \{\alpha_1^P, \alpha_2^P, \ldots, \alpha_{n_h}^P, \beta_1^{c'}, \beta_2^{c'}, \ldots, \beta_{n_f}^{c'}\}$	combined mulltiset of P_{hlm} and c'_{flm}.

$\mathbf{r}_j$, $\mathbf{t}_j$	the j^{th} element of $\mathscr{R}$, $\mathscr{T}$.
$H[\cdot]$	An objective function to be minimized.
N	the total number of elements in each of the multiset $\mathscr{R}$ and $\mathscr{T}$.
n	A general notation representing the number of control points in the TPS transform.
S	the target mask predicted by MGM.
$\mathscr{S}$	the clothing mask predicted by MPN.
R	is the combined representation of P and c'.
$\mu_r, \sigma r$	mean and covariance of the embedding of the data which is assumed to follow a continuous multivariate Gaussian distribution.
$\mu_g, \sigma g$	mean and covariance of the embedding of the generated data which is assumed to follow a continuous multivariate Gaussian distribution.
I_o	An intermediate VTON output of ISM.
I_m	A predicted mask that combines I_o and R in ISM.
$\mathbb{N}$	the set of all natural numbers.
$\omega : \mathbb{N}^2 \to \mathbb{N}^2$	a thin plate spline (TPS) transformation function.
ω_{xx}, ω_{xy}, ω_{yy}	the second-order gradients of $\omega(\cdot)$.
ν	a radial basis function kernel used in TPS.
λ	a regularization parameter.
G	a generator.
D	a discriminator.
p_{data}	the data distribution.
p_g	the data distribution learned by the generator G.
p_z	a prior noise distribution.
$\mathscr{N}(\cdot,\cdot)$	Normal distribution.
$\mathbf{x}$	random variable representing the input data following the distribution p_{data}.
$\mathbf{y}$	random variable representing the conditional information.
$\mathbf{z}$	random variable representing the noise sample following the distribution p_z.
$\mathbb{E}$	Expectation.
$\mathscr{F}$	the fashion landmark predictor network.
$F_i(\mathbf{x})$	the activation at the i^{th} layer of VGG-19.
$\tilde{F}$	a geometric matching network.
θ	parameters of the TPS transform.

$\tilde{g}(.,.)$	a feature extraction network.
$f_{tps}(.,.)$	a geometric transformation.
$T(\cdot,\cdot)$	the general notation of a warping function that relates each point in the source image to the corresponding point in the target image.
$\mathscr{F}_{fwm}$	Forward mapping function.
$\mathscr{F}_{bwm}$	Backward mapping function.
A,B,C	2D-points in the mesh corresponding to the target warp.
A',B',C'	2D-points in the mesh corresponding to the source clothing. Each corresponds to its equivalent points A,B,C in the target mesh.
ϕ_1,ϕ_2	the angles in polar coordinates of X relative to line BA and BC.
ϕ	sum of ϕ_1 and ϕ_2.
r,r'	the radius in polar coordinates of X relative to line BA and BC.
ϕ_1',ϕ_2'	the angles in polar coordinates of X relative to line $B'A'$ and $B'C'$.
ϕ'	sum of ϕ_1' and ϕ_2'.
f,g,h	some function.
X	an arbitrary point in the mesh corresponding to P.
X'	the mapping of X to the mesh corresponding to M.
$f_a,f_b \in \mathbb{R}^{1\times l}$	feature vectors.
C_{ab}	a matrix representing correlation values between the features f_a and f_b.
$(\cdot)^T$	matrix transpose.
p,a	some scalars.
l_{w1},l_{w2},l_{w3}	loss weights.

List of Abbreviations and Acronyms

TPS	Thin Plate Spline.
M2P	Model-to-Person.
C2P	Cloth-to-Person.
VTON	Virtual Try-On.
FID	Fréchet Inception Distance.
IS	Inception Score.
GAN	Generative Adversarial Network.
cGAN	Conditional Generative Adversarial Network.
SSIM	Structural SImilarity Metric.
DSSIM	Structural DiSSImilarity Metric.
PGWM	Pose Guided Warping Module.
MGM	Mask Generator Module.
ISM	Image Synthesizer Module.
MPN	Mask Predictor Network.
ISN	Image Synthesizer Network.

Chapter 1

Introduction to the Problem of Virtual Try-on

1.1 Introduction

With the proliferation of internet usage, the world of retailing has witnessed a lot of changes in the last two decades in terms of consumer expectations, new retailing technologies, and new transaction methods. The advent of e-commerce has given us easy access to purchasing products regardless of location and time of day. Statista, a company specializing in market and consumer data, anticipated a 246.15% growth in global e-commerce sales, rising from $1.3 trillion in 2014 to $4.5 trillion in 2021 [1]. That amounts to an almost threefold growth in online revenue. It is projected to grow to $6.54 trillion in 2022. However, there are some pushbacks faced by online retailers. For instance, it is difficult to keep shoppers engaged in online retail. Moreover, 30% of online shoppers deliberately over-purchase and subsequently return unwanted items, while 19% admit to ordering multiple versions of the same item so they could make their mind up when products are delivered and tried out [2]. To drive smoother customer experiences to push the growth in sales further marketers are taking interest to invest in high-end technologies like Augmented Reality (AR) and Virtual Reality (VR). Virtual try-on (VTON) is an AR application that allows a user to try a product before actually buying it. This technology aids in reducing product returns, which curbs transport expenses of retailers that were caused due to return orders [3]. In addition, VTON may also increase customer loyalty, attract new customers, provide near-to in-store shopping experiences to the customers, and many more. All these finally lead towards growth in sales [4] [5].

Initial approaches to VTON were mostly proposed based on 3D modeling; where detailed 3D models of humans and clothing are built from either depth cameras (Sekine et al., 2014) or multiple 2D

[1] https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/

[2] https://home.barclaycard/press-releases/

[3] https://www.citrusbits.com/how-augmented-reality-is-being-used-in-ecommerce/

[4] https://www.technologyreview.com/2019/10/23/238473/augmented-reality-in-retail-virtual-try-before-you-buy/

[5] https://en.wikipedia.org/wiki/Augmented_reality

images (Bogo et al., 2016a). This enables realistic clothing simulation under geometric and physical constraints, with control of the viewing direction, lighting, pose, and texture. However, the cost of data capture, annotation, and computation is overhead here. Also, the need for specialized devices, such as 3D sensors, might cause an additional cost. These large costs limit the scalability of these methods in terms of the number of customers and garments. Considering this, recently image-based VTON problem has gained the attention of the research community. This problem requires the images of the clothing and the person willing to try the product as inputs. Nowadays capturing images is easy due to the availability of high-quality cameras even on mobiles. However, discarding the depth dimension of the input data reduces the amount of input information which makes the problem more challenging. Note that from now on by virtual try-on we will mean image-based virtual try-on only.

1.1.1 Motivation and objective

Approaching the VTON problem as a data-based learning process requires a huge amount of data which is indeed available on e-commerce websites, and social media sites. However, such websites either show the images of the advertised outfit or display the image of a professional model wearing the clothing. Only a very few provide both images. Even on social media, people post their images wearing different outfits. Hence, images of the same person wearing different outfits are not available from such data sources. So the challenge in terms of data for supervised learning is clear. This motivated us to pursue our research on model-to-person (M2P) try-on methods, where the clothing source is considered to be in the form of a model (a human or a mannequin) wearing it. The other form of VTON problem popular in VTON literature and has been taken up by most researchers is cloth-to-person (C2P). This in contrast with the M2P methods is simpler where a separate clothing image is available along with a corresponding model's image wearing that outfit. Note that clothing and outfit both refer to the same things.

The existing VTON methods have mostly explored the warping-based approach (Wang et al., 2018a; Han et al., 2018; Roy et al., 2022, 2020; Yu et al., 2019; Dong et al., 2019a; Wang et al., 2018c; Hsieh et al., 2019a,b; Yang et al., 2020; Raffiee and Sollami, 2021; Jandial et al., 2020) which is mainly a two-staged approach. First, the source clothing is warped to fit the person, where warping may be considered as a mapping $T : \mathbb{Z}^2 \to \mathbb{Z}^2$ ($\mathbb{Z}$ denotes the set of integers) relating each point in the source to the corresponding point in the target, or vice versa. In the next stage, the warped cloth i.e., the predicted target warp is combined with the person's image to synthesize the final try-on output. For warping, almost all of the existing solutions took the deep neural network-based feature learning approach to predict the parameters of a predefined warping function. The estimated parameters are then used to compute the aligned source clothing that is expected to fit the person. This approach has performed well in the case of clothes with simple textures or single-colored clothes. But in presence of patterns, e.g., stripes, floral, the network predicts inconsistent transformation without additional regulariser to maintain structural consistency.

In contrast to learning the transformation directly from data, this work imposes the idea that transferring the clothing from model to person can be represented in the form of the change of body shape and pose from the model to the person. Based on this idea we explore different forms of estimates of body shape and pose to computer the target warp of the source clothing. Precisely, in this work we explored two different kinds of body shape and pose estimates, structural key points i.e., landmarks, and dense human pose representation i.e., densepose (Alp Güler et al., 2018). In terms of the warping function, we explored the thin plate spline (TPS) transform (Grimson, 1981; Duchon, 1977) for our first two works. However, this transformation has some limitations in the current problem context. In order to achieve a more generalized solution, we propose a hand-crafted feature-based warping method which is inspired by the Beier and Neely transform (Beier and Neely, 1992). In addition, we also make contributions in terms of combining the warped clothing with the person's image in order to synthesize the final try-on output.

Below we present a literature survey on the existing image-based VTON methods and thereafter discuss briefly our chapter-wise contributions.

1.2 Literature Survey

To date, a variety of image-based VTON methods have been proposed. Over the last few years, image-based VTON has obtained immense research attention, due to its various applicability in real-world problems. There exists a large body of literature Jetchev and Bergmann (2017); Han et al. (2018); Wang et al. (2018a); Roy et al. (2022, 2020); Yu et al. (2019); Dong et al. (2019a); Wang et al. (2018c); Hsieh et al. (2019a,b); Chen et al. (2018); Yang et al. (2020); Raffiee and Sollami (2021); Jandial et al. (2020); Yildirim et al. (2019); Chen et al. (2018); Zheng et al. (2019b); Men et al. (2020); Lewis et al. (2021); Kubo et al. (2019) in this domain. However, to keep this thesis concise we discuss a few of them. A detailed literature survey can be found in Cheng et al. (2021).

From the perspective of computing the target warp, the VTON methods can be overall categorized into two types, warping-based, and synthesis-based. In general, the warping-based methods generally approach VTON as a two-step problem, where first the source clothing is warped and then combined with the target person's image. The synthesis-based methods mostly employ a deep learning-based image synthesis approach. While most of the methods are based on the first approach, possibly the first image-based VTON (referred to as VTON from now on) solution CAGAN (Jetchev and Bergmann, 2017) is based on the second kind. Here the authors trained a conditional generative adversarial network (cGAN) with cycle consistency loss (Zhu et al., 2017) to learn the relation between an outfit and its appearance when rendered on a human. The concept of cycle consistency loss was first introduced in (Zhu et al., 2017) where the authors introduced a novel method of unpaired training; Where given two unpaired image domains, mapping between the image domains are learned without explicitly requiring paired examples. However, there is an underlying assumption of some relationship between the two domains. For example, consider the case when the two domains are two different renderings of the same underlying scene. In CAGAN the authors employ the concept of this loss based on the intuition that given a human image and a different clothing if the clothing region of the human image is unrelated to the clothing image then while swapping this clothing with the source clothing the generated image will be penalized for deviating from the original image. While CAGAN showed some promising results but along with the try-on clothing and the image of the person this method also requires a separate image of the current clothing of the person. Therefore not practically applicable to test on a general random user, whose clothing image is not present in the dataset. Also, this method is restricted to simple human poses and single-colored clothing with less variety in texture. VITON (Han et al., 2018) proposed a warping-based solution approach that somewhat relaxed the data constraint while achieving improved performance. In this method, during testing, only the clothing image suffices. It works in 2 stages - an encoder-decoder generator stage and a refinement stage. the first stage predicts a foreground mask of the source clothing on the target person. For warping the source clothing VITON computes the thin-plate spline(TPS) transformation (Grimson, 1981; Duchon, 1977) with shape context matching Belongie et al. (2002) between the mask of in-shop clothing and the predicted foreground mask. Shape context is a hand-crafted feature descriptor used to describe the shape and the matching of two shapes. Primarily shape context is used to measure similarity between two shapes. At a reference point, the shape context descriptor measures the distribution of the remaining points relative to it. Therefore, the corresponding points on two similar shapes will have similar shape contexts. The shape context approach considers the shape of an object can be captured by a finite set of points on the internal or external contours of the object, where the contour is computed by some edge detection method. Sampling an equal number of edge detector points on the two shapes, for each point on one shape, the goal is to find the best matching point on the second shape. Now, this matching is based on a novel descriptor i.e., the shape context. For each point on a set of points the descriptor is computed as a coarse histogram of the relative coordinates of the remaining points. This histogram is defined to be the shape context of that point. VITON has apparently two problems - limitations of the shape context warp and loss of details of the warped clothing. CP-VTON (Wang et al., 2018a) showed improvement over VITON (Han et al., 2018) by predicting a mask in order to retain better clothing details. In addition, it employed a geometric matching network (GMN) (Rocco et al., 2017) to predict the target warp and also learned a mask. GMN with TPS as the underlying transformation function to predict the target warp from the images of the source clothing and the target person; where GMN is a neural network that learns the parameters of a geometric transformation (usually TPS) from the training set of source and target image pairs. In other words, it is a learnable equivalent of traditional feature matching. However, the high flexibility of TPS, often causes GMN to produce undesirable warping results in presence of complex patterns in clothes. VTNFP (Yu et al., 2019) follows the warping approach of CP-VTON but additionally incorporates non-local mechanism (Wang et al., 2018c) in the feature extraction part of the GMN to improve feature learning and matching. While it did not address the problems of GMN but showed better results in keeping the details of non-target body parts in the final try-on output. ACGPN (Yang et al., 2020) attempts to solve the warping issues of CP-VTON by employing a second-order difference constraint to control the deformation of the target grid. It employs three modules: the first module predicts the semantic layout of the target person image, where semantic layout refers to the human parsing of the person after try-on. This is followed by the cloth warping module which predicts the target warp according to the previously predicted semantic layout and finally a try-on module. MG-VTON (Dong et al., 2019a) addresses the problem of multi-pose VTON. As its warping strategy is similar to that of CP-VTON, therefore some issues related to GMN are still observed in its output. GarmentGAN (Raffiee and Sollami, 2021) employs two separate GANs for shape generation and appearance generation to predict the final try-on output. They employ the SPADE-style normalization layer proposed by (Park et al., 2019) to more accurately transfer spatial information from the source clothing to the reference person. They showed improvement over CP-VTON in preserving better clothing details e.g., logos or patterns. However, does not explicitly mention any improvement in the case of complex textured articles of clothing. SieveNet(Jandial et al., 2020) introduces a multi-stage coarse-to-fine warping module that learns the TPS parameters in two stages instead of one showing better performance in maintaining fine intricacies.A multi-stage coarse-to-fine warping network is proposed to predict the warped target clothing. Next, a texture translation network is produced that contributes toward improving the quality of the final try-on result. To train this network a dueling triplet loss is introduced. This loss is characterized by 3 things, an anchor, i.e., the current output, a positive (the ground-truth), and a negative (result of the previous training phase). The objective of this loss is to push the anchor towards the positive and push it away from the negative. Another method Zeng et al. (2020) proposed a two-stage image generation approach for generating clothing images from model images. The authors used the implementation of CP-VTON to show that with their reconstructed clothes image ‘a model to person try-on problem’ can be reduced to ‘a clothes to person try-on problem’. However, this way of solving VTON may not be economical due to the additional stages for clothes reconstruction. PASTA-GAN (Xie et al., 2021) proposed a patch-routed disentanglement module that disentangles style and spatial information in a garment. Its key contribution is employing the source and target person’s pose-based key points to decouple a garment into normalized patches and then reconstruct them to the warped garment. While we also employ pose key points but the idea of our geometric approach to warping is quite different from their approach.

Other than warping, appearance flow based approaches are also proposed Chopra et al. (2021); Ge et al. (2021); Han et al. (2019); He et al. (2022); Liu et al. (2019a). The very first work in this direction is proposed in Han et al. (2019), which predicts the clothing flow from the source to the target images. The main idea of appearance flow is to predict the sampling grid for clothes warping. In Han et al. (2019) the authors first a conditional layout generator to predict the segmentation layout conditioned on the target pose. This layout is employed to predict the appearance flow. The predicted appearance flow estimates the visual correspondences and helps in transferring the source clothing information to the target person. In Liu et al. (2019a) Liu et al. propose a unified framework to tackle human motion imitation, view synthesis, and appearance transfer. A liquid warping

GAN is proposed with a Liquid Warping Block (LWB) that propagates the source features extracted by a denoising convolutional auto-encoder to synthesize an image concerning the reference. In Ge et al. (2021) Ge et al. proposes a "teacher-tutor-student" knowledge distillation approach for VTON. Compared to the parser-based models this work aims to reduce the dependency of VTON methods on human parsing results by employing parser-based methods as "tutor knowledge". Instead of using real images as supervision, this work formulates the try-on problem as knowledge distillation that distills the appearance flow between the person and the garment image. In simpler terms, this predicts the dense correspondence between the garment and the person's image in order to produce try-on outputs. In Chopra et al. (2021) Gated Appearance Flow is used which uses aggregation of hierarchical appearance flow. A styleGAN Karras et al. (2019) based appearance flow estimation method is proposed in He et al. (2022) which uses the style vector to capture the global context of the image. This attempts to overcome the challenges due to considering only local appearance flow Chopra et al. (2021); Ge et al. (2021); Han et al. (2019).

To remove the constraint of requiring separate clothing images, M2E-TON (Wu et al., 2019) proposed a model to target try-on strategy, requiring only a model and a person image. It works in 3 stages, the first stage is dedicated to texture synthesis of the target person using densepose representations. Before going into further details let us discuss briefly on the densepose representations, as this is going to be used in the forthcoming chapters also. Densepose Alp Güler et al. (2018) represents dense correspondences between an RGB image and a surface-based representation of the human body, such as the SMPL (Skinned Multi-Person Linear) model Bogo et al. (2016a). In other words, the DensePose system is used to associate the new image with the common surface coordinates of an SMPL model. For a given RGB image this representation contains 3 values namely, I, U, and V for each of the pixels in the image. Therefore, this is an image size 3 channel array of values, where the first channel contains the 24 human part specific segmentation labels and the other two channel contains the UV map values. UV map values are position coordinates on a 2D texture that are stored within an associated vertex. In general, this is used for projecting a 2D image to a 3D model's surface for texture mapping. UV values are generally floating point values (to maintain precision) in the range [0, 1]. The axes U, and V refer to the axes of the 2D object, as X, Y, Z are in general used to denote axes of the 3D model. The densepose proposes two major contributions, first is their manual annotation system, second, using these annotations to train a CNN for learning this representation, and also using distillation-based ground-truth interpolation. The annotation system gathers annotations for 50K humans. In the process of annotation, annotators establish a dense correspondence between images and a 3D surface model. Segmenting an image into 25 semantic regions (24 body parts+ 1 for background) the annotators sample every body part with a set of roughly equidistant points (Maximum 14 points for each part). Each part is parameterized into UV coordinates. In terms of learning, a fully convolutional neural network (FCN) is proposed which in the first step does classification to label a pixel as belonging to either background or one among several region parts. In the second step, it does a dense regression task, i.e., computes the exact coordinate of a pixel within the corresponding part predicted in the previous step. The authors also do a distillation-based ground-truth interpolation where a “teacher” network is trained to reconstruct the observed ground-truth values and then it is deployed to predict over the whole image domain.

In M2E-TON, the authors have used the correspondences between the UV mappings obtained from the densepose representations of the source and the target person, and, also have used barycentric coordinate interpolation to compute texture mapping. The rest of the stages do the refinement of lost textures and generate the final output. However, the results of this method are not photo-realistic. Some GAN-based image synthesis approach for attribute manipulation is proposed in (Yildirim et al., 2019; Men et al., 2020; Lewis et al., 2021). Yildirim et al. (Yildirim et al., 2019) and Men et al. (Men et al., 2020) proposes try-on based on StyleGAN (Karras et al., 2019). Men et al. conditioned the model on the body pose, the person’s identity, and multiple garments. It generated separate latent codes for each of those components followed by combining them into a single output by borrowing the needed parts from each image. While it produces good results, but limited to only uniform colors and textures and also fails to synthesize the correct garment shape. Yildirim et al. [2019] similarly conditions on pose and clothing items, but showed results only on single-colored outfits with no or very fewer texture details. Recently, Lewis et al. (Lewis et al., 2021) proposed a styleGAN-based novel approach of try-on on unpaired training data. Given a pair of person image and a garment image they propose a method that automatically searches for optimal interpolation coefficients per layer, such that, when applied to the two images result in a try-on. Although it produces photo-realistic output, it is limited to only simple human poses and those clothing textures which are well presented in the latent space.

1.3 Overview of the Present Work

1. The VTON problem can be addressed in various settings. Most of the existing VTON methods require the image of the clothing, in addition to the image of the model (source) wearing that clothing and the person (target). While having a separate clothing image makes the modeling of VTON systems easier, separate clothes images are rarely available. Shopping websites mostly display images of models posing with the clothes, rather than separate clothes images. Also, people tend to post images on social media platforms donning a variety of clothing items. Hence, the assumption of the availability of a separate clothing image is hard to satisfy in practice; but relaxing this constraint makes the problem more challenging. Because, if a separate clothing image is not available, clothing information has to be extracted from the image of the model.

2. Of late, VTON methods are following a two-staged approach (Han et al., 2018; Wang et al., 2018a; Yu et al., 2019; Dong et al., 2019a; Yang et al., 2020). In the first stage, the clothes are aligned according to the target person’s body shape and pose, and, in the second stage, the warped clothes are combined with the target person’s image. The use of Thin-Plate Spline (TPS)

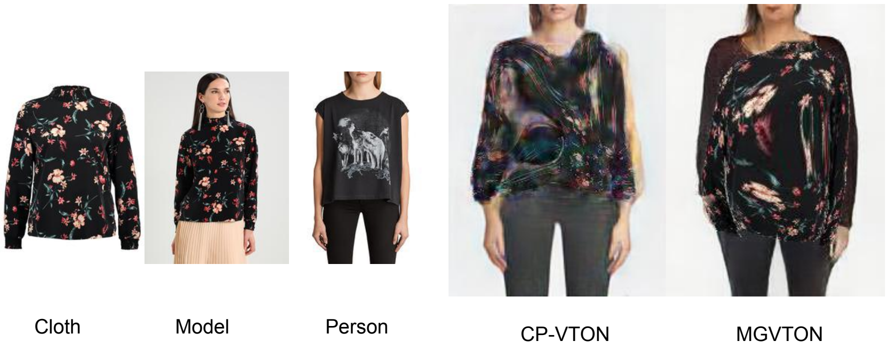


Figure 1.1: Demonstration of warping issues of the geometric matching module in MGVTON and CP-VTON.

transformation (Grimson, 1981; Duchon, 1977) is popular among VTON methods to warp the source clothes to the target shape. The parameters of this transformation are predicted either by shape context warp (Han et al., 2018) or CNN-based feature matching method called Geometric Matching (GM) (Wang et al., 2018a; Yu et al., 2019; Dong et al., 2019a; Yang et al., 2020). While the concept of GM was first proposed in (Rocco et al., 2017), CP-VTON (Wang et al., 2018a) showed its applicability in warping clothes for VTON. Although GM is a popular warping strategy, it results in inaccurate clothing deformations in the presence of complex patterns in the clothes (Jae Lee et al., 2019; Issenhuth et al., 2019; Yang et al., 2020). Fig. 1.1 demonstrates this problem in the results of CP-VTON and MGVTON. A possible reason for it might be the learned features do not encode the human body geometry well which along with the substantial flexibility in the TPS transform affects the performance. ACGPN (Yang et al., 2020) proposed to employ a second-order difference constraint to control the grid deformations. But, that is more of an avoidance measure than prevention and also does not consider any constraints related to human body geometry which will be more justifiable in this context.

3. Almost all the VTON methods (Wang et al., 2018a; Han et al., 2018; Roy et al., 2022, 2020; Yu et al., 2019; Dong et al., 2019a; Wang et al., 2018c; Hsieh et al., 2019a,b; Yang et al., 2020; Raffiee and Sollami, 2021; Jandial et al., 2020) use TPS as the warping function. However, TPS-based warping methods often produce inconsistent results when the target warp requires significant bending. To analyze the cause of it, below we first elaborate on the mathematical formulation of TPS.

   In general, given the n pairs of source and target control points, an interpolation function may be constructed by minimizing the data error given by the sum of squared difference as follows.

$$\tau[\omega] = \sum_{j=1}^{n} \|\omega(x_j, y_j) - (u_j, v_j)\|_2^2, \tag{1.1}$$

where $\{(x_j, y_j)\}$, $\{(u_j, v_j)\}$, $j = 1, \cdots, n$ represents two different sets of data points. TPS transform imposes some quality criteria, such as smoothness, that the interpolation function should satisfy. Following the thin plate spline transform $\tau[\omega]$ is considered as an energy functional defined as

$$\tau[\omega] = P(\omega) + S(\omega), \tag{1.2}$$

where

$$P(\omega) = \sum_{j=1}^{n} \|\omega(x_i, y_i) - (u_i, v_i)\|_2^2,$$

$$S(\omega) = \lambda \iint_{\mathbb{R}^2} [\omega_{xx}^2 + 2\omega_{xy}^2 + \omega_{yy}^2] dx\, dy.$$

Here $P(\omega)$, as before, is the penalty function due to deviation of data points. $S(\omega)$ is the stabilizing functional, called the second-order Sobolev semi-norm, whose minimization ensures smoothness of the function ω. λ is the regularization parameter ($\in [0, 1]$), controlling the relative importance between a close fit (first term) and the smoothness (second term) of the interpolation function.

A closed-form solution of Eq. (1.2) as proposed in Wahba (1990) is given by

$$(u, v) = \omega(x, y) = \mathbf{a}_0 + \mathbf{a}_1 x + \mathbf{a}_2 y + \sum_{j=1}^{n} \mathbf{c}_j \nu(\|(x, y) - (x_j, y_j)\|_2), \tag{1.3}$$

where $\mathbf{a}_0$, $\mathbf{a}_1$, $\mathbf{a}_2$, $\{\mathbf{c}_j : j = 1, 2, \cdots, n\}$ are vector parameters with dimension equal to the dimension of the control points, which is 2 in our case. The radial basis kernel used in TPS is $\nu(p)$ = (p^2 *ln p*). Note that, we use the bold notation for representing vectors. Rewriting the Equation. 1.3 with scalar parameters instead of the vector we get the following two equations (since the dimension of control points is 2 in our case),

$$u = a_0^x + a_1^x x + a_2^x y + \sum_{j=1}^{n} c_j^x \nu(\|(x, y) - (x_j, y_j)\|_2), \tag{1.4}$$

$$v = a_0^y + a_1^y x + a_2^y y + \sum_{j=1}^{n} c_j^y \nu(\|(x, y) - (x_j, y_j)\|_2), \tag{1.5}$$

where $\mathbf{a}_0 = (a_0^x, a_0^y)$, $\mathbf{a}_1 = (a_1^x, a_1^y)$, $\mathbf{a}_2 = (a_2^x, a_2^y)$, and, $\mathbf{c}_j = (c_j^x, c_j^y)$.

The formulation of TPS as presented in Eq. (1.2) shows that it includes an interpolation term and a second-order smoothness term which is responsible for estimating a smooth transformation from the source to the target based on the given or the estimated control point correspondences. The energy function of TPS is formulated in analogy with the bending of a physical steel plate; and most importantly, this modeling of the bending energy of a thin metal plate holds for only small displacements of the coordinates in the plane. Because for small displacements of the in-plane points, the transformation can be smooth; whereas larger displacements indicate sharp changes which are in contrast with the second-order smoothness criterion. Therefore, TPS-based warping methods often produce inconsistent results when the target warp requires significant bending, for example, consider the instances when the arm of the model or person is bent. Experimental evidence of this can be seen in the results presented in Fig. 1.2.

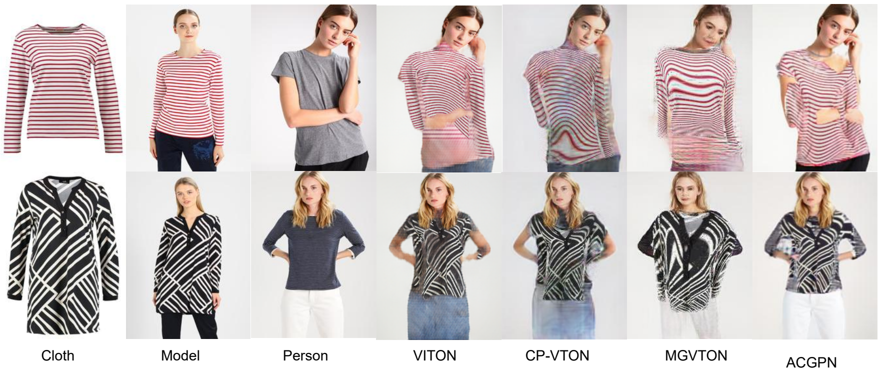


Figure 1.2: Demonstration of limitations of warping methods involving TPS transform.

4. The formulation of TPS models some rearrangements of control points in the image plane. Therefore it is not intended for applications requiring modeling of overlap or folds among the different semantic parts of the clothing Bookstein (1989). This can be verified from the results of the methods employing TPS shown in Fig. 1.2. Notice that when the arm falls over the clothing the warping approaches of the portrayed methods fail.

Based on the above discussion this work makes the following contributions,

1. Our research is focused on learning a VTON system from easily available data i.e., we propose model-to-person (M2P) try-on approaches that compete well against the cloth-to-person (C2P) methods which require both the clothing as well as the model wearing the clothing for the learning the VTON system. Thus we address a more practical problem, considering the fact of the rare availability of separate clothes images.

2. A majority of previous methods have proposed feature learning based image matching approaches for clothes warping. However, these methods do not consider any external human structural constraints which is important for ensuring human anatomy consistent prediction of the target warp. To this end, we come up with the idea of employing structural key points of humans and clothing as control points and the correspondences among them from the model's image to the person's image in predicting the target warp. Experimental evidence establishes the effectiveness of this idea in reducing the computation overhead in terms of a number of parameters while achieving an improvement in performance over the existing methods.

3. The problem of VTON is challenging in terms of data availability and problem complexity which puts a heavy burden on a single system to learn. Due to this, it is usually solved in multiple stages. In our work, we used human landmarks as a representation of body shape and pose. But the structural key points of humans are only a few and also not very accurate estimators of the body shape of a human. Hence our landmark-based warping approach produces some artifacts near the edges of the warped clothing which we call a “warping glitch”. To address this, we propose a *Mask Generator Module (MGM)* that predicts the target region of the person image which is expected to be covered by the model’s clothing after try-on. Our final module which is the *Image Synthesizer Module (ISM)* combines the results of its previous two modules with the reference person to generate the final try-on output.

4. Our previous work produces photo-realistic results but it has two limitations - (i) it employs both human and fashion landmarks as control points. However, annotations are expensive, especially, fashion landmarks, (ii) using correspondences between all the landmarks related to the source image and the target image together to compute the TPS transform fails to accurately compute the target warp when the reference person or model’s arm overlaps on the torso. To address this limitation our next work adopts the idea of geometric matching introduced by CP-VTON in the context of VTON. However, we use the densepose (Alp Güler et al., 2018) correspondences between the model and the person to learn the parameters of the TPS warping function in order to compute the target warp. A densepose representation maps all human pixels of an RGB image, to the 3D surface of the human body, thus providing a precise estimate of the human body shape under the clothing. Employing densepose correspondences aids the geometric matching network to learn the features related to the human body geometry well and thus shows improvement over other methods.

5. Our previous works employ TPS transform-based warping method to compute the target warp. However, most of the human arm movements causing the folding of sleeves can not be modeled by TPS transform. This is due to the second-order smoothness criterion of TPS which restricts the bending of the mesh. To address this, we propose a hand-crafted feature-based warping technique with constraints specific to our goal.

6. Image-based VTON approaches can not handle the cases of overlap among different semantic parts of the clothing. This is because achieving a realistic warping of the source clothing requires consideration of the depth dimension. However, we are working on image-based VTON approaches, which disregard the depth dimension. To address this we propose a part-by-part warping technique; where we divide the source clothing into parts i.e., sleeves and torso, followed by warping each of these parts separately, before finally combining them. This approach attempts to solve the overlap issue.

7. We work on the model to person VTON approaches where we consider the sole source of the

reference clothing to be tried on is the model's image wearing that clothing. For try-on, we extract the clothing segment from the model's image and compute the target warp. However, the target warp is computed from the non-occluded clothing areas only; therefore, the occluded areas of the source need to be interpolated if exposed in the target image. To address this we propose a *mask prediction network (MPN)* that predicts the target clothing mask, referring to the region of the expected try-on output containing the model's clothing. MPN learns to incorporate the correlation between the necessary features of the model and the person images, which empirically shows improvement over the semantic generation module of a previous benchmark method with a similar objective. The output of MPN aids in distinguishing the occluded areas of the target clothing. The proposed *image synthesizer network (ISN)* interpolates these occluded regions and also produces a seamless try-on image. MPN also adds to the faster computation of the target warp.

8. Our proposed methods are based on self-supervised training strategy, that removes the requirement of paired data (model and person wearing the same clothing) for training, which is often difficult to get.

1.4 Organization of the Thesis and Contributions

This thesis proposes methods for virtual try-on where the objective is to predict the image of a reference person wearing the clothing of a model. Apart from the Introduction and the Conclusion chapters, this thesis contains three contributory chapters. Their brief contents and contributions are given in the following subsections. A brief road map of this thesis is illustrated in Fig. 1.3.

1.4.1 Contributions of Chapter 2

In this chapter, we explore the efficacy of landmark guidance in the context of virtual try-on. Landmarks corresponding to the human body are the anatomical key points of humans, known as human landmarks. The functional key points of clothes e.g., the corners of the neckline, hemline, cuff, etc., are known as fashion landmarks. We attempt to use this structural information in warping the desired clothing. We employ a pose-guided warping module that predicts the fashion landmarks of the target warp and thereafter computes the target warp based on human and fashion landmark correspondences between the model and the person images. Using fashion landmarks enable our method of warping to work based on the geometric structure of the cloth; thus independent of the texture of the clothes, which otherwise confuses the warping process especially in the case of complex patterns i.e., stripes, checks, florals, etc., with unnecessary extra information. We further refine the fit of the predicted warp using our mask generator module (MGM) and thereafter predict the desired try-on image in the subsequent stage with the image synthesizer module (ISM). To tackle the problem of the lack of paired training data, we resort to a self-supervised training strategy. Here paired data refers to the

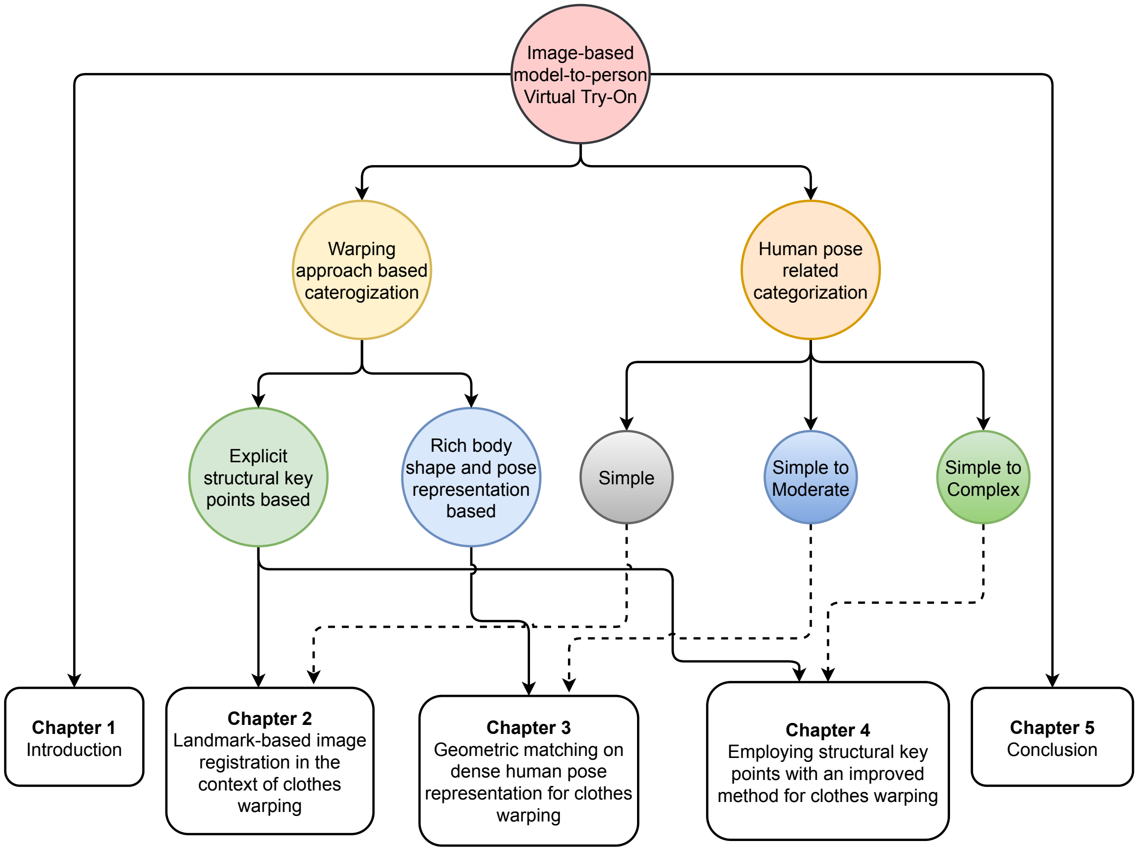


Image-based model-to-person Virtual Try-On
Warping approach based caterogization
Human pose related categorization
Explicit structural key points based
Rich body shape and pose representation based
Simple
Simple to Moderate
Simple to Complex
Chapter 1 Introduction
Chapter 2 Landmark-based image registration in the context of clothes warping
Chapter 3 Geometric matching on dense human pose representation for clothes warping
Chapter 4 Employing structural key points with an improved method for clothes warping
Chapter 5 Conclusion

Figure 1.3: Roadmap of this thesis.

image pair of model and person wearing the same cloth. This work also relaxes the constraint of the availability of separate clothing images imposed by many existing methods. This method transfers the clothing from a given person (say, model) image to the target person.

Related Publication: Roy, Debapriya, Sanchayan Santra, and Bhabatosh Chanda. ”LGVTON: a landmark guided approach for model to person virtual try-on.” Multimedia Tools and Applications (2022): 1-37.

1.4.2 Contributions of Chapter 3

In the previous chapter we have dealt with explicitly specified features i.e., landmarks, and the correspondences between the landmarks of the model image and the target image are used to compute the TPS transformation in order to predict the target warp. Moreover, employing correspondences of all human and clothing landmarks fail to give satisfactory result in case of folded or cross-armed postures. In fact, our previous method is restricted to only simple human poses. Besides, annotations are costly, and, our previous work requires both human and fashion landmark annotation. Inspired by the capability of deep neural networks in learning task-specific features, in this work we intend to learn the features from the model and person images and establish the correspondences among the learned features to compute the target warp employing a TPS transformation. Learning features from images directly has been adopted by many methods previously. However, this thesis is based on our primary observation that the transformation required on the clothing for transferring it from the model to the person can be estimated using the body shape and pose changes from the model to the person. In order to instill this idea into our feature learning and matching process, we provide the dense human pose representations of the model and the person as input instead of providing the model and the person images directly. The objective is to learn the required transformation based on the change in the densepose representation from the model to the person. Rigorous experiments establish the significance of our method compared to others.

Related Publication: Roy, Debapriya, Sanchayan Santra, and Bhabatosh Chanda. ”Incorporating Human Body Shape Guidance for Cloth Warping in Model to Person Virtual Try-on Problems.” 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ). IEEE, 2020.

1.4.3 Contributions of Chapter 4

In the previous two chapters, we have dealt with both explicitly specified features e.g., landmarks and learned features. This is followed by predicting the TPS transform based on the computed feature matching. However, employing the TPS transform as the warping function in the current problem context has some limitations due to the formulation of this transform. The energy function of TPS has two parts, a penalty function to measure the discrepancy between the predicted and the supplied control points, and a stabilizing functional, whose minimization ensures smoothness of the function. The TPS formulation imposes two limitations on its applicability in the warping of clothes in the VTON problem - (1) the formulation of TPS models some rearrangements of control points in the reference grid. This can not model the overlap among different parts of the clothing which mainly occurs due to folding or bending of sleeves. (2) the second-order smoothness criterion imposed by the stabilizing functional of TPS restricts in-plane deformation of the grid that may occur due to the flexibility of human body part movements. To this end, in this chapter, we attempt a solution approach where the clothing from the source person is segmented into semantically meaningful parts and each part is warped independently to the shape of the person. This idea of part-based warping solved the overlap issue. To address the bending issue, we propose a hand-crafted feature based warping method that is inspired from the idea of field warping (Beier and Neely, 1992). Besides, we propose two learning-based modules: a synthesizer network and a mask prediction network for refining the fit of the predicted warp and predicting the final try-on output. Experimentally we show that this method achieves state-of-the-art performance producing a photo-realistic, robust VTON solution without requiring any paired training data.

Related Publication: Roy, Debapriya, Sanchayan Santra, Diganta Mukherjee, and Bhabatosh Chanda. ”Achieving pose robustness in the Context of Virtual Try-On.” ACM Transactions on Multimedia Computing, Communications, and Applications (Under review).

1.4.4 Contributions of Chapter 5

In this chapter, we summarize the conclusions made by the previous chapters and discuss some open problems and future directions of research on the topic.

1.5 Discussion

In this thesis, we have considered a computer vision problem relevant to e-commerce or online shopping. We have presented a related literature survey and tried to identify some limitations of the existing methods. In the following chapters, we have proposed some novel solutions to address these problems and have presented their results. Finally, in the last chapter, we summarized our findings and indicated some directions for further research.

Chapter 2

Landmark-based image registration in the context of clothes warping

2.1 Introduction

The VTON problem can be addressed in various settings. Recent VTON methods (Jetchev and Bergmann, 2017; Han et al., 2018; Wang et al., 2018a; Hsieh et al., 2019a; Zheng et al., 2019a; Hsieh et al., 2019b; Sun et al., 2019; Yu et al., 2019; Dong et al., 2019a,b; Yang et al., 2020; Song et al., 2019) require the image of the clothing, in addition to the image of the model (source) wearing that clothing for the training purposes. This problem is called cloth-to-person (C2P). Having a separate clothing image makes the modeling of VTON systems easier mainly because the source clothing images are all taken in standard anatomical positions, hence, no pose variability or occlusion. But such images are rarely available on e-commerce websites or social media; which are the major data sources for this problem domain. Keeping this in mind this chapter proposes a model-to-person (M2P) VTON problem that does not require any separate clothing image. That means only an image of a model wearing the reference clothing is given along with the image of a target person who wants to try that clothing. Compared to the C2P variant, M2P problems are more challenging mostly due to the variability in the pose of the source clothing.

In general most of the M2P VTON methods follow a two-staged approach (Han et al., 2018; Wang et al., 2018a; Yu et al., 2019; Dong et al., 2019a). In the first stage, the clothes are aligned according to the target person's body shape and pose, and, in the second stage, the warped clothes are combined with the target person's image. The use of Thin-Plate Spline (TPS) transformation (Sprengel et al., 1996) is popular among VTON methods to warp the source clothes to the target shape. The parameters of this transformation are mostly computed by a CNN-based feature matching method called *Geometric Matching* (GM) (Wang et al., 2018a; Yu et al., 2019; Dong et al., 2019a). While the concept of GM was first proposed by Rocco et al. (Rocco et al., 2017), CP-VTON (Wang et al., 2018a) showed its applicability in warping clothes for virtual try-on. Although GM is a popular warping strategy, this method results in inaccurate warping when complex patterns (Jae Lee et al., 2019; Issenhuth et al., 2019; Yang et al., 2020), for instance, stripes and floral are present in the clothing. A larger range of transformations is possible to be modeled by TPS, due to which in absence of necessary constraints the network fails to realize the geometry of the human body by itself. The human body undergoes very restricted movements because of its structural constraints due to the organization of bones and muscles. Therefore warping the clothing according to the pose and shape of the human body needs the consideration of the human structural constraints, without which the transformation might produce some incorrect deformation. GMN as proposed by CP-VTON learns features from the human and the clothing images, and, correlate them to predict the parameter θ of TPS. This network is trained in a supervised way. However, since structural constraints of the human body are not explicitly imposed, the method produces inaccurate deformation in the source clothing. To address this problem, this chapter aims to use landmarks correspondence between the model and the target person images. The motivation is to impose explicitly the structural constraints of both clothing and the human body to predict the warping. The correspondence between the mask of the source clothing and that of the predicted mask has been done in VITON (Han et al., 2018), a hand-crafted feature-based traditional warping method is employed to find a corresponding set of control points between the two. This method uses a coarse-to-fine approach that refines the texture details of the clothing of the predicted warp. However, the results often lose texture details and the computation of shape context descriptor is relatively slow.

The idea of correspondences between the body shape and pose of the model and that of the person was previously used by M2E-TON (Wu et al., 2019). However, being a coarse-to-fine approach it fails in preserving the texture and color of the clothing in the output. Also since it establishes only the correspondences between the human body-related details only, it cannot predict the pixel values in the areas where clothing contour falls outside the body contour. In addition, this method has some other small issues too. For example, it fails to preserve proper background details of the person's image; second, if the source clothing contains any head patterns it predicts inaccurate warps.

In view of the above limitations, in this chapter, we propose an M2P VTON solution. It introduces a warping method that establishes the correspondences between the structural key points of humans (both the model and the target person) and also of the clothing for computation of the target warp. In addition, we propose two other learnable modules that finally synthesize the final try-on output.

2.2 Proposed Approach

We propose a Landmark-Guided Approach for Virtual Try-On (LGVTON) that learns to synthesize an image of a person wearing a model's clothing. Formally, given a model image M wearing the clothing c and a person image P, LGVTON synthesizes P', which is the new image of the person wearing the model's clothes. Here $M, P, P', c \in \mathbb{R}^{(Wd \times Ht \times Ch)}$, where Ht, Wd denotes the height and width of the images and Ch denotes the number of channels. We took, Wd = 192, Ht = 256 and Ch = 3 (for RGB image). The workflow of LGVTON is three-fold (refer to Fig. 2.1). First, it attempts to warp the model's clothes according to the shape and pose of the target person. This is done by our *Pose Guided Warping Module (PGWM)*. Second, a segmentation mask corresponding to the clothing area of the person wearing the target clothing is predicted by our mask generator module (MGM) which aids in improving the fit of the predicted clothing. Third, the image synthesizer module (ISM) synthesizes the final virtual try-on output.

The VTON problem being a challenging task is not always sufficient to be achieved by using only the model's and the person's images. Instead, the outputs of some auxiliary approaches are taken as input in the different stages of most of the existing VTON methods Wang et al. (2018a); Han et al. (2018); Yang et al. (2020); Dong et al. (2019a); Song et al. (2019). In this work also, we use some of these auxiliary methods. For instance, to compute the target clothing we warp the source clothing c extracted from M using a human parsing approach Gong et al. (2017). Our PGWM uses pose estimates of M and P in order to compute c' from c. Here to estimate the pose we employ the human pose estimation approach of Cao et al. (Cao et al., 2017) which given an image of a person predicts the location of his landmarks, i.e., the locations of different human bone joints. We also use the fashion landmarks of the clothing of the model c as input in our PGWM. Moreover, to estimate the human body shape under clothing we use the densepose representation Alp Güler et al. (2018) of the person P in the proposed ISM.

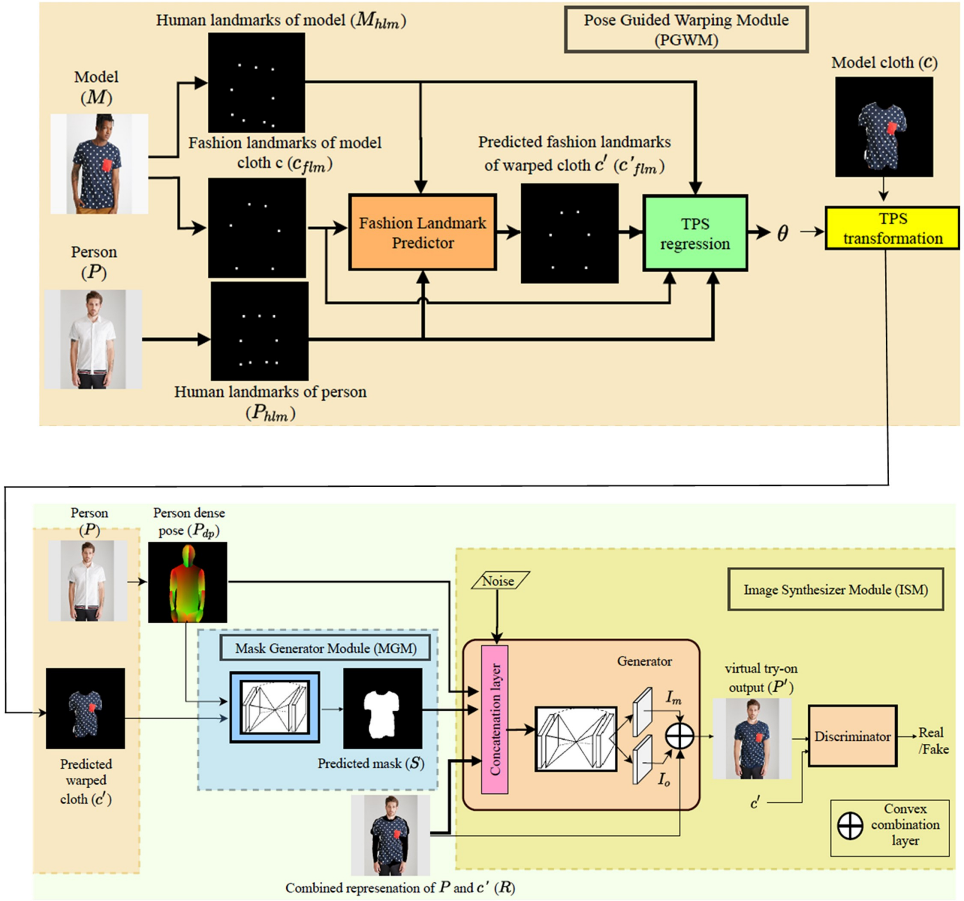

Human landmarks of model (M_{hlm})
Pose Guided Warping Module (PGWM)
Model (M)
Model cloth (c)
Fashion landmarks of model cloth c (c_{flm})
Predicted fashion landmarks of warped cloth c' (c'_{flm})
Person (P)
Fashion Landmark Predictor
TPS regression
θ
TPS transformation
Human landmarks of person (P_{hlm})
Person (P)
Person dense pose (P_{dp})
Noise
Image Synthesizer Module (ISM)
Mask Generator Module (MGM)
Concatenation layer
Generator
I_m
I_o
virtual try-on output (P')
Discriminator
Real /Fake
Predicted warped cloth (c')
Predicted mask (S)
c'
Convex combination layer
Combined represenation of P and c' (R)

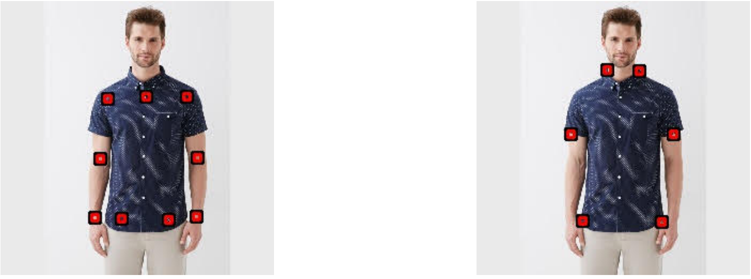
Figure 2.1: Block diagram depicting workflow of LGVTON.

2.2.1 Computing the target warp of the model clothing

This section elaborates on our Pose Guided Warping Module (PGWM). It computes a TPS transform (Duchon, 1977) $\omega(\cdot)$ to warp c into $c' \in \mathbb{R}^{(Wd \times Ht \times Ch)}$, where c' is aligned with the body shape and pose of the person P. c is obtained from M using the human parsing network proposed in (Gong et al., 2017). Our method of predicting the target warp is inspired by the concept of landmark-based image registration (Maintz and Viergever, 1998), which uses landmark (also referred to as control points) correspondences between two images to estimate the parameters of the transformation from one image to the other. In general, the idea of image registration is to geometrically align two or more images of generally the same object or scene taken from various imaging devices at different times and angles. In general, this is achieved by establishing correspondences between a set of points of one image with that of the other. An object can be defined as an infinite set of points. Now establishing correspondences between two infinite sets is not feasible. Hence, instead, an image containing an object is defined as a set of a finite number of points sampled from a large set of points defining the image. Such points are not any random selection of points but generally a set of landmark points. A landmark is a point of correspondence that matches withing objects belonging to the same category. In general, landmarks define the shape of objects, hence used to compare the shape of objects. Since landmarks define the structural information, therefore using these points allows the transformation to be interpreted in terms of the underlying anatomy.

In our case, we consider fashion and human landmarks (Fig. 2.2) as control points. PGWM predicts the fashion landmarks of the target warp, given that of the clothing of the model along with the anatomic key points called human landmarks of the model and the person. Landmarks here refer to some well-defined key points which allow a geometric shape to be characterized in a manner that corresponds across subjects. Therefore, considering fashion landmarks enable our method of warping to work based on the geometric structure of the clothing; thus independent of the texture of the clothes, which otherwise confuses the warping process especially in case of complex patterns i.e., stripes, checks, florals, etc., with unnecessary extra information.

Once the parameters of the transformation are computed, the target warp c' is computed by applying the corresponding TPS transformation on c. Therefore, the generation of c' involves two steps discussed below.



(a) Human landmarks

(b) Fashion landmarks

Figure 2.2: Demonstration of two types of landmarks used by this work.

2.2.1.1 Estimating the source and the target sets of landmarks

We extract human landmarks M_{hlm} and P_{hlm} corresponding to model M and person P, respectively, using the method proposed by Cao et al. (Cao et al., 2017). We denote the multiset of human landmarks of the person P by $P_{hlm} = \{\alpha_1^P, \alpha_2^P, \dots, \alpha_{n_h}^P\}$ and that of the model M by $M_{hlm} = \{\alpha_1^M, \alpha_2^M, \dots, \alpha_{n_h}^M\}$, where, $\alpha_i^M, \alpha_i^P \in \mathbb{N}^d$, d is the considered dimension and in our case $d = 2$. n_h is the total number of landmarks of a human used in this work and in this work, $n_h = 9$. These landmarks are anatomical key points of human which are generally used to represent the pose of a person. However, as Bogo et al. (Bogo et al., 2016b) suggested, these landmarks can also approximate the body shape of a person. Hence, the correspondence between the human landmarks of the model and the person can represent the structural change of c to c'.

However, computing the deformation of a non-rigid object such as clothes is very challenging since it includes the deformation of designs or patterns present in the clothes. Achieving a highly accurate warping, which is very essential for realistic VTON output, usually requires many landmarks. Consequently, along with human landmarks, we consider fashion landmarks as well so that the warping becomes more precise. We use the annotated ground-truth for fashion landmarks of c. However, that of c' is not available. Hence, to compute the corresponding landmarks of c' we propose a fashion landmark predictor network $\mathscr{F}$ (Fig. 2.3) that predicts c'_{flm} given M_{hlm}, P_{hlm} and c_{flm}. We denote the multiset of fashion landmarks of the clothes of the model by $c_{flm} = \{\beta_1^c, \beta_2^c, \dots, \beta_{n_f}^c\}$ and that of the target warp of the model clothes by $c'_{flm} = \{\beta_1^{c'}, \beta_2^{c'}, \dots, \beta_{n_f}^{c'}\}$, where n_f is the number of fashion landmarks of a clothing used in this work, and in practice $n_f = 6$. This network is trained using L2 loss.

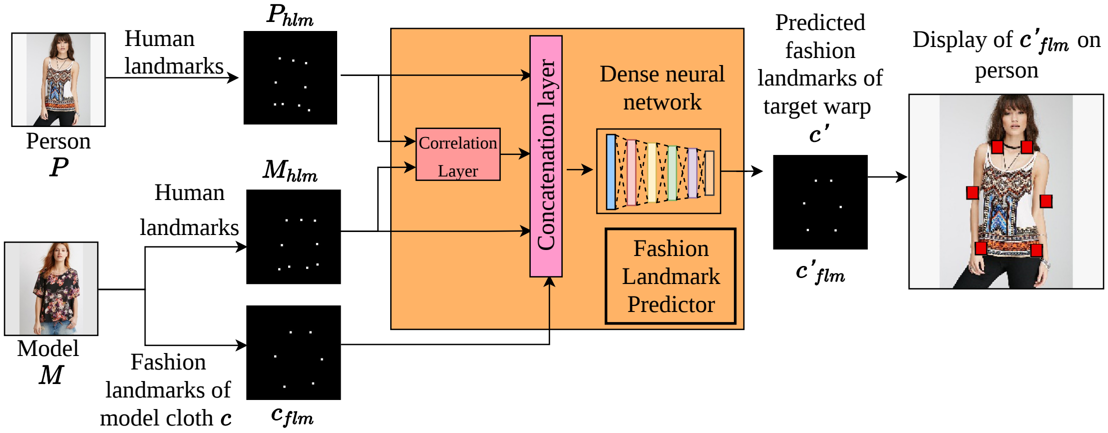

Person
P
Human landmarks
P_{hlm}
Model
M
Human landmarks
M_{hlm}
Fashion landmarks of model cloth c
c_{flm}
Correlation Layer
Concatenation layer
Dense neural network
Fashion Landmark Predictor
Predicted fashion landmarks of target warp c'
c'_{flm}
Display of c'_{flm} on person

Figure 2.3: Block diagram of fashion landmark predictor network ($\mathscr{F}$) in Pose guided warping module (PGWM).

Here, we are transferring the clothes from model to person, hence, it has to undergo deformation according to the way the body shape and pose change from model to person. Therefore, we model the warping of the source clothes as a function of the correlation between the human landmarks of the model and the person, and fashion landmarks of the source clothes. We employ a correlation layer in $\mathscr{F}$ in order to model this change by the correlation between the human landmarks of the model and the person i.e., M_{hlm} and P_{hlm}. Correlation is a statistical technique that can show whether and how strongly pairs of variables are related and this is measured by the linear relationship between a pair of variables. Coming into more detail, given the data $f_a, f_b \in \mathbb{R}^{1\times l}$, the correlation layer produces the correlation map $C_{ab} \in \mathbb{R}^{l\times l}$, where,

$$C_{ab} = {f_a}^T f_b. \tag{2.1}$$

Therefore a correlation map contains the pairwise similarity of all the values of its inputs. The objective of putting a correlation layer here is to leverage the similarities between the locations of the human landmarks of the model and the person and use that information to predict the matching between the fashion landmarks of the model and the person. Intuitively, this matching will aid in predicting the fashion landmarks of the person. The motivation here is that changes in the location of the human landmarks from the model to the person can be a good estimator to measure the change in terms of the locations of the fashion landmarks. Now, how the correlation layer measures this similarity needs to be clarified. In the classical method of estimating correspondence between two images, there are three steps in general, feature extraction, feature matching, and, robust estimation of global geometric transformation. Here, in our case we already have the features extracted i.e., we have the locations of the landmarks. In classical methods, matching is done by computing similarities between all pairs of descriptors between two images. Note that, as we have already discussed, the correlation layer does the same thing. Now unlike classical geometry estimation where this matching is used to estimate the transformation from the source to the target, here we are interested in leveraging the matching between the human landmarks to predict the fashion landmarks. However, the idea of leveraging this matching for the current objective is by intuition, and inside the network how this information is processed can not be explicitly explained as we know analyzing the weights learned by the neural network is difficult.

Training $\mathscr{F}$ is tricky since paired data, i.e., human and fashion landmarks of two persons wearing the same clothes in any arbitrary pose, is not available. We discuss more on this in Section 2.3.

2.2.1.2 Computing the target warp

Once we have the source landmarks M_{hlm}, c_{flm}, and the target landmarks P_{hlm}, ${c'}_{flm}$, we now define $\mathscr{R} = \{\alpha_1^M, \alpha_2^M, \ldots, \alpha_{n_h}^M, \beta_1^c, \beta_2^c, \ldots, \beta_{n_f}^c\}$ and $\mathscr{T} = \{\alpha_1^P, \alpha_2^P, \ldots, \alpha_{n_h}^P, \beta_1^{c'}, \beta_2^{c'}, \ldots, \beta_{n_f}^{c'}\}$ (refer to the description of these notations given at beginning of Section 2.2). Our idea is to utilize the correspondences between the source $\mathscr{R}$ and the target landmarks $\mathscr{T}$ to transform c to c'. Hence, we find a smooth interpolation function $\omega : \mathbb{N}^2 \to \mathbb{N}^2$, such that, $\omega(\alpha_i^M) = \alpha_i^P$; $i = 1, 2, \ldots, n_h$; and $\omega(\beta_i^c) = \beta_i^{c'}$; $i = 1, 2, \ldots, n_f$; hold. Then we employ $\omega(\cdot)$ to transform c to c'; which means basically $\omega(\cdot)$ is applied on the mesh grid containing c to get c'. For notational convenience, we use $\mathbf{r}_j$ and $\mathbf{t}_j$ to denote the j^{th} element of $\mathscr{R}$ and $\mathscr{T}$ respectively, where $j = 1, 2, \ldots, N$. Here N denotes the total number of elements in each of the multiset $\mathscr{R}$ and $\mathscr{T}$. In practice, $N = 15$.

Now, the deformation of non-rigid objects such as clothes should be smooth. Considering this, we choose $\omega(\cdot)$ to be a thin-plate spline (TPS) transform (Duchon, 1977), which is a widely used transform representing coordinate mappings with a penalty term for imposing smoothness [1]. As we are dealing with only estimates of the true landmark locations which may be noisy, so instead of exact interpolation in this step, we remain content with approximation. This is accomplished in the TPS transform by minimizing the following objective function (Sprengel et al., 1996; Donato and Belongie, 2002),

$$H[\omega] = \sum_{j=1}^{N} \|\omega(\mathbf{r}_j) - \mathbf{t}_j\|_2^2 + \lambda \iint_{\mathbb{R}^2} [\omega_{xx}^2 + 2\omega_{xy}^2 + \omega_{yy}^2] dx\, dy, \tag{2.2}$$

where λ is a regularization parameter, which determines the relative weight between the approximation behavior and the smoothness of the transformation. It is a positive scalar. ω_{xx}, ω_{xy}, ω_{yy} are second-order gradients of $\omega(\cdot)$ as we consider each landmark as a 2-tuple.

Note that, this work is targeted to upper body clothes only, so for warping, we use only upper body human landmarks which in total is 9, and fashion landmarks corresponding to upper body cloths which in total is 6 (refer to Fig. 2.2). Therefore, we have $N = 15$ landmarks, i.e., $n_h = 9$ and $n_f = 6$. We select $\lambda = 0.01$, experimentally for satisfactory results. Some target warps generated by PGWM are shown in Fig. 2.4. A closed-form solution of (2.2) as proposed by Wahba (1990) is given by

[1] a detailed study on TPS is given in the Appendix A

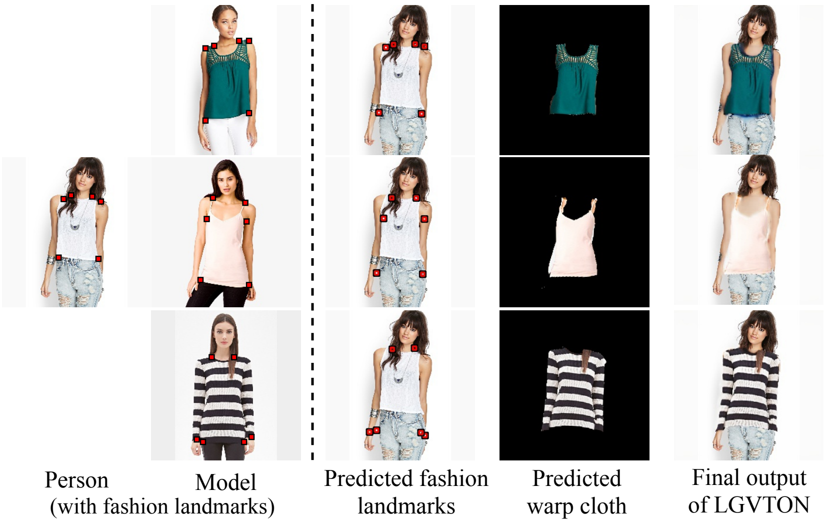


Figure 2.4: Results of the two steps of PGWM and final output of LGVTON.

$$(u,v) = \omega(x,y) = \mathbf{a}_0 + \mathbf{a}_1 x + \mathbf{a}_2 y + \sum_{j=1}^{N} \mathbf{c}_j \nu(\|(x,y) - (x_j,y_j)\|_2), \tag{2.3}$$

where $\mathbf{a}_0$, $\mathbf{a}_1$, $\mathbf{a}_2$, $\{\mathbf{c}_j : j = 1,2,\cdots,n\}$ are vector parameters with dimension equal to the dimension of the control points, which is 2 in our case. The radial basis kernel used in TPS is $\nu(p) = (p^2\ ln\ p)$.

Our TPS regression part of the PGWM as shown in Fig. 2.1 estimates the values of the parameter θ from the source and target sets of control points. Here θ is a vector with the values of $\mathbf{c}_1,\cdots,\mathbf{c}_N,\mathbf{a}_0,\mathbf{a}_1,\mathbf{a}_2$. Hence we have in total $N+3$ number of vector parameters. In this work, we used 6 fashion landmarks and 9 human landmarks. Therefore here $N = 9 + 6 = 15$. Therefore we have total $N+3 = 18$ vector parameters. Now as each parameter is a 2 dimensional vector hence, we have $18 \times 2 = 36$ scalar parameters. On the other hand, we have 15 control points or landmarks, each point has x and y coordinate values hence a total of 30 (15×2) values. Therefore, we have 36 parameters to estimate with 30 values, which is an under-conditioned situation. Hence, a constrained least square estimation approach is used to solve.

2.2.2 Generating Segmentation Mask of Target Clothing

Human landmarks are a sufficient representation of pose and a good approximator of body shape (Bogo et al., 2016b) but not a highly accurate estimator (Shigeki et al., 2018). That means they result in a good overall warping but it might not be very precise near the edges of the cloth, as illustrated in

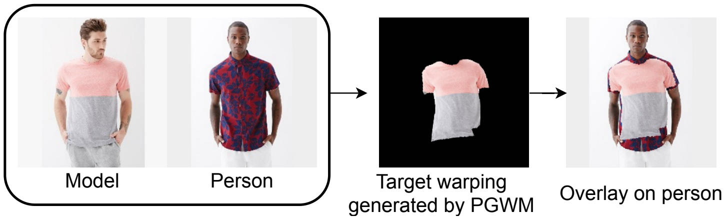


Figure 2.5: Demonstration of warping glitch. The target warping generated by the PGWM is not completely accurate as it is observable when overlaid on the person's image. Although the overall fit is good the fit is not precise near the areas of the collar and sleeves.

Fig. 2.5. We address this problem as a warping glitch. This is addressed by our Mask Generator Module (MGM). MGM predicts the segmentation mask corresponding to the region of the clothes of the target person after wearing the model clothing. This in turn guides the next module (ISM) to handle the warping glitches.

Formally, given c' and densepose (Alp Güler et al., 2018) $P_{dp} \in \mathbb{R}^{(Wd \times Ht \times 3)}$ of the person (refer to Fig. 2.1), MGM generates the target clothing segmentation mask S corresponding to c' (ideally in cases of no warping glitch, mask of c' should be same as S). The first channel in the densepose representation contains the 24-part labels of humans and the other two channels contain the UV parametric values of the body surface corresponding to each of the parts. Similar to the previous module, paired data is also not available for training in this module. However, we tackle this using a landmark perturbation-based approach, discussed in Section 2.3. An ablation study related to this module is conducted in Sec. 2.4.5.

2.2.3 Synthesizing VTON image

Our final module, the *image synthesizer module* (ISM) generate the final virtual try-on output P' from the inputs c' and P.

This module takes in the following inputs: (i) The generated target clothes mask S from MGM, (ii) a combined representation R of P and c', Densepose P_{dp} of P. Note that instead of providing P and c' separately, we provide a combined representation R as input as it enhances the perceptual quality of the output. This is discussed more in detail later in this section. This representation is obtained by setting pixel values in the upper body area including the clothing region of P to zero (using the human parsing of P) and then combining c' with it. Such a representation can be viewed in the block diagram given in Fig. 2.1. This module is implemented as a conditional generative adversarial network (cGAN). Therefore, an additional input which is the random noise $\mathbf{z}$ sampled from a noise distribution $p_{\mathbf{z}}$ is provided.

ISM when trained without S generates image artifacts. This is due to the great variety in the types of designs of the clothes, which confuses the cGAN to distinguish a warping glitch from a clothing design. Therefore, the network can not identify a warping glitch itself. Whereas, giving S as an input makes the network identify the regions of warping glitches. For a more detailed discussion on this please see supplementary material.

The objective of our cGAN may be expressed as

$$L_{\text{cGAN}}(G,D) = \mathbb{E}_{\mathbf{x}\sim p_{data}}[\log D(\mathbf{x}|\mathbf{y})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}[\log(1\text{–}D(G(\mathbf{z}|\mathbf{y})))], \tag{2.4}$$

where the generator G learns a distribution p_g over data $\mathbf{x}$. It builds a mapping function from a prior noise distribution $p_{\mathbf{z}}$ with conditional information $\mathbf{y}$ to the data space. While the discriminator D represents the probability of $\mathbf{x}$ given $\mathbf{y}$, to be coming from the training data rather than the generator distribution p_g (Mirza and Osindero, 2014). cGAN is trained to minimize an objective L_{cGAN} against an adversarial D that tries to maximize it. The optimum G denoted by G^* is

$$G^* = \ \text{argmin}_G \ \text{argmax}_D \ L_{\text{cGAN}}(G,D). \tag{2.5}$$

In our model, c' is the condition given to both G and D.

G contains an hourglass network (Newell et al., 2016) (a convolutional neural network with skip connections [2]), followed by two parallel convolution layers, giving activation I_o, an intermediate VTON output, and I_m, a mask, which helps to retain the necessary details from R. The last layer of G is a convex combination layer that combines I_o and R using I_m. Generally, the averaging tendency of convolution operation causes the loss of fine details in the output. The network tackles this with the help of I_m as it aids in preserving necessary details from R in the final output. This can be validated from our results presented in Sec. 2.4 which shows that our result retains cloth, person, and background details better in comparison to the results of the other methods. The discriminator D is a patchGAN discriminator (Isola et al., 2017).

It has been experimentally observed by different past works on conditional GAN (Johnson et al.,

[2] more detailed in the Sec. A.2.2 of Appendix A

2016),(Xian et al., 2018) that having other loss functions, such as perceptual loss (Xian et al., 2018), along with adversarial loss gives a better output. Therefore, we incorporate structural dissimilarity index (DSSIM) and VGG perceptual loss (Johnson et al., 2016) as additional loss functions in the generator. The inclusion of these additional loss terms keeps the task of the discriminator unchanged while the generator, in addition to the task of fooling the discriminator, has to generate data instances closer (in L2 sense) to the ground-truth.

SSIM (Structural similarity index) (Wang et al., 2004) is an image metric that measures the structural similarity between two images. However, in the neural network, the objective is to minimize the value of the loss function, so instead of SSIM we take DSSIM that is related to SSIM the following way, $\text{DSSIM}(\cdot,\cdot) = \frac{(1-\text{SSIM})}{2}(\cdot,\cdot)$. DSSIM loss for the generator is defined as follows

$$L_{\text{DSSIM}}(G) = \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}},\mathbf{x}\sim p_{data}} \text{DSSIM}(\mathbf{x}, G(\mathbf{z}|\mathbf{y})), \tag{2.6}$$

where $\mathbb{E}$ denotes expectation.

VGG perceptual loss (Johnson et al., 2016) is also an L2 loss between the features of generated and ground-truth images, obtained from different layers of the pretrained classification network (VGG-19). Instead of exactly matching the pixel values of the generated and ground-truth images, this loss matches their feature representations. This encourages the network to produce images that are perceptually similar to their corresponding target images. Formally, this loss is defined as

$$L_{\text{VGG}}(G) = \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}},\mathbf{x}\sim p_{data}} \left[\sum_{i=1}^{\rho} \frac{1}{C_i H_i W_i} \|(F_i(\mathbf{x}) - F_i(G(\mathbf{z}|\mathbf{y}))\|_2^2 \right], \tag{2.7}$$

where $F_i(\mathbf{x})$ denotes the activation at the i^{th} layer of VGG-19 for the input image $\mathbf{x}$. ρ is the total number of layers of VGG-19 that we are using. Physically $F_i(\mathbf{x})$ is a feature map of shape $C_i \times H_i \times W_i$, where C_i, H_i, and W_i denote the number of channels, height, and width of the i^{th} feature map respectively. We take the features from conv1_2, conv2_2, conv3_2, conv4_3, conv5_1 layers of VGG-19.

Hence, our final objective function becomes

$$G^{**} = \text{argmin}_G \, \text{argmax}_D \, (l_{w1} L_{\text{cGAN}}(G,D) + l_{w2} L_{\text{DSSIM}}(G) + l_{w3} L_{\text{VGG}}(G)), \tag{2.8}$$

where l_{w1}, l_{w2}, l_{w3} are the loss weights.

2.3 Training details

Training of LGVTON is tricky since paired data as shown in Fig. 2.6 is not usually available in the publicly available datasets (Liu et al., 2016a; Dong et al., 2019a) related to fashion. In this section, we discuss our training strategies for different modules.

2.3.1 Pose Guided Warping Module (PGWM)

This module has one trainable component which is the fashion landmark predictor network $\mathscr{F}$. Given the human landmarks of the model and the person and the fashion landmarks of the clothing of the model, $\mathscr{F}$ predicts the fashion landmarks of the target warp of the model's clothing. While $\mathscr{F}$ deals in landmarks only, the inputs and the corresponding ground-truths are extracted from the respective model and person images for preparing the training data. Below, while discussing the training data, we refer to the images corresponding to the landmarks instead of the landmarks, for maintaining the simplicity of the explanation.

We train $\mathscr{F}$ using the data pairs of the same person wearing the same clothing in different poses. In the way we utilize them, these data pairs have sufficient variability of poses between the model and the person. Moreover, such data pairs are available in most of the fashion datasets (Liu et al., 2016a; Dong et al., 2019a). Please see Fig. 2.6 where we have given an example of our training data as well as ideal training data for a better understanding of the reader. However, our training data pairs are devoid of basic clothing structure variability, as the model and person are wearing the same cloth. By clothing structure, we mean design style like sleeve length and neck shape, etc., which are supposed to be encoded by fashion landmarks. At this point, we disregard the texture and color of clothes as $\mathscr{F}$ deals with landmarks only. We observed that our network generalizes well across different input clothing shapes. Note that we can not measure the performance of this network separately due to the lack of ground-truth data; however, the final try-on outputs in a way reflect the performance of each of its component modules including this network. Some results of $\mathscr{F}$ on different clothing shapes and poses are shown previously in Fig. 2.4.

2.3.2 Mask Generator Module (MGM)

Given the target warp along with the densepose of the person, the objective of this module is to predict the mask of the clothing region on the target VTON output. The given target warp (estimated

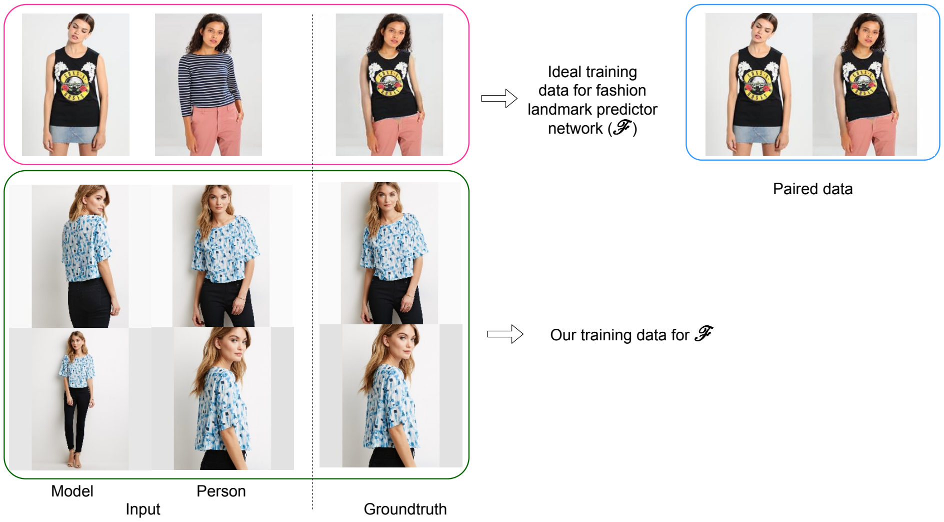

Ideal training data for fashion landmark predictor network ($\mathscr{F}$)
Paired data
Our training data for $\mathscr{F}$
Model
Person
Input
Groundtruth

Figure 2.6: Example of an ideal training data of LGVTON and our training data for $\mathscr{F}$. Although $\mathscr{F}$ operates on landmark data only, for keeping the illustration simple we have shown only the images of the models and the persons from which the landmarks are extracted for preparing the training data. Note that since the ground-truth corresponding to the ideal training data does not exist, hence, for illustration, we used the output of LGVTON.

by the PGWM) may contain warping glitches. So the idea is to make the network learn to identify warping glitches and discard their effects in the predicted mask. An example of a training scenario of this mask generator module (MGM) is shown in Fig. 2.7. To train this network, for each model image of the training dataset, we extract its clothing region (segment) with the help of a human parsing method (Gong et al., 2017) and artificially induce a warping glitch on it. The ground-truth corresponding to this input is the mask of the clothing segment before inducing the warping glitch. An example of the training data of MGM is shown in Fig. 2.7.

To induce a warping glitch artificially on a clothing segment, say, c, we compute $\hat{c}$ by random perturbation of c_{flm}. In this work, we perturb c_{flm} by adding random noise $\mathscr{N}(0, 0.001)$ to it and denote it by $\hat{c}_{flm}$. A TPS transformation mapping from c_{flm} to $\hat{c}_{flm}$ is computed and used to warp c to $\hat{c}$. Multiple such random warps corresponding to each model image and the associated densepose representation of the model are computed to prepare the training data for this module. Training with multiple such perturbed clothing segments corresponding to each clothing segment, make the network learn to extract features required for predicting the desired mask drastically reducing the effect of landmark perturbations, i.e., warping glitches.

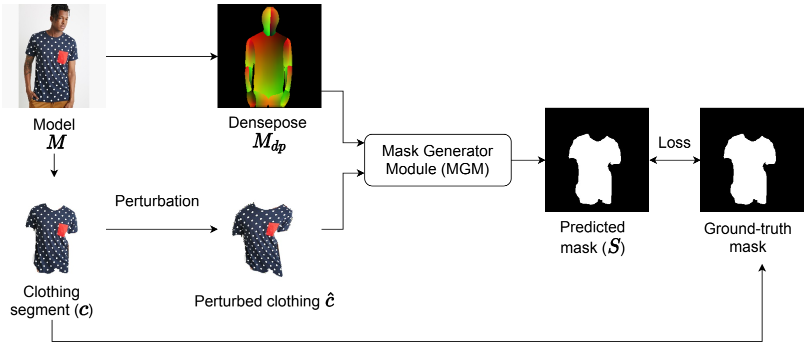


Figure 2.7: Illustration of the training scenario of mask generator module (MGM).

2.3.3 Image Synthesizer Module (ISM)

The objective of ISM is to combine the warped clothing with the person's image seamlessly. Now for a pair of images of a distinct model and person, the ground-truth image of the person wearing the clothing of the model is not available. In this case, we train our image synthesizer module (ISM) using a self-supervised training strategy. An example of training data of this module is illustrated in Fig. 2.8.

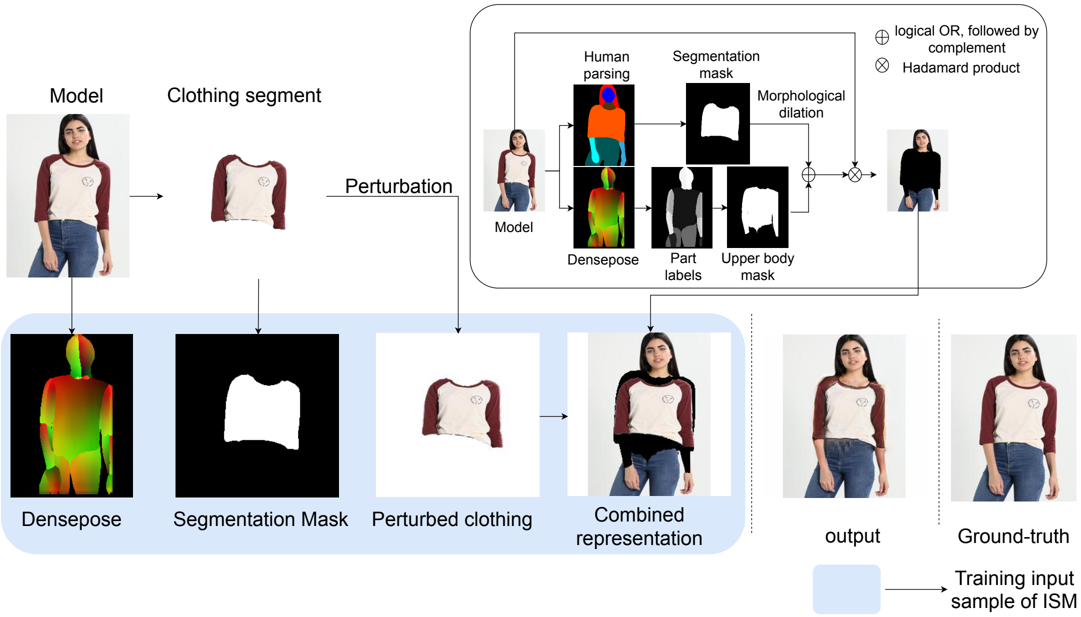


Figure 2.8: Example of a training data sample of image synthesizer module (ISM).

The objective of our self-supervised training is to make the network of ISM learn to fit the model's segmented clothes back onto him. Therefore, for a given model image we first remove all the upper body details from it. However, during testing, if the model and the person are wearing clothes of different structural shapes (e.g., jacket and vest), the network may generate artifacts in the regions of the previous clothes of the person which do not overlap with the clothing regions of the target warp. This is because such a region does not exist in the training data. To overcome this, we adopt a couple of tricks. First, while removing the respective clothing details from the model image, we intentionally change the contour of the clothing region of the model image in the training data by dilating the segment. Next, we perturb the segmented model cloth by applying the same perturbation method specified in the training strategy of MGM. These are done because due to the self-supervision, the network does not encounter many difficult situations during training. In other words, to make the network robust to warping glitches, we induce such problems artificially during training. We then combine the perturbed clothing segment with the modified model image. Fig. 2.8 illustrates the preparation of training data of ISM step by step for better understanding. Note that the mask of the clothing segment is also provided as input. This makes the network understand the true clothing area of the target. Along with this, the densepose representation of the model is also provided that aids the network in understanding the body shape under clothes.

It should be noted that all three modules: PGWM, MGM, and ISM are trained independently of each other. This has two advantages. First, unlike the previous methods (Wang et al., 2018a; Dong et al., 2019a; Yu et al., 2019; Han et al., 2018) our modules can be trained in parallel. This is advantageous in terms of training time in case a sufficient amount of resources is available. Second, any training error of one module does not affect the training of other modules. In other words, each module is trained optimally.

2.4 Experiments

In this section, we first introduce the experimental details of the proposed method, and then we present a comparative study of LGVTON (our method) with other comparing methods such as CP-VTON (Wang et al., 2018a), VITON (Han et al., 2018), MG-VTON (Dong et al., 2019a) and M2E-TON (Wu et al., 2019). With VITON and M2E-TON only qualitative analysis is conducted because of the unavailability of the code of M2E-TON and some errors in the implementation of the VITON's official code.

2.4.1 Dataset

We have reported our results on two datasets: the In-shop clothes retrieval dataset of DeepFashion (Liu et al., 2016a) and the MPV dataset (Dong et al., 2019a). In-Shop Clothes Retrieval benchmark dataset contains multiple views of each person (front, side, back, and full). It has in total of 52,712 images and each image is annotated with fashion landmarks for either the upper body or lower body. Since we are focusing on upper body clothes only, we did the experiments on 33,536 upper body annotated samples. We made two test sets from this dataset. The first test set DeepFashion-I contains 3000 randomly selected image pairs. We selected only the front pose image pairs to maintain a fair comparison with all the comparing methods (CP-VTON, VITON, are limited to front pose human images only). This dataset does not contain any ground-truth data, therefore we prepared a test set with 1800 images where the model and person images are the same. Therefore the ground-truth for each model person pair in this dataset is also the same as these images. We used such a dataset for training the ISM.

Table 2.1: Testset Details.

Dataset	Number of image pairs
DeepFashion-I	3,000
DeepFashion-II	1800
MPV-I	3,000
MPV-II	15,321
MPV-III	3,000
MPV-front	20,034

MVP dataset (Dong et al., 2019a) contains in total 35,687 images. For each image, the corresponding human segmentation, human pose estimates, and the corresponding separate clothes images are also provided. Since the baseline methods (Han et al., 2018; Wang et al., 2018a; Yu et al., 2019; Dong et al., 2019a) require separate clothing images which are not available in DeepFashion, we have done a comparative study on this dataset only. We collected four test sets from MPV. The first test set (MPV-I) is collected by random selection, while for the second test set (MPV-II) we randomly selected 133 person images and made all possible unique triplets ($15321 = 133^2 - 2235 - 133$) from these images, where 2235 images are repeating entries (repetition in person-cloth pairs). This is due to the same cloth image corresponding images of the same model in different poses. The reason for creating MPV-II is to check the performance of the methods on all possible model-person and cloth-person combinations of a set of images. These two datasets do not contain corresponding ground-truths. So, we collected MPV-III which contains 3000 same-person multiview image pairs i.e., the model and the person images in each of the data samples are the different views of the same person. MPV-III is a randomly collected subset of a large set of its type (58,968 data samples). Both MPV-I and MPV-II have variability in terms of the pose. We collected MPV-III to experiment specifically on complex posture because same-person multi-view images often include person images from various angles. MPV-front contains 20,034 same-person front pose image pairs i.e., the model and person images in each of the data samples are the same images. It is the training set of the proposed ISM, that is trained using self-supervision. Table. 2.1 shows the number of images in each of these datasets.

2.4.2 Quantitative Comparison

For quantitative evaluation including comparison with other methods, we have reported the scores obtained with two metrics, namely, Fréchet Inception Distance (FID) (Heusel et al., 2017) and SSIM (Wang et al., 2004). Note that there is no specific metric related to virtual try-on. However, recent methods (Dong et al., 2019a; Han et al., 2018) have used Inception score (IS) (Salimans et al., 2016) and SSIM (Wang et al., 2004). While both Inception score (IS) and Fréchet Inception Distance (FID) (Heusel et al., 2017) are evaluation metrics of GAN but unlike FID, IS does not consider real data at all. As a result, IS can not estimate how well the generator approximates the real data distribution.

Fréchet Inception Distance (FID) is a measure of similarity between two sets of images. It extracts the features embedded in both real and generated images from a layer of inception v3 model (Szegedy et al., 2016) pretrained on ImageNet (Deng et al., 2009). Considering the embedding as a continuous multivariate Gaussian, the mean and covariance are estimated for both the generated (μ_g, σ_g) and the real data (μ_r, σ_r). Then the FID is computed as: $\|\mu_r – \mu_g\|_2^2 + Tr(\sigma_r + \sigma_g – 2(\sigma_r\sigma_g)^{1/2})$. A lower value of FID indicates better results.

The concept of Frechet inception distance is derived from the idea of Frechet distance. If we dig a little into the background of Frechet distance, we see, in Dowson and Landau (1982) as elaborated M. Frechet introduced a metric on the space of probability distributions on $\mathbb{R}$. Between two probability distributions, D_1, D_2, the Frechet distance is defined by, $d_2(D_1, D_2) = \underset{R_1, R_2}{\text{minimize}}\, E\|R_1 - R_2\|$, where R_1, R_2 are random variables having the distributions D_1, D_2 respectively. When D_1, D_2 belongs to a family of distributions closed with respect to changes in location and scale, the Frechet distance takes the following simple form, $d^2 = (\mu_{R_1} - \mu_{R_2})^2 + (\sigma_{R_1} - \sigma_{R_2})^2$. In $\mathbb{R}^n$ this equation generalises to, $d^2 = \|\mu_{R_1} - \mu_{R_2}\|_2^2 + Tr(\sigma_{R_1} + \sigma_{R_2} - 2(\sigma_{R_1}\sigma_{R_2})^{1/2})$.

The objective of generative learning (Goodfellow et al., 2014) is to generate data that match observed data. Intuitively, this distance between the probability of observing real-world data and the probability of generating model data may serve as a performance measure for generative models. However, considering the difficulty of defining appropriate performance measures for generative models the concept of ”Inception score” (Salimans et al., 2016) (IS) was proposed. In computing this score the generated samples are fed into an inception model (Szegedy et al., 2016), (a deep neural network) pretrained on ImageNet (Deng et al., 2009) dataset. This network is a classification network that given an image provides labels for 1000 classes of the ImageNet dataset. Now the goal of IS is to have low output entropy, i.e., localization of the labels towards some specific classes when the image has a meaningful object. However, across a set of images, the entropy should be high indicating the diversity of the generated samples. While this score correlates well with human judgment in terms of generated sample quality but it does not compare the statistics of real-world samples with that of the generated samples. To address this FID was proposed. It compares the distribution of the real data samples and generated samples. However, GAN does not explicitly predicts the distribution, instead generates only samples from the underlying distribution. Hence, moments of the distributions, in practice, mean and covariance are compared instead. Now, instead of considering the intensity values of images, in order to capture the vision-related features, the images are passed through the inception model and features from the coding layer are obtained. The mean and covariance on these coding layers' feature values for both the real and the generated samples are computed to compute FID. Since the Gaussian distribution is a maximum entropy distribution hence the features are considered to follow a multidimensional Gaussian distribution. Finally, the Frechet distance of these two multidimensional Gaussians is used to compute the Frechet inception distance (FID). Since the data on which moments are computed are the feature values from a layer of the inception model hence this distance is called Frechet inception distance.

Experimental evidence shows FID is consistent with increasing disturbances and human judgment. On different kinds of disturbances like gaussian noise, gaussian blur, swirls, and implanted black rectangles FID shows consistent performance when compared with human judgment. Briefly, in terms of discriminability, robustness, and computational efficiency FID performs well. Unlike IS, the FID degrades betters with various kinds of artifacts (Borji, 2019). Since FID can capture artifacts and degradations in images well over IS and also complies well with human judgment, hence in general this measure is widely used as a performance metric in VTON (Xie et al., 2021) (Neuberger et al., 2020).

Table 2.2: Quantitative evaluation on different datasets.

Dataset	Methods	FID↓	SSIM↑
DeepFashion - I	Ours	**33.67**	-
DeepFashion - II	Ours	**26.40**	**0.86**
MPV - I	MGVTON	44.44	-
	CP-VTON	28.38	-
	Ours	**22.59**	-
MPV - II	MGVTON	42.76	-
	CP-VTON	32.56	-
	Ours	**25.10**	-
MPV-III	MGVTON	38.58	0.77
	CP-VTON	25.78	0.81
	Ours	**24.88**	**0.82**
MPV - front	MGVTON	35.70	0.76
	CP-VTON	21.03	0.74
	Ours	**12.06**	**0.89**

The values of FID due to different methods are shown in Table 2.2. We report SSIM scores only for those test sets containing ground-truths i.e., DeepFashion-II, MPV-III, MPV-front. The results demonstrate that our method outperforms others in terms of both metrics.

2.4.3 Qualitative Comparison

Here, we present a visual comparative study of our method (LGVTON) with some baseline algorithms, such as VITON (Han et al., 2018), CP-VTON (Wang et al., 2018a), M2E-TON (Wu et al., 2019) and MG-VTON (Dong et al., 2019a) in Figs. 2.9 and 2.10 [3]. It is observed that in terms of preserving the person and clothes details, LGVTON performs better compared to the baselines. In both Fig. 2.10 and Fig. 2.9 it is observed that based on the quality of the target warp, the results of VITON come after LGVTON in terms of quality; which indicates that the point correspondence-

[3] More results of LGVTON are given in the Appendix A.

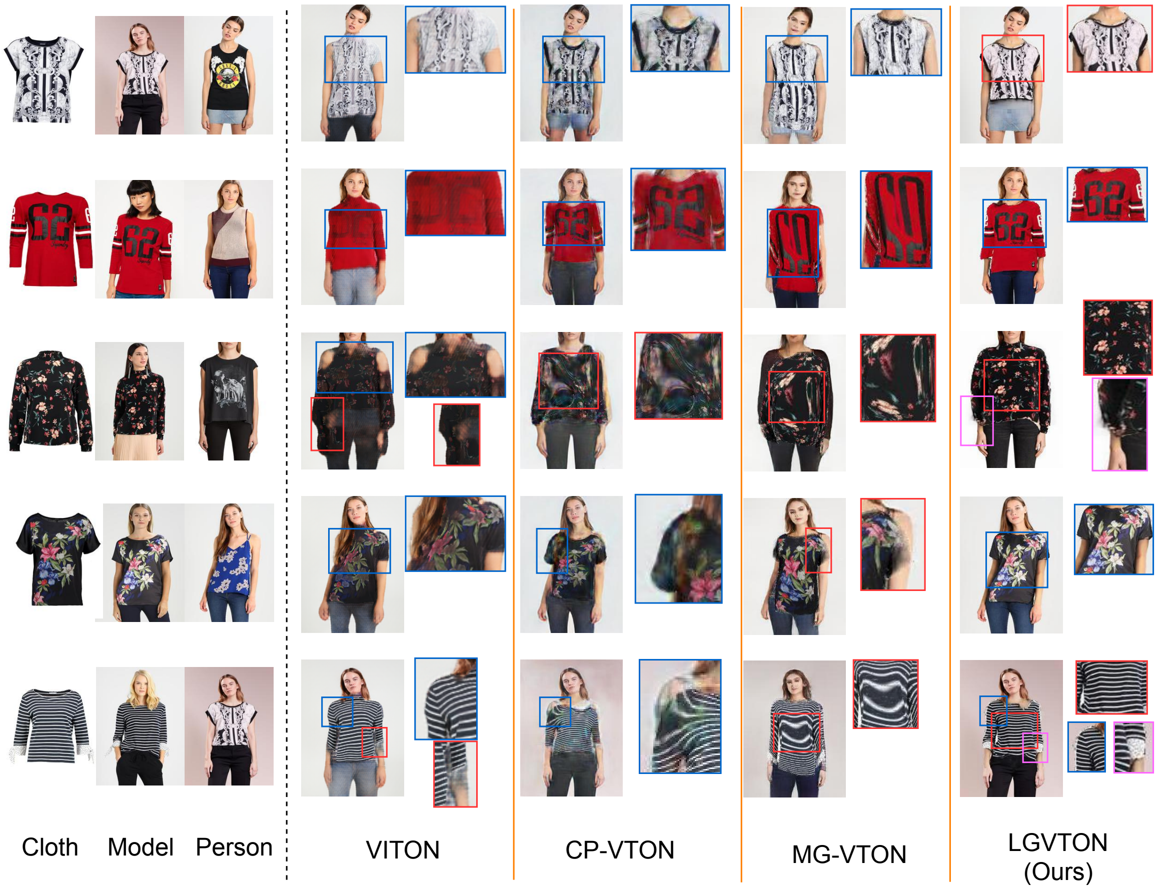


Figure 2.9: Comparative study on MPV dataset. Significant details are zoomed in and shown right after each output.

based methods (both LGVTON and VITON) perform better compared to geometric matching (e.g., CP-VTON and MG-VTON). The geometric matching network employed in both CP-VTON and MG-VTON learns the TPS transformation of the clothes to its corresponding target warp. This network learns the transformation of a predefined mesh grid of the source clothing considering the grid points as control points. For CP-VTON the size of the mesh grid is 3 × 3, therefore the number of control points is 9. Now, a grid can be deformed in numerous ways. However, the human body undergoes restricted deformations due to the presence of limbs, bone joints, and muscles. Considering this, instead of leaving the entire job of learning the feasible set of TPS transformations to a neural network, we choose to restrict the grid transformation implicitly to a plausible range of transformations. To achieve this, our method infers the target control points from a set of source control points, and these points are anatomically meaningful in terms of the human body (human landmarks) as well as clothes

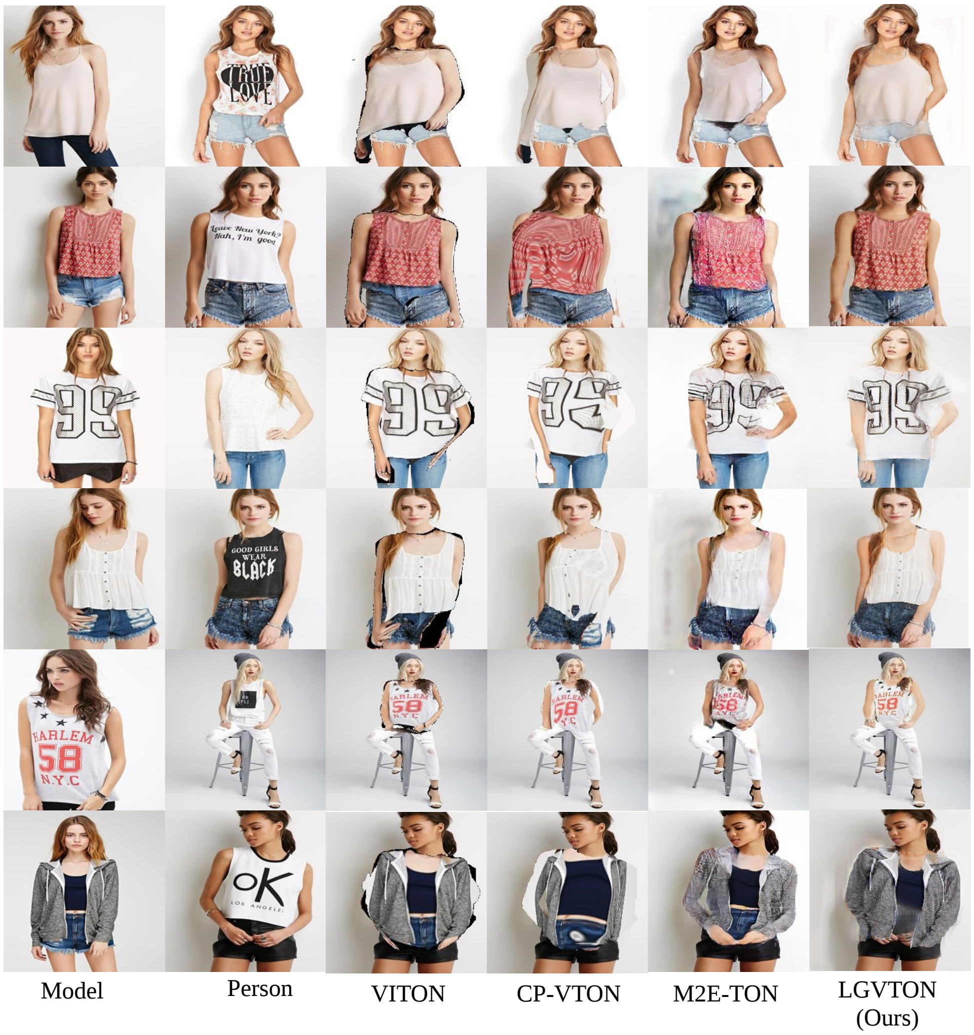


Figure 2.10: Qualitative comparison with other methods. The first five columns are taken from M2E-TON paper.

(fashion landmarks). In other words, the constraint on the possible set of transformations is implicitly imposed in our method by using key points as control points.

Other than the quality of the target warp, VTON also necessitates preserving the details of the other clothing of the person in order to maintain realism in the output. However, the problem of loss of details of other clothes, e.g., pants, etc., is evident in the results of CP-VTON and VITON. On the other hand, the results of MG-VTON show a poor reconstruction of the original person's details. For example, face details are not well retained by MG-VTON. Based on the overall quality of the outputs, we observe that LGVTON outperforms the other methods, which is due to the combined effect of (i) landmark guided warping of the model clothing leading to better target warps and (ii) the convex combination layer in the ISM that helps to retain the details of the warped clothing and the person in the final try-on output. In the results of M2E-TON (Fig. 2.10), it can be observed that the colors of the target clothes are not the same as that of the model clothes and the human faces look brighter. However, no such artifact or photometric change is observed in our results.

We also did some experiments (as shown in Fig 2.11) to show that LGVTON's performance is unaffected in the presence of background clutter, which establishes its applicability in an in-the-wild setting. To generate these results, we took a random image from the MPV dataset and modified it by adding backgrounds collected from the internet then applied virtual try-on to it. In addition, we also did a comparative study on similar setting for other different compared methods e.g., CP-VTON and MGVTON which is shown in Fig. 2.12. We observe that unlike ours both CP-VTON and MGVTON fails to keep proper background details in their output [4].

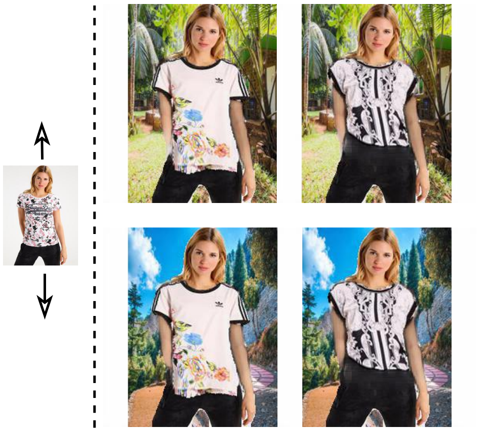


Figure 2.11: Results on images in cluttered background. Observe that the performance of LGVTON is unaffected by the background clutter.

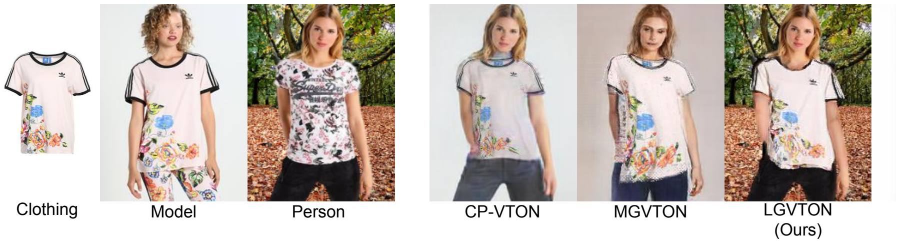


Figure 2.12: Comparative study in case the person's image contains background clutter.

[4] Note that we do not compare with VITON in this case since it is already mentioned in the corresponding paper Han et al. (2018) that it works well in case of background clutter.

2.4.4 On Different poses of the Model and the Person

Here we examine the performance of LGVTON on different poses (other than the front as it has been already discussed) of the model and the person. As shown in Fig. 2.13 it may be observed that LGVTON is not only constrained to the front pose but works even for the back pose also.

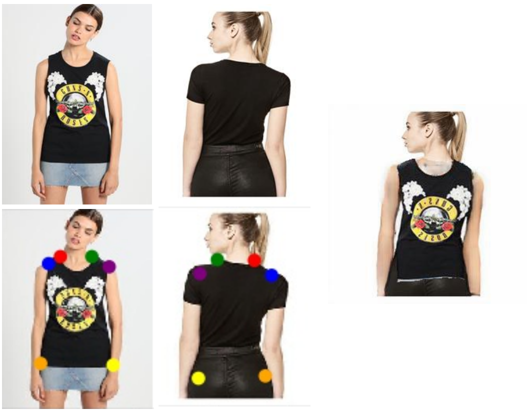


Figure 2.13: Results on the back pose shows LGVTON is not constrained to the front pose only.

2.4.5 Ablation Study

Here, we present a quantitative and qualitative study on the significance of each component of LGVTON. This study is conducted on the MPV-I test set (explained in Section 2.4.1) based on the FID metric. The values of this metric are reported in Table. 2.3, which shows complete LGVTON achieves the best FID score.

Table 2.3: Quantitative ablation on MPV dataset.

Methods	FID↓
LGVTON (w/o MGM)	23.02
LGVTON (w/o correlation layer in PGWM)	23.21
LGVTON (w/o fashion landmarks)	22.74
LGVTON (w/o fashion landmarks, w/o MGM)	23.03
LGVTON	**22.59**

2.4.5.1 Utility of Different Loss Functions for Training ISM

We run a comparative study on the effect of different loss functions used to train ISM. Keeping the other settings the same, we train 3 different instances of ISM with different combinations of loss functions as shown in Fig. 2.14. As GAN tries to approximate the data distribution, so it generates

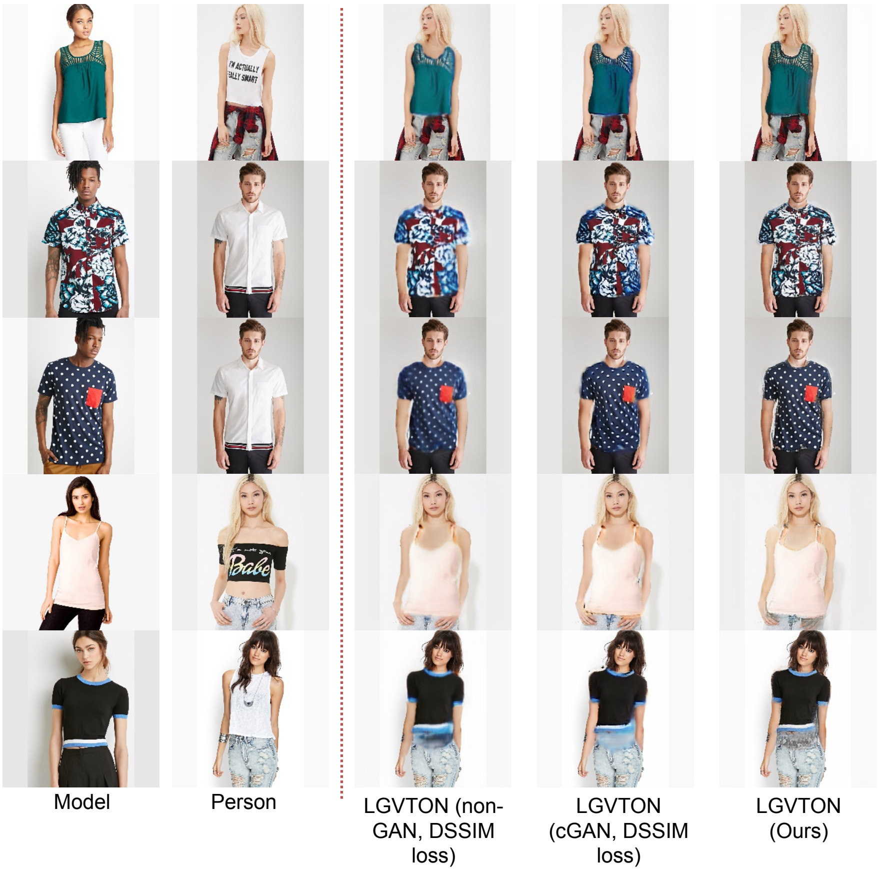


Figure 2.14: Effectiveness of different losses during training the Image Synthesizer module (ISM). (1) LGVTON (non-GAN, DSSIM loss) - a non-GAN variant of ISM trained with DSSIM loss, (2) LGVTON (cGAN, DSSIM loss) - cGAN with the generator trained with DSSIM loss, (3) LGVTON - similar to (2) but the generator is trained with DSSIM and VGG perceptual loss both. Notice the clarity of output increases from LGVTON (non-GAN, DSSIM loss) to LGVTON.

better output than its non-GAN variant (Goodfellow et al., 2014). This is evident in the results of the non-GAN variant of LGVTON where the ISM is trained with DSSIM loss. We call this variant LGVTON (non-GAN, DSSIM loss) and the corresponding GAN variant LGVTON (cGAN, DSSIM loss). Comparing the results of LGVTON(cGAN, DSSIM loss) and LGVTON (ours) (a cGAN, where the generator is trained with DSSIM and perceptual loss both), it can be observed that the result improves in the presence of VGG perceptual loss.

2.4.5.2 Significance of Fashion Landmarks in PGWM

We conduct a study on PGWM when the warping is done with human landmarks only instead of both human and fashion landmarks. Fig. 2.15 shows two cases portraying the effectiveness of fashion landmarks around the collar and hem. Having only human landmarks might serve the purpose but at the cost of an increase in the amount of warping glitches, which is tackled to some extent in ISM with the help of the mask generated by the MGM. The scores are reported in Table. 2.3 (LGVTON (w/o fashion landmarks)) show that without fashion landmarks the performance of LGVTON degrades. However, for obvious reasons, the performance degrades even more if the support of the MGM is also removed (observe the score of LGVTON (w/o fashion landmarks, w/o MGM) in Table. 2.3.

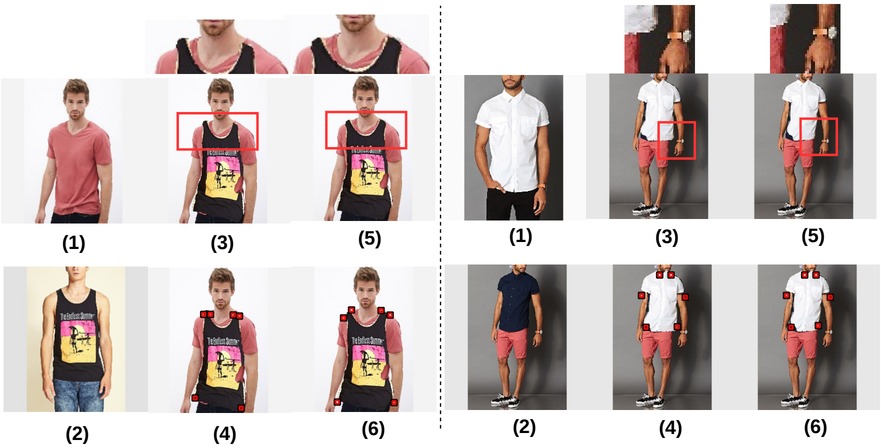


Figure 2.15: Role of different fashion landmarks in predicting target warp of the model clothes in PGWM. The figure shows the utility of fashion landmarks around the collar (left) and hem (right) respectively. For a better understanding of the viewer, instead of showing the generated warped clothes image only, we show the overlay of it on the person image. (1) Person, (2) model, (3) clothes warped using only human landmarks, (4) transformed locations of fashion landmarks of the model clothes, obtained by the corresponding transformation function of (3), (5) clothes warped by PGWM using both human and fashion landmarks, (6) fashion landmarks predicted in PGWM, which is observed to be more accurate than (4). This results in better warping of the model clothes as shown in (5) in comparison to that in (3).

2.4.5.3 Effectiveness of Correlation Layer in PGWM

We study the effectiveness of the correlation layer in the fashion landmark predictor network of PGWM (refer to Fig. 2.16). Empirically from Table. 2.3 it is observed that the presence of the correlation layer improves the performance, as we see without this layer the performance degrades. As previously discussed the correlation layer is used to compute the matching between the human landmarks of the model and the person. Empirically we see these matching values aid in predicting better estimates of fashion landmarks of the target warp clothes which in turn assists in predicting better estimates of fashion landmarks of the target warp clothes.

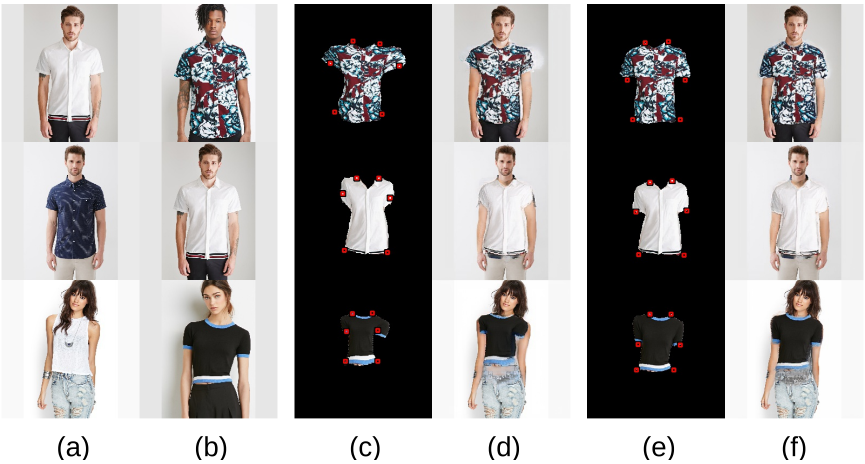


Figure 2.16: A study on the effectiveness of correlation layer in fashion landmark predictor network $\mathscr{F}$ of PGWP in LGVTON. (a) Person image, (b) Model image, (c, e) predicted locations of fashion landmarks and warp clothes generated by PGWM when $\mathscr{F}$ is trained w/o correlation layer and with correlation layer respectively. (d, f) final VTON result for the warp cloths shown in (c) and (e) respectively. We observe that the results in (f) are better than that in (d), which justifies the potency of the correlation layer in PGWM.

2.4.5.4 Studying the Role of Target Mask as Input to ISM

For understanding the effect of the target mask, we train an instance of ISM without providing the target mask as input. Now it is observed that w/o the target mask, the network can not identify the warping glitches. This can be verified from Fig. 2.17(h) where the effect of the warping glitch is propagated to the output. However, when a target mask is given as input, ISM can identify the areas of warping glitches and take the necessary action according to the type of glitch. The reason being the variety of clothing types makes the network confused to distinguish a warping glitch from the design of the clothes, e.g., in the first two rows, the inappropriate estimation of c'_{flm} causes the sleeves to be stretched more outwards. While that in Fig. 2.17(j) is not observed as the network removes those areas of extra stretch and replaces them with the background. In the example of the third row Fig. 2.17(h), the effect of the warping glitch exposes some body parts near the right neckline of the person (better viewed when zoomed in), while this gets filled with a clothes texture and color in Fig. 2.17(j).

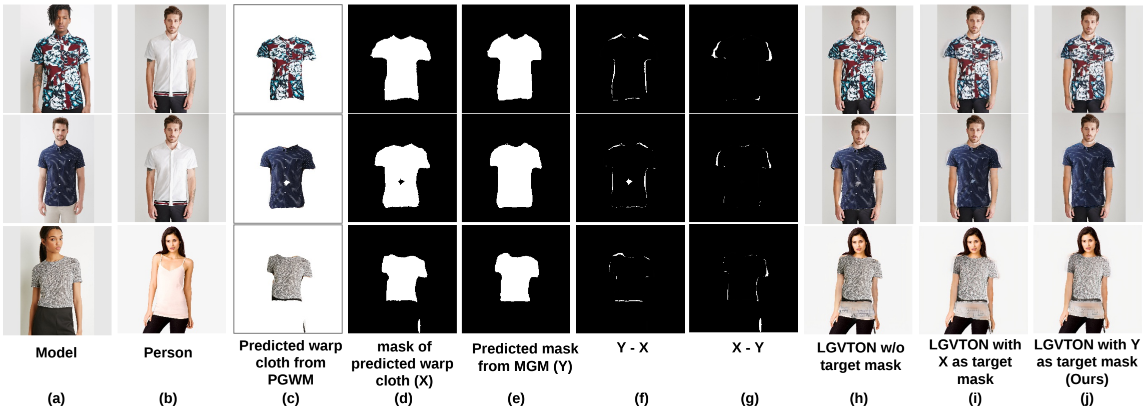

Model
(a)
Person
(b)
Predicted warp cloth from PGWM
(c)
mask of predicted warp cloth (X)
(d)
Predicted mask from MGM (Y)
(e)
Y - X
(f)
X - Y
(g)
LGVTON w/o target mask
(h)
LGVTON with X as target mask
(i)
LGVTON with Y as target mask (Ours)
(j)

Figure 2.17: Effectiveness of target mask generated by our mask generator module (MGM). We notice significant differences (as shown in (f), (g)) between the mask (d) of the predicted warped clothes (c) and the mask (e) predicted by MGM corresponding to the warp clothes (c). (h) shows the final try-on result contains artifacts when ISM is trained without the target mask. To show the effectiveness of the target mask we also show the result where instead of (e) we give (d) as target mask input to ISM; which is shown in (i). As we observe the artifacts still remain in (i). Whereas, when (e) is given to ISM as a target mask, the result (j) improves, e.g., the areas with pixel value 1 in (f) are filled with necessary color and texture details, and those in (g) are replaced with background details resulting in a better VTON result. Note that the hole in the warp clothes in row 2 is not due to a warping glitch. This is due to inaccurate human parsing. However, LGVTON handles this also.

2.4.6 Implementation Details

We trained PGWM, MGM and ISM approximately for 316, 9000, 20 number of epochs respectively, with Adam optimizer keeping $lr = 0.001$, $\text{beta1} = 0.9$, $\text{beta2} = 0.999$. The network architectures of our modules have been discussed in the Appendix related to this chapter. The values of the loss weights are $l_{w1} = 1$, $l_{w2} = 1$, $l_{w3} = 10^3$. To report the computation cost of our method, below we compare the number of parameters of our method as well as that of the different comparing methods.

Number of parameters: Number of parameters is an implementation-independent metric that is a reasonable factor for comparing the computation cost. Therefore, we report the number of parameters of our method as well as that of the comparing methods in Table. 2.4. Each of these methods uses human parsing (Gong et al., 2017) and human pose estimation (Cao et al., 2017) methods to get the corresponding annotations, whereas our method uses an additional annotation of dense-pose (Alp Güler et al., 2018). We report the count of parameters of (Gong et al., 2017; Cao et al., 2017; Alp Güler et al., 2018) in the third column. The total number of parameters of the annotation methods corresponding to each of the compared methods is reported in the fourth sub-column of the third column. Let us call this parameter count P_2. While the number of parameters in the try-on methods excluding that of the annotation method's is reported in the second column. Let us call this parameter count P_1. Finally, the total number of parameters is reported in the fourth column (P_1 + P_2). Note that although the total count of parameters in our method is a little higher than that of VITON and CP-VTON, the P_1 parameter count is quite less in our method compared to others. The number of parameters in our warping module is 1.69M compared to that of CP-VITON's 19.05M and MGVTON's 45.80M. We do not compare VITON here as it uses shape context matching based on TPS warping (Belongie et al., 2001), a traditional warping method that has its limitations, especially in terms of the time of execution. However, the point we want to make here is that our way of warping reduces the computation cost of warping significantly.

Table 2.4: Comparing the number of parameters (in millions) of different methods. The total number of parameters in the proposed method is given in column P_1. The parameters required to compute different inputs are specified in the columns under P_2. The sum of the values reported in P_1 and P_2 are reported in the column Total. We used the following abbreviations for the different input estimation methods, HP - Human parsing, PE - Pose estimation, DPE - Densepose estimation.

Methods	P_1	P_2				Total
		HP	PE	DPE	Total P_2	
VITON	29.34	75.65	52.31	-	127.96	**157.30**
CP-VTON	40.40	75.65	52.31	-	127.96	168.36
MGVTON	224.46	75.65	52.31	-	127.96	352.42
Ours	**2.45**	75.65	52.31	59.73	187.69	190.14

2.5 Discussion

This chapter presents a self-supervised landmark-guided approach to virtual try-on which synthesizes the image of a person wearing model clothes. Unlike many existing works, this work requires only the images of the model and the person without requiring any separate clothing image. This makes it more effective, as having a separate clothing image is difficult. Our method contains three modules. The first module utilizes the correspondence between the estimated landmark sets of the model and the person to predict the target warping of the model clothes. Employing structural key points i.e., landmarks for computing the target warp enables the warping function to be defined in terms of the geometric structure of the human body and the clothing. Employing human landmarks implicitly maintains the feasibility of the target warp. Moreover, consideration of fashion landmarks enables our method of warping to work based on the geometric structure of the cloth; thus independent of the texture of the clothes, which otherwise confuses the warping process, especially in case of complex patterns. In order to refine the fit of the predicted warp, we propose a mask generator module. Our final module, the image synthesizer module, combines the aligned model clothes and person to synthesize the final output. We conducted an ablation study that establishes the necessity and efficacy of each of these modules. This is worth mentioning that this work explores a new research direction involving landmarks in the domain of virtual try-on.

The main contributory idea of this work is proposing the idea of landmark-based image registration for warping the source clothing. In this context, we use both human and fashion landmarks correspondences from the source mesh to the target mesh and compute the target warp of the source clothing using TPS transform. However, the clothing necessarily does not cover all the parts of the human body (upper body in our case). Hence consideration of those landmark correspondences that do not belong to the clothing region does not aid in the computation of the warp, rather sometimes puts unnecessary constraints on the warping function which affects the result. This can also be verified from the examples shown in Fig. 2.18. Moreover, considering all the landmarks at once to compute the TPS parameters fails in case the arms of the reference person are bent significantly. As can be seen in Fig. 2.18 when the target person poses with his arms bent or the arm falls on the torso the warps computed by our method show unnecessary deformation of the source clothing.

The issue of warping glitches as discussed earlier in this chapter is also a concern with this approach. Warping glitches occur because human structural key points (human landmarks) do not always estimate the human body shape well. Due to this, artifacts are getting generated near the boundary of the warped clothing. Although we employ our mask generator module (MGM) to address this issue, it does not achieve success all the time. So we foresee that employing a better estimate of the human body shape may address this issue more aptly.

One of the major limitations of this method is that it does not apply to the cases when any of the images of the model or the person is in side view. In Fig. 2.19 we have demonstrated two cases. In the first case (first row) - the model is in front and the person in the side pose. Here we see that due to the occlusion of the hand of the person, the related human landmarks estimated are wrong (d), which

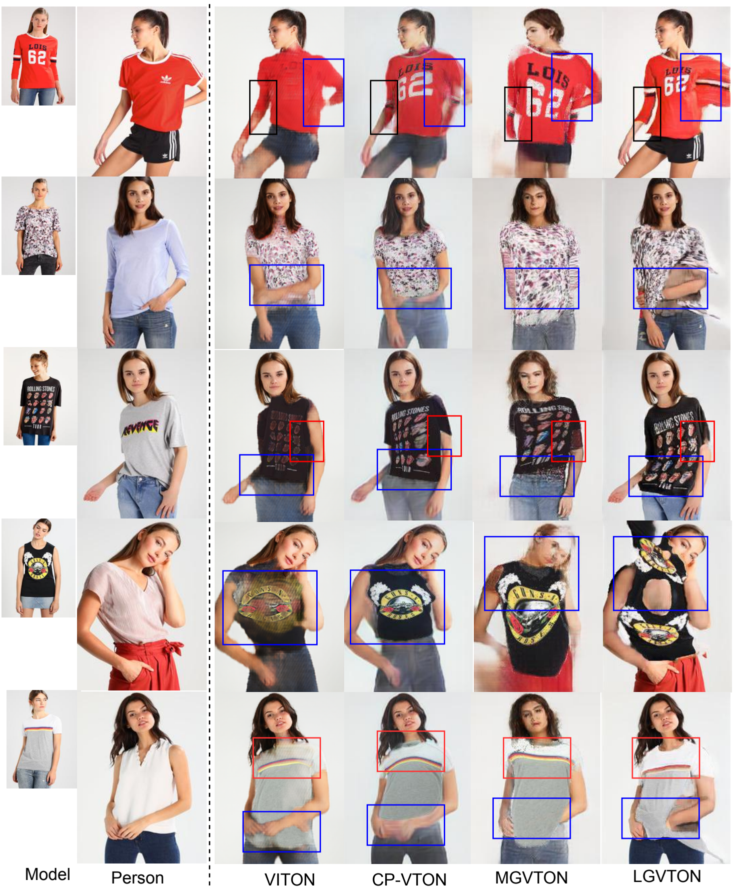


Figure 2.18: Our method performs poor in case of significant bending of arm in the reference person.

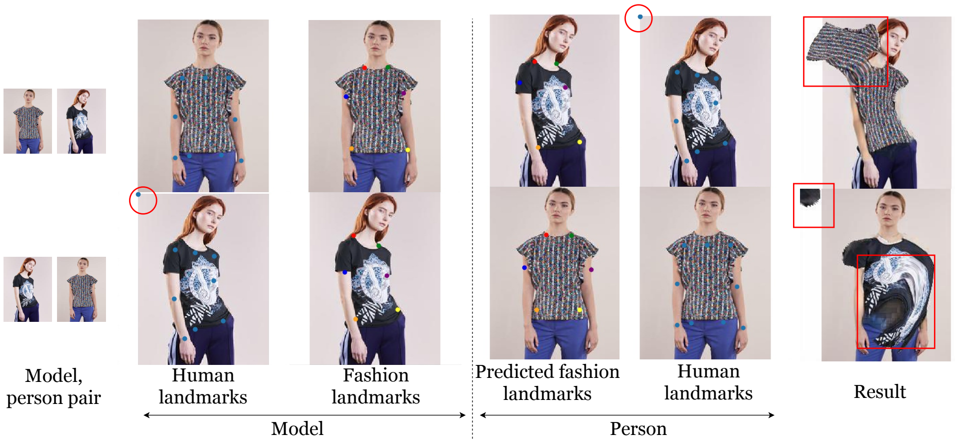


Figure 2.19: LGVTON shows poor performance in the case at least one of the model or the person is in side pose; as the predicted landmarks in occluded regions are often incorrect.

affects the results (e). In the second case (second row) - the model is in the side pose and the person is in the front pose. Similar to the previous case, an incorrect landmark estimation affects the result (e). Based on this explanation it may be understood that whenever any of the model or person is in a side pose then due to the occlusion of certain parts of the body landmarks related to those parts are estimated incorrectly. Since our method of warping is based on the assumption that the estimated landmarks are correct therefore incorrect landmark estimates affect the results.

In this chapter, we have used the predefined features i.e., the locations of two types of landmarks and the correspondence between these two sets for computing the target warp. Owing to the problem associated with this approach in the next chapter we intend to explore a neural network-based feature learning approach. More specifically, instead of using explicit location-based feature (i.e., landmarks) correspondences, in the next chapter, we intend to learn the features necessary for such a task from a rich collection of shape and pose representations of humans.

Chapter 3

Geometric matching on dense human pose representation for clothes warping

3.1 Introduction

In general, any virtual try-on (VTON) system takes as input a source clothing along with a target person representation. The source clothing can be in two different forms, either as a separate cloth image or a model wearing it. The VTON problem related to the former case is called cloth-to-person (C2P) and the latter case is called model-to-person (M2P). While most of the existing approaches take a separate cloth image as input, from the data availability perspective, the latter case is more realistic as elaborated in Chapter. 1 and 2. In view of this, this chapter proposes a novel M2P VTON solution.

In Chapter 2 we opted for a landmark-based image registration approach for clothes warping in the context of the VTON problem. In general landmark-based image registration contains 3 steps, where the first two steps involve identifying the landmark points in the source and the target images followed by establishing the correspondences between the two sets of landmarks. The third step involves computing the transformation between these two sets. In Chapter 2, we followed recognizing structural key points of human and that of the clothes as landmarks and employed TPS transform for computing the transformation between the landmark sets of the source and the target images. Therefore, for image registration, we used some predefined features and their established correspondences in the first two steps there. While this idea was simple and effective compared to the geometric matching method proposed by CP-VTON, it also has some limitations. Hence, in this chapter, instead of predefined features and their established correspondences, we explore deep neural network-based feature learning and matching.

The estimation of correspondences between images is a primitive computer vision problem. Image matching has applications in various domains, for example, image retrieval, where a query image is needed to be matched with the database images in order to find a match. Many other problem domains like satellite imaging or medical imaging use image matching. Before the advent of the deep neural network, the traditional method of estimating correspondences included detecting local features using SIFT (Lowe, 2004) or HOG (Dalal and Triggs, 2005). CNN-based image matching was first proposed by (Rocco et al., 2017), where the authors proposed an architecture called Geometric Matching Network (GMN) mimicking the traditional matching approach. In general, in traditional feature matching between the images of two objects, there are 3 steps included, detecting local features, matching them followed by pruning incorrect matches using some constraints, and then estimating global geometric transformations to transform from one image to the other. In general in traditional feature extraction SIFT (Lowe, 2004) or HOG (Dalal and Triggs, 2005) features were popularly used. In global geometric transformations, RANSAC and Hough transforms, etc., have been used previously. However, with the advent of neural networks instead of some predefined feature extraction methods now objective specific learning of features has been seen to produce better results. Considering this in GMN there are 2 feature extraction networks that learn to extract features from the two input images. In classical methods, the feature matching step includes computing similarities between all possible pairs of features between the two images. Pruning of matches is done by thresholding or nearest neighbor matching. In GMN this is done using the correlation layer, where the similarities in terms of the dot product are computed between all pairs of features computed from the previous step. A channel-wise normalization of the correlation results is done in order to remove ambiguous matches. This is followed by a regression network that takes the results of the normalized correlation map and then estimates the parameters of the geometric transformation. In the context of the VTON problem, CP-VTON (Wang et al., 2018a) employed this architecture for warping the source clothing. In contrast to handcrafted feature generation and matching algorithms, the CNN-based matching architecture contains trainable convolution layers for feature learning as well as matching. Finally, a transformation estimation layer with some predefined transformation function is employed. Here, estimation essentially means predicting the parameters of the transformation. Compared to Rocco et al.'s approach (Rocco et al., 2017), where the two inputs are the two images between which the match is established. In CP-VTON the inputs are the clothing image and a clothe-agnostic person representation. The agnostic representation contains (i) the pose heatmap of the person containing the human landmarks of the person, (ii) body shape representation, which is a 1-channel feature map of a blurred binary mask that roughly covers different parts of the human body, and (iii) the face and hair of the person that is a part of the identity of the person. A similar matching method has been used by many (Dong et al., 2019a; Yu et al., 2019; Raffiee and Sollami, 2021) VTON approaches. MGVTON (Dong et al., 2019a) adopted the CNN-based geometric matching with the mask of the body shape of the person obtained from the segmentation map of the person and the mask of the clothing image. The difference with CPVTON is that it excludes the consideration of texture from the clothing image. Compared to CP-VTON, VTNFP (Yu et al., 2019) have one difference. Here instead of a blurred binary mask to represent the body shape, a 1-channel body part map containing class labels of 6 body parts is used. In comparison to VTNFP GarmentGAN (Raffiee and Sollami, 2021) additionally used the previously generated segmentation map of the target clothing on the reference person.

Previous approaches have provided the clothing image or its mask along with the cloth-agnostic representation of the reference person as inputs to the geometric matching network (GMN). In the case of the M2P VTON problem where clothing image is not available, sometimes formulating the objective as a trivial supervised learning problem may be difficult since paired data (images of model and person wearing the same cloth) for such objective is not available. An alternative training strategy may use synthetic data as used in (Rocco et al., 2017). But generating a synthetic dataset in this scenario may be complex. This is because the human body undergoes very constrained deformation due to the presence of limbs, skin, and muscles, and mimicking this using any 2D geometric transformation is difficult.

Compared to the previous approaches, our warping approach in this thesis is based on the observation that the transformation of the model's clothing can be estimated from the changes in body shape and pose between the model and the person. We impose this idea in our geometric matching network by providing dense human pose representations of the model to the person as inputs. Therefore, instead of learning to match between the clothing and the person's shape and pose, our idea is to learn to establish the matching between the body shape and pose of the model and the person. In the previous approaches, the expected output is the transformation of the input clothing that results in the desired target warp. In contrast, our objective is to predict the transformation of the densepose of the model to the densepose of the target person, which also happens to be one of the inputs to the matching network. In order to compute the target warp, we apply that estimated transformation to the segmented clothing of the model to get the desired target warp. Our network training is self-supervised as the expected output is also one of the inputs during training; whereas for other methods (Dong et al., 2019a; Yu et al., 2019; Raffiee and Sollami, 2021) it is supervised requiring paired training data, i.e., the clothing image and the image of a model wearing that clothing. Self-supervised learning is a class of supervised learning which does not require any explicit labeled input-output pairs. Rather knowledge is extracted from the inputs and used for learning. Here we call our training self-supervised because our approach does not use paired training data, as we have already discussed. Rather, one of the inputs is used as the ground-truth. Whereas, for other methods (Dong et al., 2019a; Yu et al., 2019; Raffiee and Sollami, 2021) image of a model is paired with a separate image of the corresponding article of the clothing the model is wearing. We call such clothing-model pair as paired data in the current problem context, which we have also elaborated on earlier.

3.2 Proposed Approach

We propose a cloth warping method in the M2P VTON framework. Formally, given an image of a model M wearing the desired clothing c and an image of the person P willing to try the cloth of the model virtually, the objective of the current problem is to find a warping of c, say c' so that it fits the target person P.

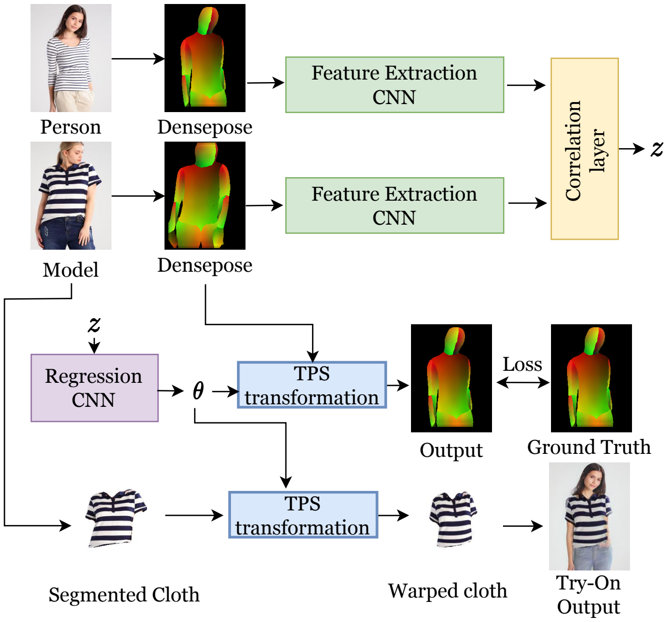


Figure 3.1: Block diagram depicting the workflow of our method.

We model the transformation of c to c' as the body shape and pose change from the model to the target person. The reason is obvious; while wearing the clothing it gets warped in compliance with the body shape and pose of the person. Therefore, the warping of the clothing can be modeled by the deformation of the human body. Here we are assuming the cloth is made of flexible and stretchable material.

Considering the representation of the body shape and pose of the model and the target person as M_{dp} and P_{dp} respectively, we formulate this problem as

$$\begin{aligned}\hat{P}_{dp} &= \tilde{F}(M_{dp}, P_{dp}) \\ &= f_{tps}(\tilde{g}(M_{dp}, P_{dp}), M_{dp}).\end{aligned} \tag{3.1}$$

Here M_{dp} and P_{dp} denote the densepose representation of M and P, respectively, and $\hat{P}_{dp}$ denotes the predicted estimate of P_{dp}. $\tilde{F}(.,.)$ is a geometric matching network (Rocco et al., 2017) (GMN) which is a convolutional neural network-based matching feature learning and matching network. It estimates the parameters of TPS transformation ($f_{tps(.,.)}$) between the two inputs M_{dp} and P_{dp}. This network consists of a feature extraction CNN mimicking the traditional feature extraction methods, followed by a correlation layer and a regression network in accordance with the concepts of feature matching and pruning of incorrect matches. The regression network predicts the parameter estimates of a geometric transformation (e.g., thin plate spline (TPS) (Sprengel et al., 1996) in our case). The final layer applies the transformation to the respective input M_{dp} and the gradient of the loss between the output $\hat{P}_{dp}$ and the corresponding ground-truth P_{dp} is used to train this network. We denote $\tilde{g}(.,.)$ as the pipeline of $\tilde{F}(.,.)$ containing the feature extraction network, correlation layer, and feature regression network.

A block diagram of $\tilde{F}(.,.)$ is given in Fig. 3.1. We use DSSIM loss to train this network, where DSSIM is related to SSIM (Wang et al., 2004) as DSSIM $(\cdot,\cdot) = (1-\text{SSIM}(\cdot,\cdot))/2$.

Once $\tilde{F}(.,.)$ is trained, we use this network to compute c'. Computation of c' can be formulated as follows,

$$\begin{aligned}\theta &= \tilde{g}(M_{dp}, P_{dp}), \\ c' &= f_{tps}(\theta, c),\end{aligned} \tag{3.2}$$

where θ is the parameter of the TPS transform.

3.3 Experiments

In this section, the experimental results of our proposed method applied to our test sets are presented. Comparative study with state-of-the-art methods, CP-VTON (Wang et al., 2018a), MG-VTON (Dong et al., 2019a), VITON Han et al. (2018) and M2E-TON (Wu et al., 2019) and LGVTON (ours, ch2) is also carried out. With VITON and M2E-TON only qualitative analysis is conducted because of the unavailability of the code of M2E-TON and some errors in the implementation of the VITON's official code.

3.3.1 Quantitative Analysis

We report our results based on two metrics, namely, Fréchet Inception Distance (FID) (Heusel et al., 2017) and SSIM (Wang et al., 2004). Note that, this chapter proposes the clothes warping method which is the solution to a part of the VTON problem. However, the evaluation metrics FID and SSIM are applicable only to the final try-on output. Hence, we compute the final try-on output over our generated warps using the try-on module of CP-VTON, trained on the MPV dataset.

In the quantitative analysis shown in Table. 3.1, it may be observed that compared to CP-VTON and MGVTON, our method secures better scores in both metrics. However, this method achieves comparable performance to our previous approach LGVTON (ours, ch2). In fact, other than MPV-III in all the other test sets our previous approach secures a little better scores. However, recall that the evaluation is based on the final try-on output, and the try-on network of CP-VTON which we employed to compute the final output fails to preserve the other clothing details very accurately. This might cause the degradation of scores of this method compared to our previous approach in chapter 2. The MPV-III dataset contains more pose variability compared to other test sets and this method scores better on this test set. Keeping in mind the limitation of the try-on network of CP-VTON this shows that this method can handle pose variability much better compared to the other methods.

The reason for the poor FID and SSIM scores on DeepFashion is due to the underperformance of the try-on module of CP-VTON which is not robust across datasets. As a result, poor reconstruction of the existing clothing details is observed when applied to some datasets (e.g., DeepFashion) other than its training dataset (e.g., MPV). This may also be verified from our results presented in Fig. 3.5.

Table 3.1: Quantitative evaluation on different datasets.

Dataset	Methods	FID↓	SSIM↑
DeepFashion - I	Ours (ch2)	**33.67**	-
	Ours	64.35	-
DeepFashion - II	Ours (ch2)	**26.40**	**0.86**
	Ours	63.16	0.75
MPV - I	MGVTON	44.44	-
	CP-VTON	28.38	-
	Ours (ch2)	**22.59**	-
	Ours	24.11	-
MPV - II	MGVTON	42.76	-
	CP-VTON	32.56	-
	Ours (ch2)	**25.10**	-
	Ours	26.43	-
MPV-III	MGVTON	38.58	0.77
	CP-VTON	25.78	0.81
	Ours (ch2)	24.88	0.82
	Ours	**22.53**	**0.86**
MPV - front	MGVTON	35.70	0.76
	CP-VTON	21.03	0.74
	Ours (ch2)	**12.06**	**0.89**
	Ours	14.34	0.80

3.3.2 Qualitative Analysis

Here, we present some visual comparisons with the benchmark methods. Figs. 3.2, 3.3 shows the comparison on MPV dataset. Fig. 3.4 presents the comparative study with M2E-TON. We also presented our results on the DeepFashion dataset in Fig. 3.5. It may be observed from Fig. 3.2 and Fig. 3.3 that in the case of clothes with patterns the results of CP-VTON and MG-VTON degrade showing unrealistic warps. Our goal in this work was to make the network learn better correspondences without resorting to any additional constraints. This is possible by learning better features through the feature extraction layer. It is observed from the results that our formulation of the problem makes the network learn better features that justify the precise warps predicted by our method.

Intuitively, the parameters of warping of the clothing should be independent of its texture and color. That means warping should be related only to the structure of the clothing. Our method satisfies this because our formulation training of GMN does not require any clothing details. In this work, we

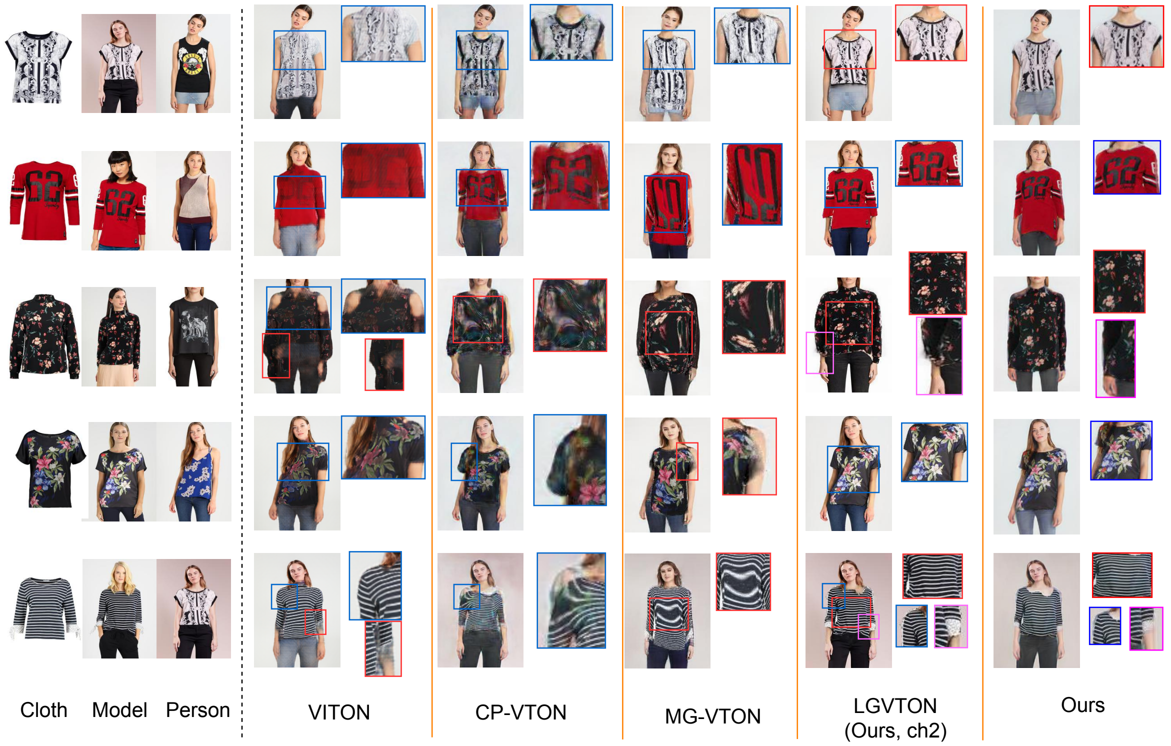


Figure 3.2: Visual comparison (Please zoom in for details).

use densepose for representing human body shape and pose. However, existing methods (Dong et al., 2019a; Wang et al., 2018a) use blurred body shape images instead. That is obtained by downsampling the mask of the image of the person to a resolution of 1/16 of the original and then upsampling to its original resolution. The reason for doing this process is to remove the shape details of the existing clothing of the person. Note that compared to the downsampled shape representation, densepose is richer in information and thus causes better learning of our warping network.

In the case of VITON, the loss of texture details is observed which is due to its coarse-to-fine architecture. As observed from Fig. 3.2 compared to the previous approach we see similar performance in the case of less complex human poses where arm folding or bending is very less. On the contrary, for the opposite cases as shown in Fig. 3.3 where folding or bending of arms are observed, our approach shows better performance compared to the approach in chapter 2. This establishes the usefulness of learned features and their matching based on a deep neural network. Fig. 3.4 shows that in terms of color and texture of cloths, our warps are better compared to that of M2E-TON.

Another point is that this method of warping learns based on the body shape and pose, hence the performance of this method is robust to background clutter in the model image or clothing type (i.e.,

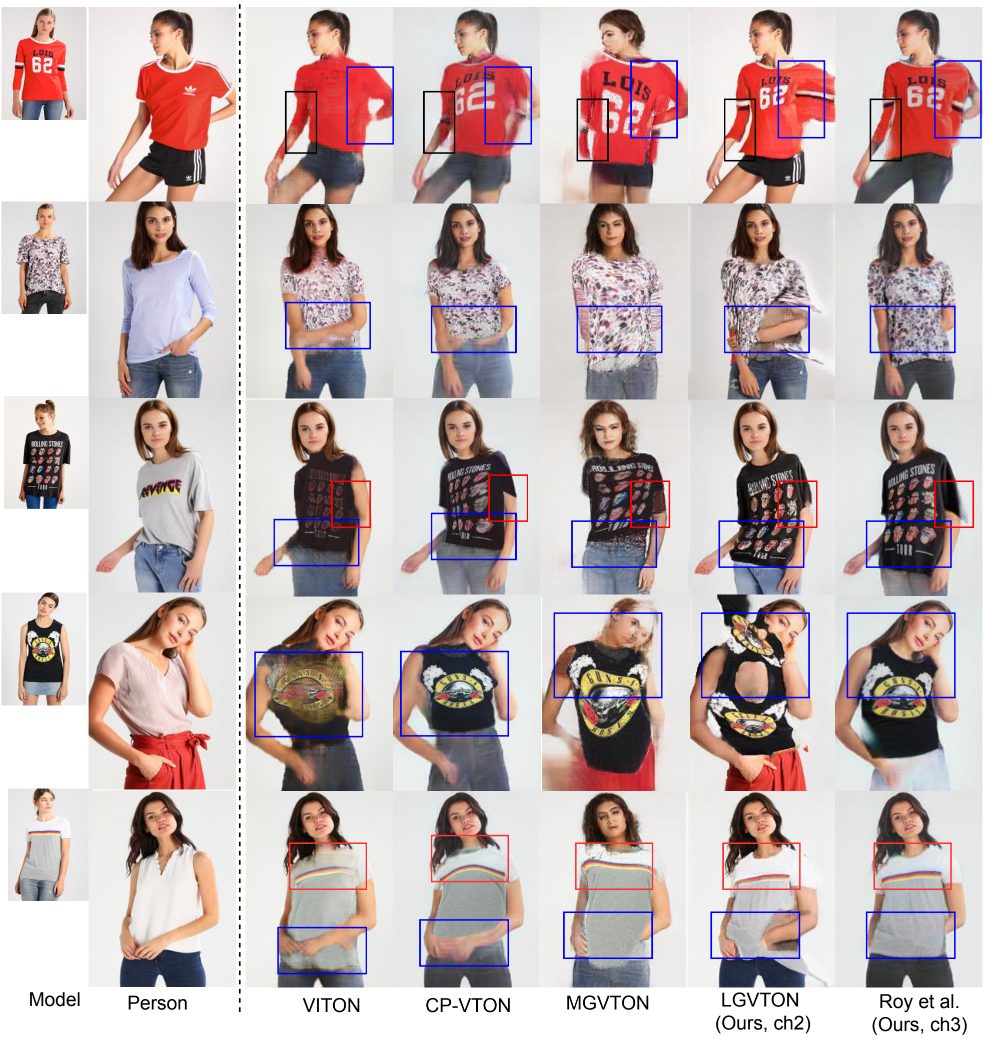


Figure 3.3: Comparative study on MPV dataset. Significant details are zoomed in and shown right after each output.

upper, lower, full-body dresses).

3.3.3 Implementation Details

Here we elaborate on the network architecture of our geometric matching network shown in Fig. 3.1. This network contains 3 trainable components - the two feature extraction CNNs which are of the same structures and a regression CNN. The feature extraction CNN contains 6 convolution layers each followed by a relu activation layer. The first 4 convolution layers contain 64 filters each with a 4×4 kernel and stride 2. The rest of the convolution layers contains 512 filters with the kernel of size 3×3 and stride of 1. The last layer is an L2 feature normalization layer as proposed in Wang et al.

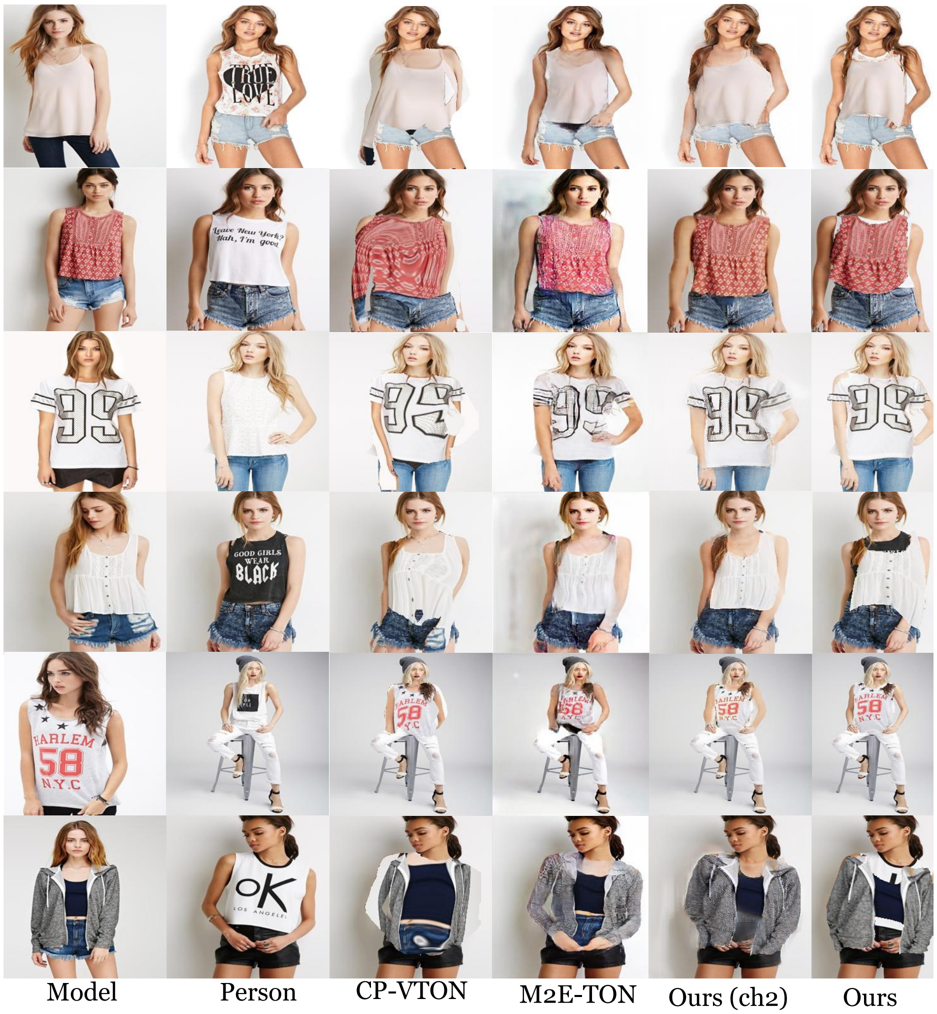


Figure 3.4: Comparative study on MPV dataset. Significant details are zoomed in and shown right after each output.

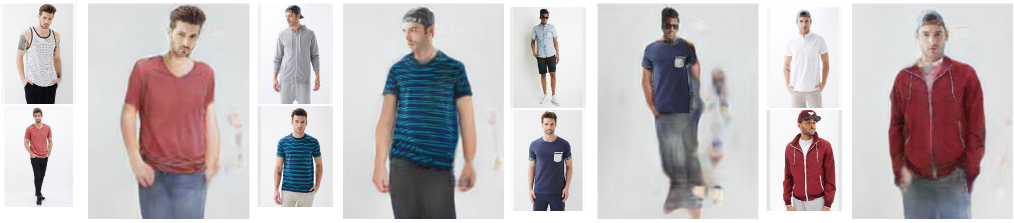


Figure 3.5: Results on DeepFashion dataset.

(2018a). The first 3 layers of the regression CNN are convolution layers with the number of filters 512, 256, and 128 respectively where the first two layers have kernels of size 4×4 and stride of 2. The third convolution layer has the kernel of size 3×3 and stride of 1. These layers are followed by a relu activation layer and a dense layer of 64 nodes with relu activation. This is followed by 2 consecutive dense layers of 18 and 50 nodes respectively with tanh as the activation function.

Number of parameters: We report a cost analysis of different methods in terms of the number of parameters in Table. 3.2. Compared to VITON, CP-VTON, and, MGVTON we achieve a better parameter count (P_1^{warp}) in the warping stage. However, our geometric matching network (GMN) is costlier compared to the PGWM of our previous method.

Table 3.2: Comparing the number of parameters (in millions) of different methods. The total number of parameters in the proposed method is given in column P_1. The number of parameters in the warping stage and the rest of the stages are reported in columns P_1^{warp} and P_1^{rest} respectively. The parameters required to compute different inputs are specified in the columns under P_2. The sum of the values reported in P_1 and P_2 are reported in the column Total. We used the following abbreviations for the different input estimation methods, HP - Human parsing, PE - Pose estimation, DPE - Densepose estimation. The best and the second-best results are highlighted with bold notation and blue color respectively.

Methods	P_1^{warp}	P_1^{rest}	P_1	P_2				Total
				HP	PE	DPE	Total P_2	
VITON	29.26	0.076	29.34	75.65	52.31	-	127.96	**157.30**
CP-VTON	19.06	21.34	40.40	75.65	52.31	-	127.96	168.36
MGVTON	28.97	195.49	224.46	75.65	52.31	-	127.96	352.42
Ours (ch2)	**1.69**	0.76	**2.45**	75.65	52.31	59.73	187.69	190.14
Ours	9.82	21.34	31.16	75.65	52.31	59.73	187.19	218.35

3.3.4 A visual analysis of the features learned

We study the 512 filters of kernel sized 4 × 4 from the first convolution layer of the Regression CNN of our GMN. This layer operates directly on the output of the correlation layer i.e., the matching layer. As shown in Fig. 3.1 our GMN contains 2 Feature Extraction CNN each taking the densepose of the model and the person as input respectively. Without loss of generality let us denote the outputs of these two Feature Extraction CNNs as F_A and $F_B \in \mathbb{R}^{(12\times16\times512)}$ respectively. The correlation layer of GMN computes all possible pairwise dot products of its inputs F_A and F_B. We denote the output of the correlation layer i.e., the correlation map as $C_{AB} \in \mathbb{R}^{(12\times16\times192)}$. The computation of correlation map is illustrated in Fig. 3.6 (i). Each 1D slice at any location (i, j) through this correlation map sized

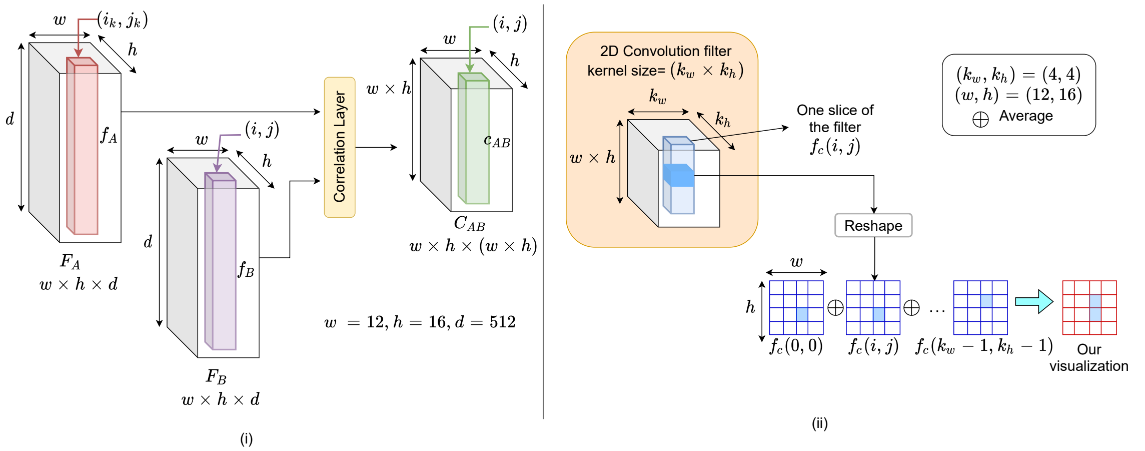


Figure 3.6: (i) Computation of the correlation map on the learned features, (ii) Processing of the convolution filter of the first layer of the regression CNN for visualization purposes.

$(1 \times 1 \times 192)$ (say, c_{AB}) contains the pairwise dot product of F_B at (i, j) denoted by f_B with all such slices in F_A. The depth 192 in c_{AB} represents the combination with 192 (i.e., 12×16) 1D slices in F_A.

Now, we study the 512 filters of kernel sized 4×4 from the first convolution layer of the Regression CNN. Each convolutional kernel of these 512 filters is of size $4 \times 4 \times 192$. The kernel depth 192 is the same as that of the correlation map as these filters work directly on the correlation map. We first normalize each of these filters. Then, we take each $1 \times 1 \times 192$ 1D slice through the channels of one convolutional filter at a particular spatial location (say, $f_c(i, j)$) and reshape this to 16×12 image. This image shows the filter's preference for the association between the particular feature in image B to that of the features in image A. Now, if we take the maximum of the filter values for such an image and visualize it, we get to see an image with one peak at some location and all other values as zero. For all the other slices corresponding to one filter i.e., the other 15 slices in our (4×4) filter we can get the peaks in a similar fashion. Now, if we get these peaks in a localized region then this implies that the filter responds strongly when spatially co-located features in image B match with spatially persistent features in image A.

In order to see what happens in our filters, we pick the maximum weight from each of the images corresponding to the 1D slice for a filter and average them together to produce a single image. A visual illustration of this computation is shown in Fig. 3.6 (ii). Note that, we highlighted one location of $f_c(i, j)$ with blue, to denote the maximum value. With this visualization, we get 512 images for 512 filters. Note that, the pixel values in these images represent the convolutional filter's preference to the matches of features in F_A with f_B. Among the 512 filters, we randomly selected (without replacement) 20 filters for visualization. These filters are shown in Fig. 3.7. Here for visualization, we used the matplotlib library of python and yellow indicates the highest filter weight.

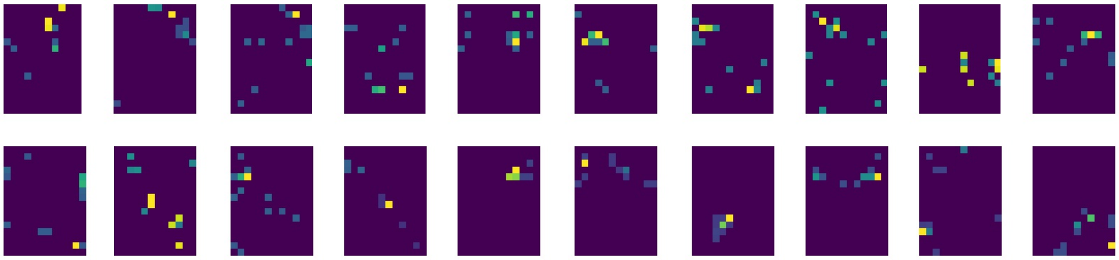


Figure 3.7: Visualization of filters from the first layer of the regression CNN of the geometric matching network of ours which acts directly on the output of the correlation layer.

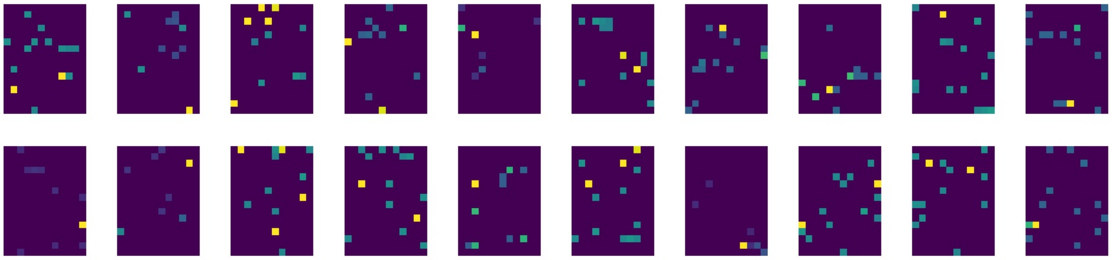


Figure 3.8: Visualization of filters from the first layer of the regression CNN of the geometric matching network of CP-VTON.

From Fig. 3.7 we see that the weights of most of the filters are highly localized. This indicates that these filters specialize in detecting matches in specific positions in image A (associated with F_A). In other words, this localization indicates that the filters prefer when spatially co-located features in image B correlate well with the spatially persistent features of image A. Thus based on our previous elaboration of local neighborhood consensus, we see that our network has learned to mimic local neighborhood consensus for robust match estimation. We also see that the neighborhood size is similar in all the filters. We believe, that since we have used densepose representation as inputs that are consistent for both the model and person images hence we are getting very localized filter responses. We also show a similar visualization of the same filters from the geometric matching network (GMN) of CP-VTON in Fig. 3.8. Notice that the filter weights are more localized in our network compared to that of CP-VTON. This also explains the improved performance of our approach compared to

CP-VTON.

3.4 Discussion

In this chapter, we propose a cloth warping method for the image-based model to person virtual try-on (VTON) problem. In this context, we employ a neural network-based image matching network called the geometric matching network, which is a trainable equivalent of classical feature matching methods. The network learns the TPS transformation required to transform the clothing of the model to fit the shape and pose of the person. A comparative study with state-of-the-art methods shows the potency of our method in producing precise warps. Our approach shows notable improvement in handling pose variability and is also invariant to clothing texture variation.

However, this method has a few limitations. While the method proposed in this chapter can handle bending and folding of arms or pose variability between the model and the person better compared to other methods in many cases, it fails to achieve photo-realistic results. This problem becomes severe in the case of long-sleeved outfits. Then, for crossed arm postures, the overlap between different clothing parts might happen. The existing approaches including the proposed approaches till this part of the thesis, especially the warping-based methods employing *thin plate spline (TPS)* transform, cannot tackle such cases. For instance, observe the results shown in Fig. 3.9. The formulation of 2D TPS transform is such that it can handle only in-plane organization of points. The overlap between different parts of the plane indicates some 3D displacement of the grid points. Hence, TPS cannot model such cases. We investigate this issue more elaborately in our next chapter and propose a solution to this problem.

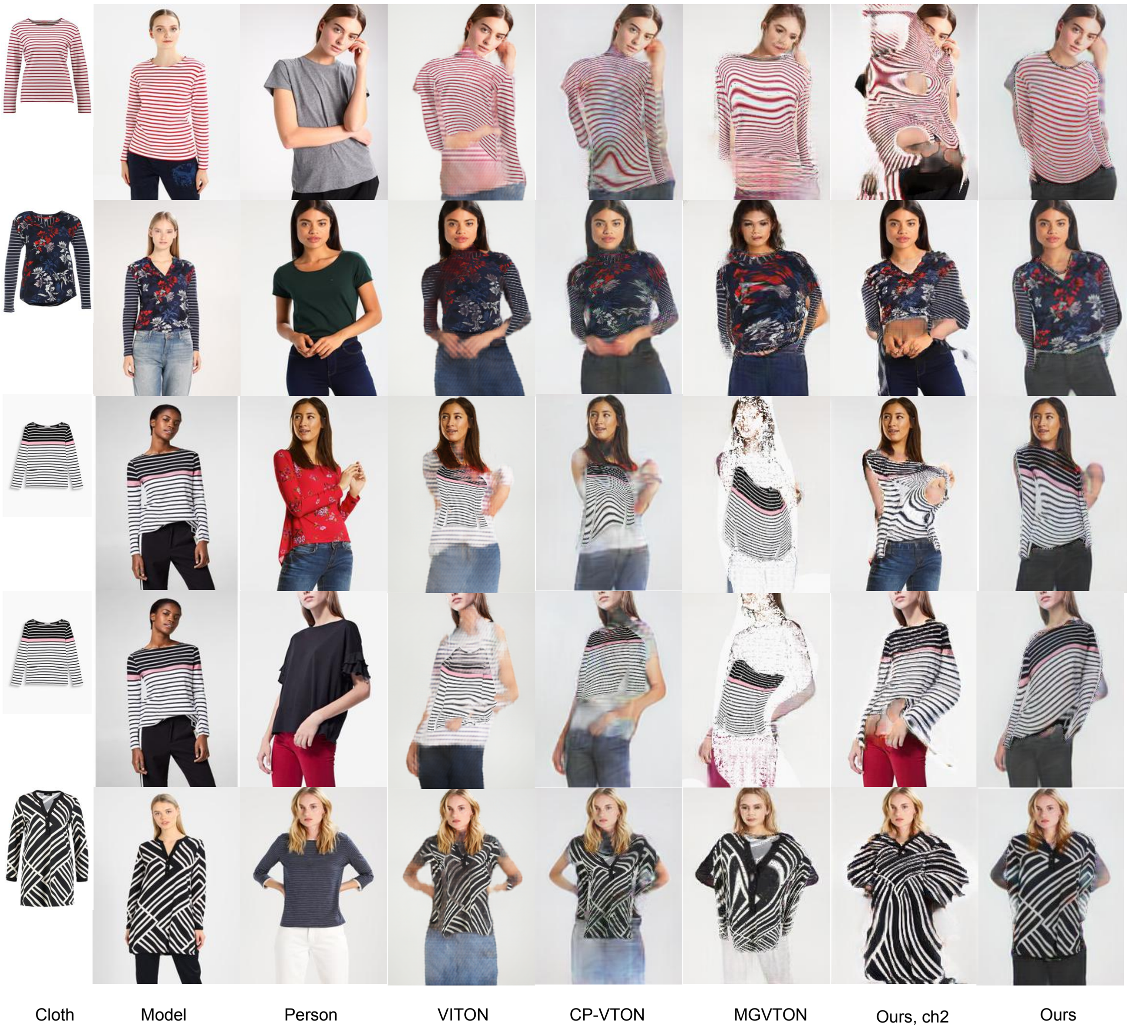


Figure 3.9: Limitation of ours as well as other approaches in handling pose variability, especially in case of folded or bent arm postures.

Chapter 4

Employing structural key points with an improved method for clothes warping

4.1 Introduction

As discussed in Chapter 1 the problems with cloth-to-person (C2P) VTON, briefly speaking, are related to available data as the training phase requires both the image of the clothing (C) and that of a model (M) wearing that clothing (Wang et al., 2018a; Han et al., 2018; Yu et al., 2019; Dong et al., 2019a; Yang et al., 2020). Moreover, in general, these clothing images are taken in flat backgrounds and are placed in a way that resembles the anatomical pose of humans. Such images make the formulation of the problem comparatively easier, without any issue of occluded clothing parts. Here we consider a more challenging version of the VTON problem, where the clothing from the model is transformed to the target person. This version is called the model-to-person (M2P) VTON problem. In the last two chapters, we have proposed two different solutions to this problem. However, the most challenging aspect of the M2P is that the model (M) can be in any pose resulting in significant occlusion. Therefore, a solution that is robust in terms of pose variation between M and P is crucial here. The solutions proposed in Chapters 2 and 3 have addressed this for simple to moderate pose variability but fail to handle the postures with folded or crossed arms along with occlusion. That means these algorithms could not handle the cases when parts of the model's clothing is occluded.

It is already mentioned that employing thin plate spline transform (TPS) as the warping function is common among the warping-based VTON methods (Han et al., 2018; Wang et al., 2018a; Dong et al., 2019a; Yu et al., 2019; Roy et al., 2020, 2022; Yang et al., 2020). Given a set of source and corresponding target control points TPS transform computes an interpolation function $\mathbb{Z}^2 \rightarrow \mathbb{Z}^2$ by minimizing a second-order smoothness constraint, called bending energy. Some VTON methods Wang et al. (2018a); Dong et al. (2019a); Yu et al. (2019); Yang et al. (2020); Roy et al. (2020) employed a deep neural network based geometric matching network (GMN) to predict the TPS transformation to be applied directly on the input images, without explicitly specifying the source and target control point pairs. Our previous method proposed in Chapter 2 opted for the landmark-based image registration (Maintz and Viergever, 1998). It computes the TPS transform using explicitly specified source to target control points correspondences. However, due to the smoothness constraint of TPS, both types of methods often produce inconsistent results when the target warp requires significant bending, which mainly occurs due to the flexible arm movement of humans. For instance observe the results (a) - (d) in Fig. 4.1.

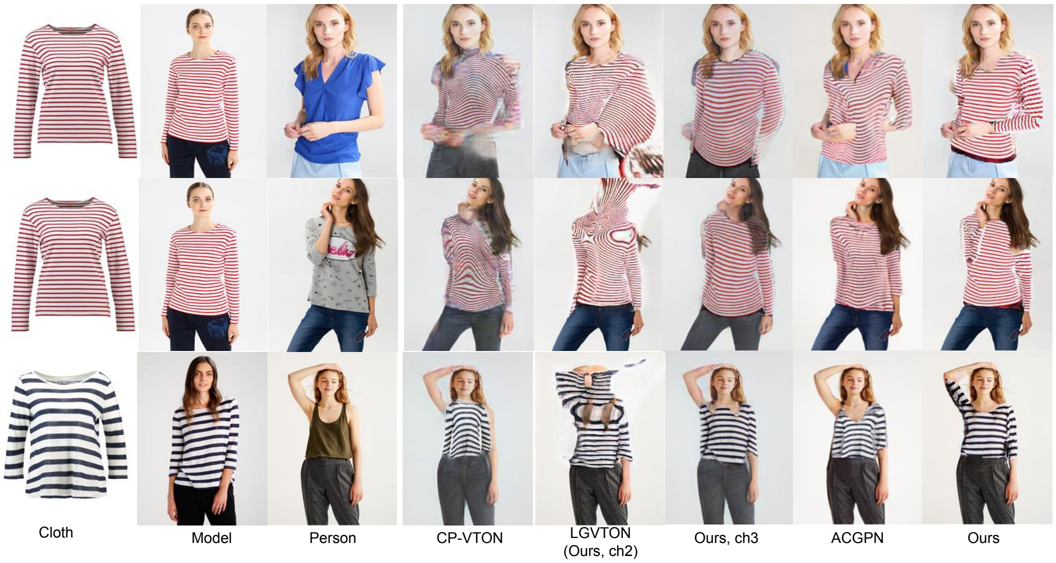


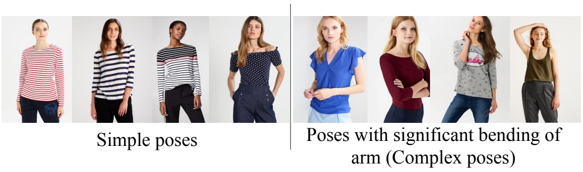
Figure 4.1: Results of different methods for different input clothing, model, and person combinations.

The formulation of TPS models some rearrangements of control points in the image plane; therefore, it is not fit for applications that require modeling of overlap or fold (Bookstein, 1989). For instance, the results (a) - (d) shown in the top two rows of Fig. 4.1 demonstrate when the arms of the person P cross her torso requiring the sleeve to overlap on the torso. The TPS-based warping methods (Wang et al., 2018a; Roy et al., 2022, 2020; Yang et al., 2020) fail to compute consistent target warp in such cases.

Some GMN based methods (Wang et al., 2018a; Dong et al., 2019a; Yu et al., 2019) often produce undesirable grid deformations, which mainly occur due to the inability to learn the constraints of human anatomy. For example, observe the results of CP-VTON (Wang et al., 2018a) in the first two rows of Fig. 4.1. ACGPN proposed to employ a second-order difference constraint to control the grid deformations. But, that is more of an avoidance measure than prevention. On the other hand, LGVTON showed that if structural key points (e.g., human landmarks, demonstrated in Fig.4.3) are considered as control points in computing the warps, this problem does not arise. It suggests that considering some anatomically meaningful control points, at least, implicitly imposes constraints to keep the target warp consistent with the feasible range of human body movements. In Chapter 3 we followed this idea. However, instead of using structural key points, we used the correspondences of a body shape representation, namely, densepose (Alp Güler et al., 2018) between M and P and learned the TPS transform by employing a GMN. Therefore, here we try to take the benefit of both anatomy-based and GMN based ideas. Though the method proposed in Chapter 2 is strictly restricted to simple human poses only, it showed comparatively better flexibility in terms of the pose. However, it is still not free from the limitations due to employing TPS. Some simple and complex human poses are shown in Fig. 4.2.

Similar to the method proposed in Chapter 2, M2E-TON (Wu et al., 2019) and Liu et al. (Liu et al., 2021) also used the idea of body shape correspondences in model-to-person VTON, but differently. Liu et al. proposed a coordinate-prior map leveraging the partial UV texture map obtained from densepose representations of M and P. In addition, they proposed a spatial-aware texture generation network to complete the partial UV texture and thus infers the unobserved clothing appearance. Compared to M2E-TON, Liu et al.'s solution can handle cases of large pose variability between M and P and also the texture-variability of clothes. However, in the case of complex textures on the clothing, e.g., stripes and floral, it falls short in preserving the details. In addition to the above-mentioned approaches, the problem of VTON is also explored using GAN-based (Goodfellow et al., 2014) attribute manipulation approaches (Jetchev and Bergmann, 2017; Yildirim et al., 2019; Men et al., 2020; Lewis et al., 2021). While these methods often meet the standards of photo-realism in many cases, they still fail to retain the exact texture patterns of the source clothing in the try-on output.

Considering all these, in this chapter, we attempt to achieve a pose robust model-to-person (M2P) VTON solution without requiring any paired training data, i.e., images of the same person with and without wearing the model's clothing. To achieve a robust solution that can handle pose variability our first idea is to divide the source clothing into parts i.e., sleeves and torso, followed by warping each of these parts separately, and then finally combining them. This approach attempts to solve the


Simple poses

Poses with significant bending of arm (Complex poses)

Figure 4.2: Demonstration of simple and complex human poses.

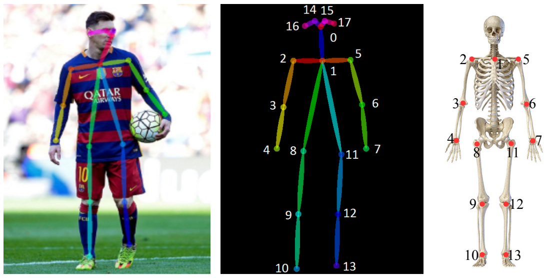


Figure 4.3: Demonstration of human landmarks. Best viewed in electronic version.

overlap issue. Moreover, it opens the flexibility of employing clothing part-specific transformation. Here also we adopt LGVTON's idea of using landmark-based image registration for warping for its effectiveness and applicability in our formulation. Second, similar to LGVTON we employ TPS transform as the warping function, However, this transformation is applied only on the torso part as the human torso undergoes restricted as well as smooth deformations. Since the human arm can move in various ways causing significant deformations in the sleeves, in most cases, they cannot be modeled by TPS transform due to its bending constraint. Moreover, each arm contains only 3 human landmarks (Fig. 4.3), which is insufficient for estimating TPS transform (Bookstein, 1989). So, we propose a hand-crafted feature-based warping technique. Our idea is inspired from a popularly known warping method called *field warping* which was proposed by *Beier and Neely* (Beier and Neely, 1992). Instead of landmark correspondences, here we consider the correspondences of the straight line segments between the consecutive landmarks. This imposes an additional constraint that conforms to human anatomy (see Fig. 4.3). Keeping the perspective similar we propose some major methodological modifications to the original idea of the field transform to make it applicable in the current problem context along with addressing some of its limitations.

Note that we compute the target warp from the non-occluded clothing areas only. Therefore, the occluded areas of the source need to be interpolated, if exposed in the target image. We propose a *mask prediction network (MPN)* that predicts the target clothing mask, referring to the region of the expected try-on output containing the model's clothing. MPN learns to incorporate the correlation between the necessary features of M and P, which empirically shows improvement over ACGPN's semantic generation module with a similar objective. The output of MPN helps in distinguishing the occluded areas of the target clothing. The proposed *image synthesizer network (ISN)* interpolates these occluded regions to produce a seamless try-on image. MPN also adds to the faster computation of the target warp.

Before presenting our proposed method, we discuss some preliminary ideas for a better understanding of the former.

4.2 Preliminaries

Here we discuss some preliminary ideas. First, we discuss the idea of forward and backward mapping. Then we briefly explain the Beier and Neely warping method (Beier and Neely, 1992) and its limitations to aid the reader in understanding the motivation behind our revised formulation.

A warp may be considered as a mapping $T : \mathbb{Z}^2 \rightarrow \mathbb{Z}^2$ relating each point in the source image to the corresponding point in the target image, or vice versa. In more detail, it can be written in the following form

$$(x,y) = \mathscr{F}_{bwm}(u,v) = (\mathscr{X}(u,v), \mathscr{Y}(u,v)), \tag{4.1}$$

$$(u,v) = \mathscr{F}_{fwm}(x,y) = (\mathscr{U}(x,y), \mathscr{V}(x,y)), \tag{4.2}$$

where $(x,y) \in \mathbb{Z}^2$ refers to the target image coordinate corresponding to the source image coordinate $(u,v) \in \mathbb{Z}^2$. $\mathscr{X}, \mathscr{Y}, \mathscr{U}, \mathscr{V}$ refer to the arbitrary mapping functions that specify the spatial transformation. Since X, Y map from the source to the target, they are referred to as *forward mapping*. Likewise, U, V are referred to as *backward mapping*. In practice, backward mapping is preferred over forward mapping as it relates every pixel in the output to its corresponding input pixel through closeness.

Feature based warping algorithm of Beier and Neely.

This algorithm defines a mapping from one image into the other based on a pair of corresponding straight-line segments, one relative to the source image and the other relative to the destination. This algorithm relies on backward mapping, where each grid position in the target image gets mapped to a corresponding pixel in the source image.

This algorithm was proposed as a method for morphing, which warps and cross-dissolve one digital image into another. Based on the scope of this paper, we restrict our discussion to the warping part only. For a pair of given line segments PQ, $P'Q'$ belonging to the target and the source image respectively, a coordinate mapping from the target image pixel X to the corresponding source image

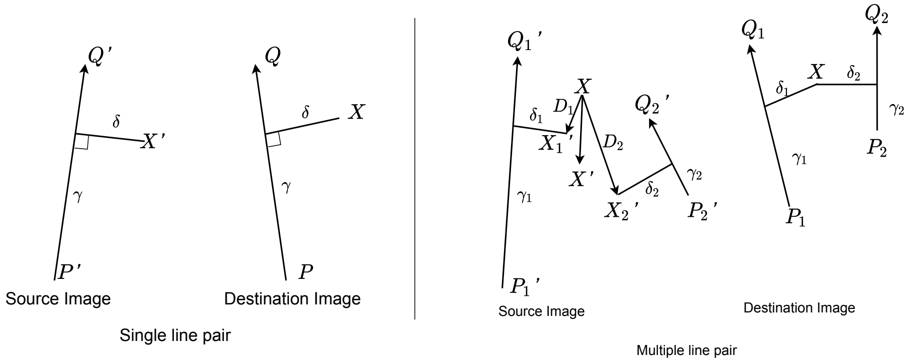


Figure 4.4: Beier and Neely transformation for single and multiple line pairs.

pixel X' is computed as,

$$X' = P' + \gamma.(Q' - P') + \frac{\delta.Perpendicular(Q' - P')}{\|(Q' - P')\|}, \tag{4.3}$$

$$where$$

$$\gamma = \frac{(X-P).(Q-P)}{\|(P-Q)\|^2}, \tag{4.4}$$

$$\delta = \frac{(X-P).Perpendicular(Q-P)}{\|(P-Q)\|}. \tag{4.5}$$

$\gamma \in \mathbb{R}$ refers to the distance of the pixel along the line and $\delta \in \mathbb{R}$ refers to the perpendicular distance from the pixel to the line. For a visual clarification please refer to the single line pair case shown in Fig. 4.4. In presence of multiple such line pairs, a weighted combination of the locations of X's with respect to each line pair is computed. The weight assigned to the coordinates of a pixel for a line pair is computed as

$$weight_{BN} = \left(\frac{length^p}{a + dist}\right)^b, \tag{4.6}$$

where '$dist$' refers to the shortest distance from the pixel to the corresponding line. a, b, p are the constants and '$length$' is the length of the line PQ. An example of a multiple line pair case is shown in Fig. 4.4.

Over the years field warping has been appreciated for its expressiveness compared to mesh warping. Also unlike TPS, this algorithm has no limitations on the number of line segments considered. However, for two or more line segments, while the algorithm tries to guess what should happen far away from the line segments, sometimes it makes a mistake (Beier and Neely, 1992; Lerios et al., 1995). This occurs due to the weighted influence of the lines in the adjoining region of the line segments. As a result, unexpected interpolations are generated. The authors called this problem 'the ghosting problem'. Some examples of this can be observed in Fig 4.5. In addition, Beier et al. mentioned the time complexity of this transformation which increases with the number of lines considered.

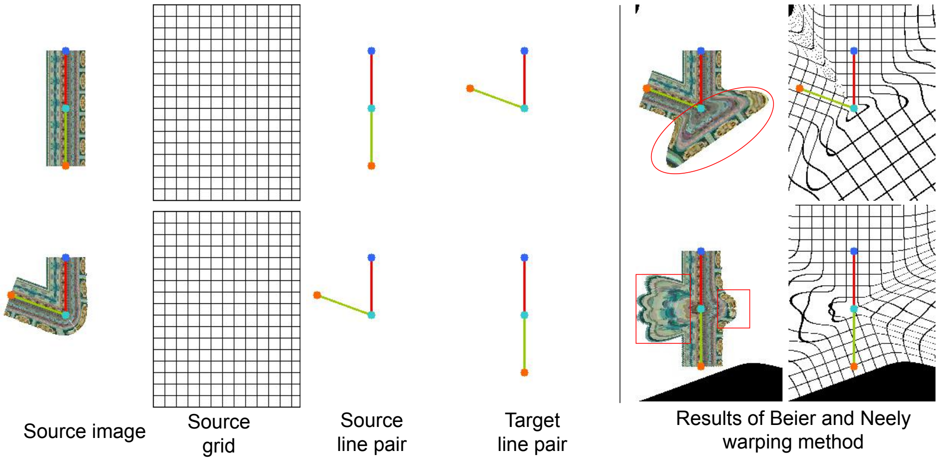


Figure 4.5: Demonstration of the ghosting problem of the Beier and Neely warping method. Areas showing the ghosting problem are highlighted.

4.3 Proposed method

4.3.1 Overview

We compute P' in mainly three steps. First, We employ a mask predictor network (MPN) (Sec. 4.3.2) that predicts the target clothing mask, denoting the trialed clothing region of the expected try-on output. Second, we extract the model's outfit c using a human parsing method (Gong et al., 2017)[1] and compute the target warp c' that fits the person P, using our modified Beier and Neely warping method (Sec. 4.3.3). Third, we propose a novel synthesis step employing an image synthesizer network (ISN) that seamlessly combines c' with P to generate the final try-on output P' (Sec. 4.3.4). Below we discuss all the steps in detail. A simplistic block diagram demonstrating the different steps of our approach is shown in Fig. 4.6.

[1] Human parsing is the task of segmenting a human image into different fine-grained semantic parts e.g., head, torso, arms, legs, upper-clothes, etc.

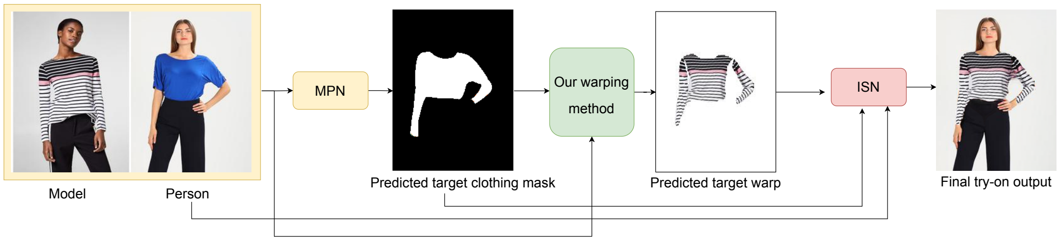


Figure 4.6: Illustration of our overall virtual try-on approach.

4.3.2 Predicting the target clothing mask

This step predicts the target clothing mask serving two purposes in the subsequent stages of this work: (1) it aids the synthesis module to identify the occluded regions of the source clothing which needs to be interpolated to produce a seamless try-on output, (2) it helps our proposed warping method in reducing the time of computation of the target warp corresponding to the sleeves of the source clothing. This will be elaborated on later during the discussion of the warping method.

To predict the mask, we train a convolutional neural network, named the mask prediction network (MPN), with the following inputs - (1) the underclothing body shape of the model and the person, encoded using their densepose representations (Alp Güler et al., 2018) [2]. In addition, we also provide the face and head segments of both the model and the person extracted using (Gong et al., 2017). (2) the mask of c, and (3) semantically segmented human parts (Lin et al., 2020) of the model. A block diagram of MPN along with the inputs and output are demonstrated in Fig. 4.7. Note that providing the face and head segments along with the densepose as input gives the whole detail of a human body. Whereas, a more refined level of details is provided by the segmented human parts.

The architecture of MPN is demonstrated in Fig. 4.7. We incorporate correlation layers in MPN that computes the linear relationship between the body shape and pose features of the model and that of the person (Observe Fig. 4.7). The idea of incorporating such a layer is inspired by our method proposed in Chapter 2, where it is shown to improve results. We added an extra-human parsing branch at the end of this module (demonstrated in Fig. 4.7). The objective of this is to predict the human parsing of P' given the non-target clothing details of P e.g., the lower body clothing of P, the face and hair segments of P, and the predicted mask from the previous branch. This is enforcing another constrain on the main branch to predict a mask consistent with the final human parsing. We

[2] A densepose representation maps all human pixels of an RGB image, to the 3D surface of the human body, thus providing a precise estimate of the human body shape under the clothing.

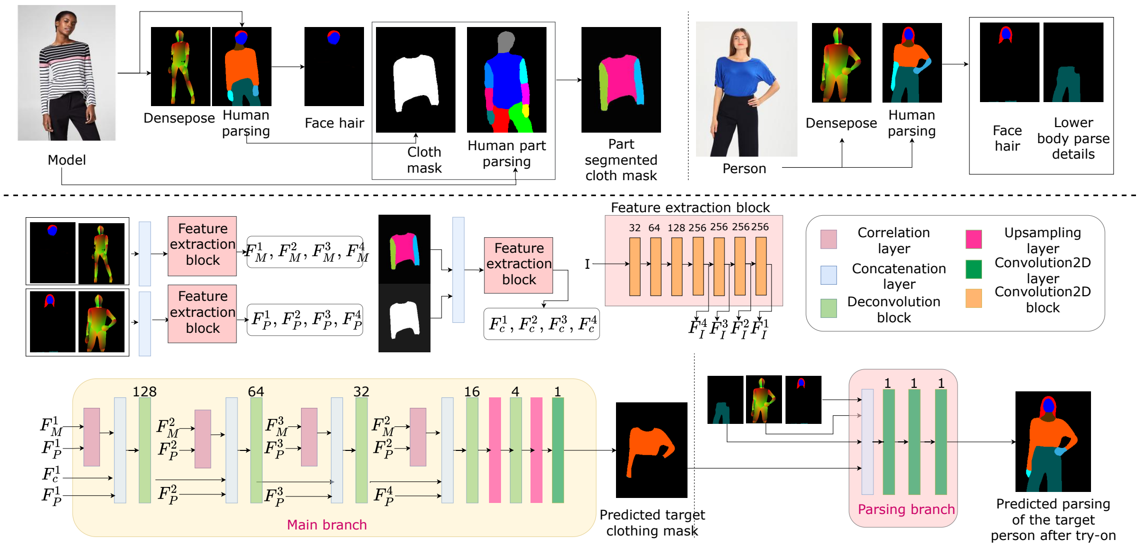


Figure 4.7: Block diagram of the proposed mask predictor network (MPN). Please see the electronic version for better view.

empirically show the effectiveness of this branch in Sec. 4.4.

4.3.3 Predicting the target warp

This step predicts c' from c. We propose a part-by-part warping method where we first segment c into 3 semantic parts - torso (c_{torso}), left sleeve ($c_{lsleeve}$), right sleeve ($c_{rsleeve}$), using the human part segmentation method proposed in (Lin et al., 2020) (a result of (Lin et al., 2020) is given in Fig. 4.9). Our objective is to compute the corresponding target warps c'_{torso}, $c'_{lsleeve}$, $c'_{rsleeve}$. A block diagram of this step is presented in Fig. 4.8.

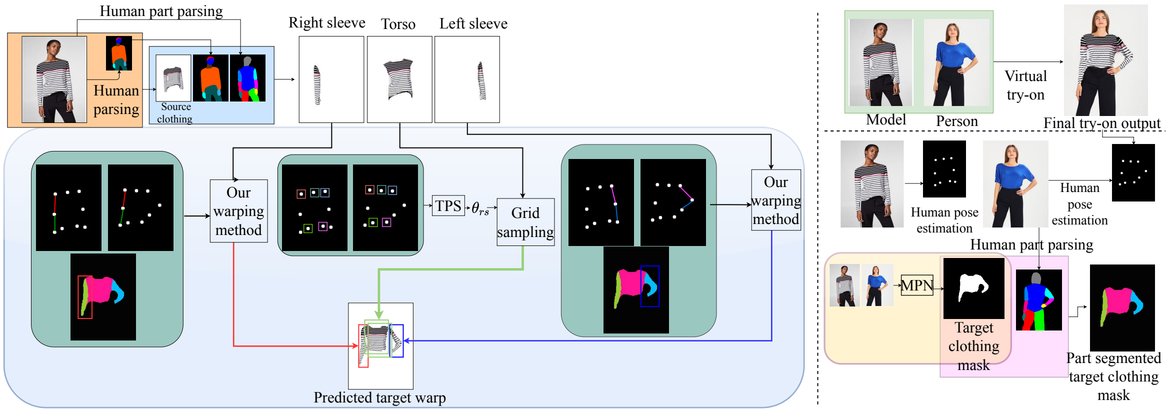


Figure 4.8: Illustration of our warping method.

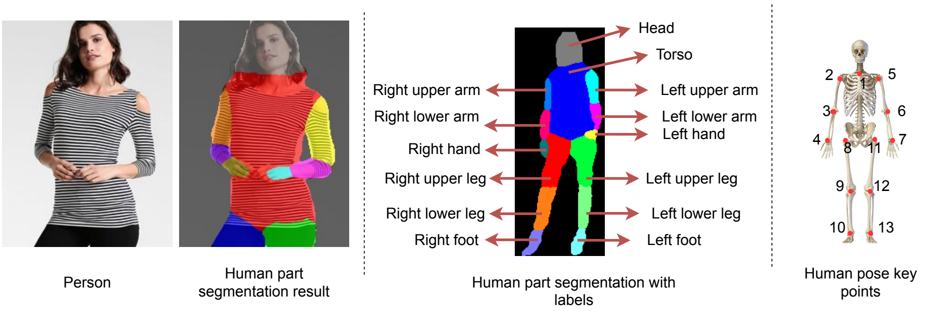


Figure 4.9: Part segmentation results of human part parsing method (Lin et al., 2020) and demonstration of pose key points. Best viewed in electronic version.

To compute c'_{torso} from c_{torso} we partially adopt the landmark-based image registration (Maintz and Viergever, 1998) method proposed by LGVTON that computes TPS transformation function from the correspondences of some kind of structural key points between M and P. While LGVTON used fashion landmarks as well as human landmarks, but, we use human landmarks only to cut the cost of annotations [3]. However, the limitations due to this are later taken care of by our synthesis network. Unlike LGVTON as we are computing the target warp corresponding to the torso only, therefore, we used five human landmarks localized on the torso of humans for this step. This can be visually verified from Fig. 4.8. Notice the landmark correspondences, demonstrated by matching colors.

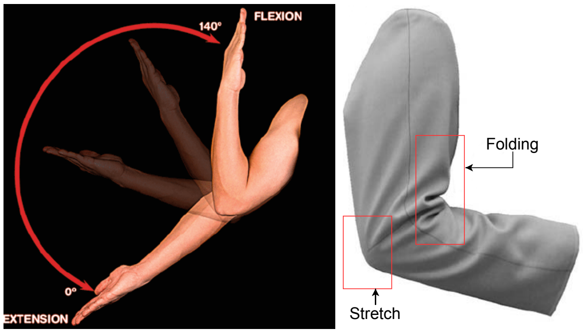


Figure 4.10: (Left) Elbow flexion and extension (picture courtesy (Malagelada et al., 2014)), (right) stretch and folds in clothing sleeve due to elbow flexion i.e., arm bending (picture courtesy (Masteikaitě et al., 2014))

.

In the following part we discuss our warping method that we adopt to warp the sleeves, $c_{lsleeve}$ and $c_{rsleeve}$. Warping a part of clothing is influenced by only the landmarks of the corresponding region of the human body. Therefore, to warp the sleeves we consider only the three landmarks localized

[3] Annotating fashion landmarks is a difficult task considering the variety of clothing types, textures, and colors compared to the task of localizing human landmarks.

on each of the arms. Since Beier and Neely warping is a line transform, hence, unlike TPS here we consider the correspondences of the line segments joined by two consecutive landmarks. As shown in Fig. 4.9 the line segments that we are considering are - (2, 3), (3, 4) (for right arm) and (5, 6), (6, 7) (for left arm). Landmarks 3, 6 represent the location of the right and left elbow joints of a human. Observe in Fig. 4.9 the line segments we are considering are anatomically meaningful also e.g., (2, 3) and (5, 6) can be considered equivalent to the humerus bone, and, (3, 4) and (6, 7) are equivalent to the combined representation of radius and ulna bones. Consideration of human anatomy imposes additional constraints on the warping process leading towards a more justified warping. Based on this, geometrically we can represent an arm as 2 line segments connected at a common endpoint. Let us denote an arm of a person by the line pair (BA, BC) and the corresponding arm of the model by the line pair ($B'A'$, $B'C'$). Given this, we define a coordinate mapping from any arbitrary pixel X of P (target image) to the corresponding pixel X' of M (source image) in compliance with the concept of backward mapping. A demonstration is given in Fig. 4.11.

Now, let us introduce some notations required to elaborate our approach. Considering, ϕ_1 = $\angle XBA$ and ϕ_2 = $\angle CBX$, and $r = BX$, the polar coordinate of X relative to the line BA and BC are (r, ϕ_1) and (r, ϕ_2) respectively. Similarly, considering ϕ_1' = $\angle X'B'A'$ and ϕ_2' = $\angle C'B'X'$, and $r' = B'X'$, the polar coordinate of X' relative to the line BA and BC are (r', ϕ_1') and (r', ϕ_2') respectively. Here, $\phi_1, \phi_2 \in (-\pi, \pi)$ and ϕ = $\angle CBA = \phi_1 + \phi_2$. $\angle X'B'A'$ = ϕ_1' and $\angle C'B'X'$ = ϕ_2'. Similarly, $\phi' = \angle C'B'A'$ = ϕ_1' + ϕ_2'. Note that we compute ϕ_1 and ϕ_2 in such a way that the signs of them are the same. We assume that ϕ_1', ϕ_2' also have the same sign and compute them accordingly. We use capital letters to denote the vectors and small letters for scalars. Here, $A, B, C, A', B', C', X, X'$ are all 2D-vectors. Now given A, B, C, A', B', C' our objective is to find $X'(r', \phi_1')$ corresponding to each $X(r, \phi_1)$. Below, we first discuss our method to compute the angular coordinate ϕ_1' followed by discussing our idea of computing the radial coordinate r'. Note that computing the coordinate relative to any of the lines is consistent with our idea, however, without loss of generality, we considered here line BA only.

Anatomically, the bending of the arm is called *elbow flexion* as shown in Fig. 4.10. The amount of bending can be represented quantitatively using the *flexion angle* which is 0° when there is no bending and may increase up to a maximum of 145° (Islam et al., 2020) as we bend our arm. When we bend our arm the sleeve around each of the upper and lower parts of the arm aligns being influenced mainly by the part it is associated to. For example, keeping our upper arm fixed if we move our lower arm, the upper part of the sleeve will not show any significant movement. This implies that due to bending

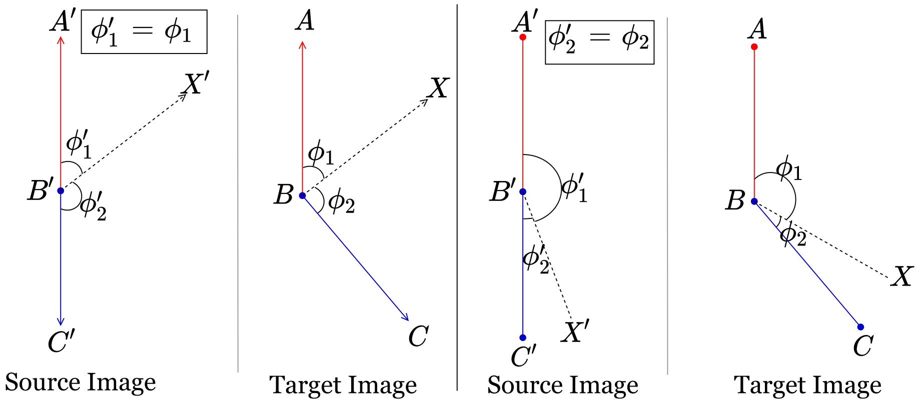


Figure 4.11: Geometrical illustration of our warping method for sleeves warping.

the relative position of the pixel X with its closest line remains unchanged. Remember, this part of the discussion is limited to the angular position only. Considering this we include our observation in this method in the form of two assumptions,

- Assumption 1 - when X is closer to the line BA in terms of angular distance i.e., when $|\phi_1| < |\phi_2|$, the relative angular position of X with respect to line BA which is ϕ_1, will be equal to the angular position of X' with respect to line $B'A'$ which is ϕ_1', i.e., $\phi_1 = \phi_1'$.

- Assumption 2 - in a similar sense, when X is closer to the line BC i.e., when $|\phi_2| < |\phi_1|$, the relative angular position of X with respect to line BC which is ϕ_2, will be equal to the angular position of X' with respect to line $B'C'$ which is ϕ_2', i.e., $\phi_2 = \phi_2'$. Now since we represent the coordinate of X related to line BA only therefore $\phi_2 = \phi_2'$, will be equivalent to saying $\phi_1' = \phi' - \phi_2$.

Let us now understand the effect of elbow flexion on the sleeve area near the elbow joint. Due to bending of the arm, folding and stretching of the sleeve occurs near the elbow joint which is shown in Fig. 4.10. The sleeve gets stretched based on the clothing material near the outer side around the elbow. Whereas in the inner part of the elbow due to bending the clothing gets folded [4]. While the phenomenon of bending is consistent with our previous assumptions, but, stretching occurs by the relative pose of the upper and lower arms. Hence, considering a weighted combination of the positional influence of each of the upper and lower arms leads to a realistic warp of the sleeve here. This is technically realized by a weighted combination of the effects of both the assumptions, where

[4] Note that, we do not consider the case of cloth wrinkling in this work, to keep the idea simple for now.

the weight is computed by function h. Therefore to combine the idea of folding and stretching both in one expression, we compute ϕ_1' as,

$$\phi_1' = \phi_1(1 - h(\phi_1, \phi_2)) + (\phi' - \phi_2)h(\phi_1, \phi_2). \tag{4.7}$$

The value of $h(\cdot,\cdot)$ lies in the range $[0,1]$ and it represents the smooth transition effect while satisfying both the assumptions. Based on our first assumption ϕ_1' should be equal to ϕ_1 when $\angle XBA \ll \angle XBC$, and for the first assumption ϕ_2' should be equal to ϕ_2 when $\angle XBA \gg \angle XBC$. Notice Eq. 4.7 is formulated based on this idea only, where for $h(\cdot,\cdot) = 0$ we get $\phi_1' = \phi_1$ and for $h(\cdot,\cdot) = 1$ we get $\phi_2' = \phi_2$. This implies $\phi_1' = (\phi' - \phi_2)$ considering $\phi' = \phi_1' + \phi_2'$. The values of $h(\cdot,\cdot)$ in between 0 and 1 signifies the weighted influence of the transformations for each of the upper and lower parts of the arm which in fact avoids sharp corners at the elbow joint.

We define h in terms of two functions f and g as

$$h(\phi_1, \phi_2) = g(\phi) f(\phi_1, \phi_2) + (1 - g(\phi)) Round(f(\phi_1, \phi_2)), \tag{4.8}$$

given that

$$\begin{aligned} f(\phi_1, \phi_2) &= \frac{\phi_1^2}{\phi_1^2 + \phi_2^2} \quad \text{and} \\ g(\phi) &= \frac{1}{1 + e^{a(\pi - |\phi|)}}, \end{aligned}$$

where $\phi = \phi_1 + \phi_2$, and $f(\phi_1, \phi_2) \in (0,1]$. We have chosen experimentally $a = 100$. The role of f is to select which assumption will hold for a pixel X. See Fig. 4.12 where the plot of f is shown. Observe that the functional value of f takes smooth transition from 0 to 1 from the region $|\phi_1| < |\phi_2|$ to $|\phi_1| > |\phi_2|$. This smooth transition ensures a softer version of our assumptions is implemented to handle stretching. On the contrary, to impose the idea of folding of the sleeves, which does not require the combined effect of our assumptions, we round-off the function f which ensures a very steep transition from 0 to 1 between the regions $|\phi_1| < |\phi_2|$ and $|\phi_1| > |\phi_2|$. Thus, the warping follows the assumptions strictly for the sleeve areas near the inner part of the elbow when the elbow flexion happens.

While elbow flexion occurs, for the pixels in the inner part, $|\phi| < \pi$ and in the outer part, $|\phi| > \pi$.

Therefore the transition from inner to outer part can be determined by a sigmoid function that varies from 0 to 1 near π. The idea of taking a sigmoid instead of a step function here is to make a smooth transition between the regions $|\phi| < \pi$ and $|\phi| > \pi$ so that a realistic warping result is obtained.

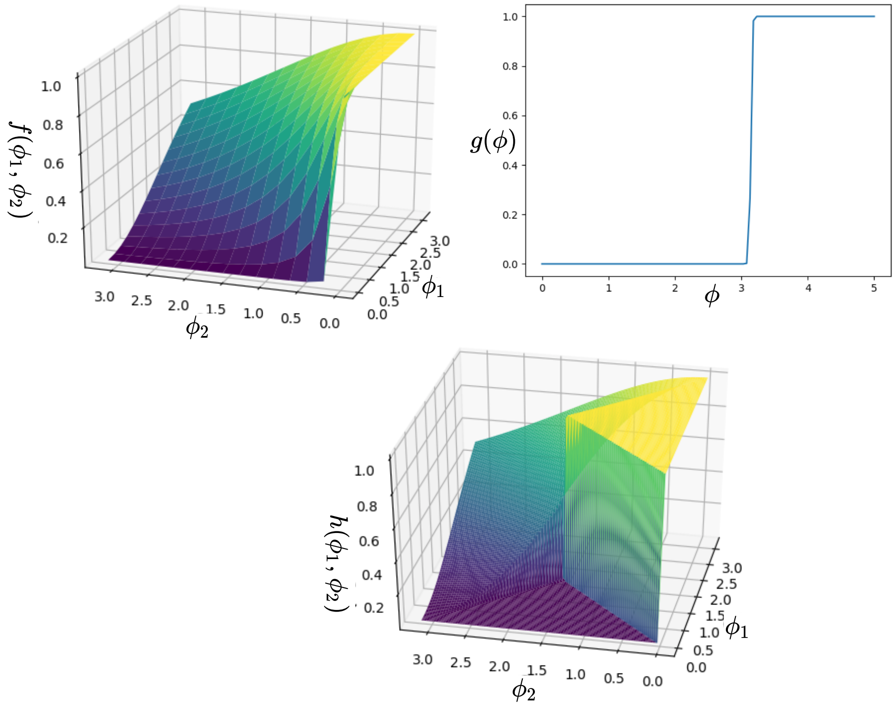


Figure 4.12: Plot of functions $f(\phi_1, \phi_2)$, $g(\phi)$ and $h(\phi_1, \phi_2)$.

As already mentioned our objective is to compute $X'(r', \phi_1')$ corresponding to $X(r, \phi_1)$. where, $r' = \|B'X'\|$ and $r = \|BX\|$. Now, that we have computed ϕ_1', we need to compute r' to get the corresponding source pixel X'.

Before we compute r' let us understand, a point closer to $B'C'$ (or $B'A'$) should be proportionally similar in distance to BC (respectively BA). To maintain this scaling effect we multiply r with the ratio of the source and the corresponding target line i.e., $B'C'$ and BC (resp. $B'A'$ and BA). For example, if $\|B'A'\| < \|BA\|$ then, $r' < r$, as $\frac{\|B'A'\|}{\|BA\|} < 1$. A demonstration of this case is given in Fig. 4.13. In this example, the line BA (red-colored) is longer than $B'A'$, hence, the upper part of the sleeve (rectangle) in the transformed image is scaled accordingly. (Naming convention of the labels is the same as that shown in Fig. 4.11).

Based on this concept, we obtain the value of r' using the following formula.

$$r' = r\left\{(1 - h(\phi_1, \phi_2))\frac{\|B'A'\|}{\|BA\|} + h(\phi_1, \phi_2)\frac{\|B'C'\|}{\|BC\|}\right\}. \tag{4.9}$$

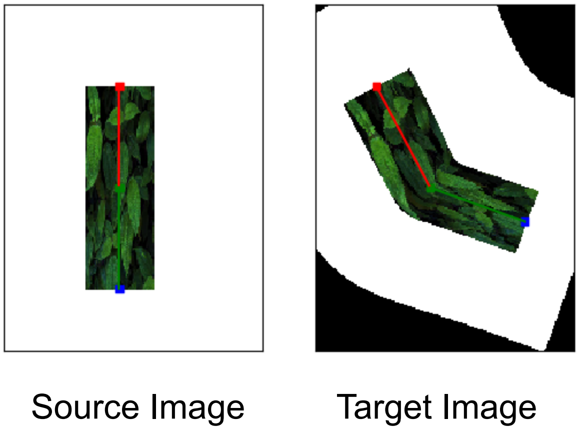


Figure 4.13: Our result depicting the effect of scaling of lines in our method. As the length of the line increases from source to target the corresponding part of the rectangle looks zoomed-in in the result.

Now, we have computed the radial and angular coordinates r' and ϕ_1' of X' respectively and, we already have the coordinates of A' and B'. Therefore, computing the location of X' is straightforward. So, we do not describe it in detail.

Beier and Neely method computes the source pixel for every pixel in the target image. But in this application, since we are interested in the target clothing region only, so, we compute the source pixels corresponding to the pixels belonging only to the target clothing region predicted by the proposed MPN. This improves the execution time. Additionally, the computation time of the Beier and Neely algorithm increases with the number of line segments, but as we are dealing with two lines only, this is not an issue here. The Beier and Neely transform suffers from the ghosting effect (discussed in Sec. 4.2), but our formulation solves this problem because of our human anatomy-inspired constraints. This can be observed from the results shown in the top two rows of Fig. 4.14. The predicted warps are consistent in our results unlike that of Beier and Neely's where regions of inconsistent warps are highlighted.

Occlusion is a major concern in the case of M2P try-on methods. Consider a case where the model's arm is more bent compared to the person's arm. For instance see the source and target line pairs are shown in the last two rows of Fig. 4.14. As already discussed, the bending of the arm increases flexion angle and causes folds in the sleeve near the elbow. Whereas when the flexion angle is less in the target compared to that of the source, some folded region of the model's sleeve might become visible in the target warp. Now, folds imply occlusion which results in loss of details. As a result, some parts in the adjoining region of the target line pairs can not be deterministically predicted, because pixel values in such parts are occluded in the source. For instance observe the regions highlighted in the results shown in the 5^{th} and 7^{th} columns of the last two rows of Fig. 4.14. Comparing with the ground-truths (1^{st} column of the top two rows) you may understand the predicted pixel values are inappropriate. We call such regions 'ambiguous regions' owing to the ambiguity in deciding the pixel values in such regions. Our formulation aids the identification of such regions. We denote a pixel X in the target as ambiguous if either of the following two conditions holds, $|\phi_1| > \frac{|\phi'|}{2}$ (in case the assumption 1 holds, i.e.,$|\phi_1| < |\phi_2|$), or, $|\phi_2| > \frac{|\phi'|}{2}$ (in case the assumption 2 holds i.e.,$|\phi_2| < |\phi_1|$). We choose not to interpolate the pixel values in these regions, rather leave them to be filled by our proposed ISN through inpainting. Notice our results shown in the 2^{nd} last column of the last two rows of Fig. 4.14 where we identified such regions and excluded the corresponding pixel values from being predicted. A more detailed discussion on deriving these conditions are elaborated in the Appendix. B. Note that deformation of the grid corresponding to each of the computed warp is shown next to it in Fig. 4.14 for better understanding.

4.3.4 Try-on image synthesis

The objective of this stage is to combine c' with P to generate a plausible and photo-realistic try-on result. Our warping method computes c' from only the visible clothing areas of the model's clothing. But due to the difference in pose between model and person some occluded areas of the model's clothing might become visible in the target. Also, long hairs often occlude clothing areas. Considering the target clothing mask predicted by MPN as $\mathscr{S}$, and the mask of the predicted warp c' as c'_m, we obtain such areas as, $\mathscr{S} \cdot \bar{c'_m}$, where $\cdot$ denotes dot product and $\bar{(.)}$ denotes complement. The main goals of ISN are to inpaint this region and produce a seamless combination of c' and P, while also removing the previous clothing details from P.

To achieve the above objectives we trained an encoder-decoder-based convolutional neural network (CNN). We call it the image synthesizer network (ISN). The inputs to this network are: (i) the predicted target clothing mask from the MPN, (ii) the body shape of the target person represented by its densepose representation, (iii) instead of providing the predicted target warp and the person image

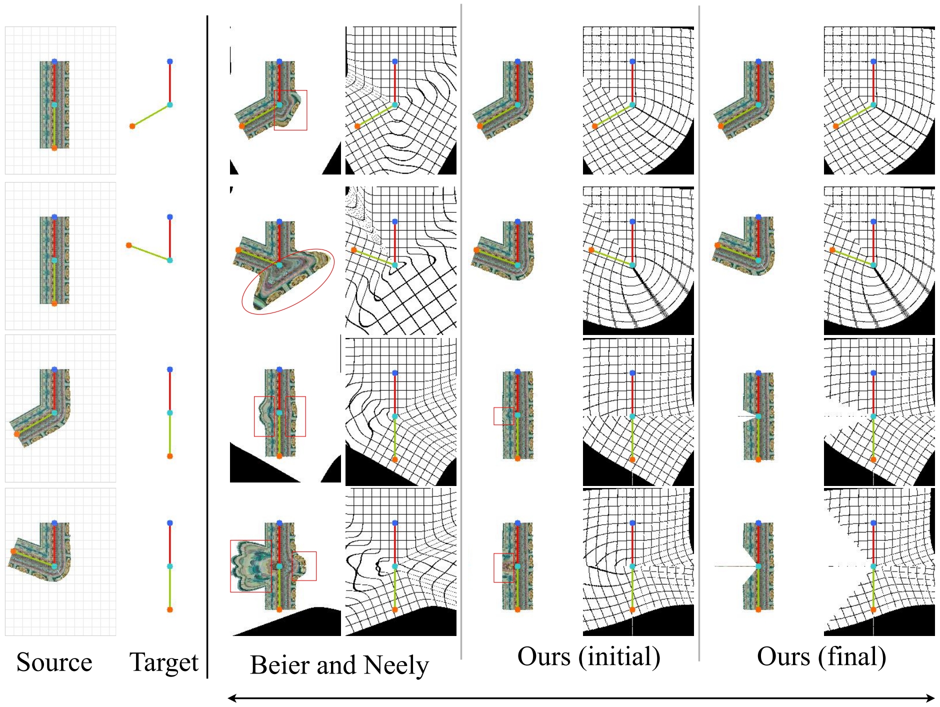


Figure 4.14: Comparison of the results of our approach with that of the Beier and Neely algorithm on different arrangements of two line segments, shown with red and green lines. Note that we have shown the source image with the line segments and their end points overlaid on it. Observe the highlighted regions showing the incorrect pixel values.

separately, we provide a combined representation of them. To compute this, the upper body details i.e., previous clothing (e.g., t-shirt, etc.), exposed skin areas are masked in the person image, and the predicted warp is overlaid on it. (iv) The inpainting mask $\bar{\mathscr{S}} \cdot c_m^{\bar{I}}$ (0 denotes the pixel with missing details).

We trained this network using self-supervision employing inpainting related loss functions as suggested in (Liu et al., 2018). Training in our case is tricky as we do not have paired data. To meet the restriction in the length of the paper we discuss the details of the training of ISN in the supplementary material.

4.4 Experiments

Dataset. We evaluate our method on MPV-I, MPV-II, MPV-III, MPV-front test sets from the MPV (Dong et al., 2019a) dataset and DeepFashion-I and DeepFashion-II test sets from DeepFashion (Liu et al., 2016a). We compare this method with VITON, CP-VTON, MGVTON, ACGPN, and the methods in our previous two chapters.

4.4.1 Quantitative analysis

We quantitatively evaluate our results on two metrics - FID and SSIM. We report the SSIM and FID scores in Table. 4.1 (SSIM score is reported only for MPV-III and MPV-front as ground-truths are not available for others). Table. 4.1 shows that our method outperforms the other methods in terms of both the metrics. Other than comparing only the final performance, we also compare the individual performance of our proposed MPN with the corresponding semantic generation module (SGM) of ACGPN on the MPV-front and MPV-III test sets. The respective SSIM scores reported in Table. 4.2 suggests that we compete well against ACGPN.

4.4.2 Qualitative Analysis

Before comparing the final try-on outputs we present a visual comparison among the results of our MPN and the corresponding semantic generation module (SGM) of ACGPN in Fig. 4.15. The results show some significant features (e.g., neck pattern, sleeve length) of the target clothing mask are not predicted accurately by SGM. Features of the predicted mask match more with the existing clothing of P, while it should be the same as the reference try-on clothing. For example, in the 2^{nd} row, the neck pattern should be 'round' in the output like the model's clothing, but, SGM wrongly predicts it to be 'V'-shaped, which is the shape of the existing clothing of P. Similarly, observe the sleeve lengths in rows 1, 3, and 5. Examples in the last 2 rows show the occluded and exposed areas are also predicted incorrectly by SGM. Compared to SGM, MPN keeps better features and thus predicts more accurate masks.

Some qualitative comparison of our final try-on output with that of some competitive methods is presented in Fig. 4.16. The results depict, the method proposed in Chapter 2 struggles to cope with the folding of the sleeve as this method is strictly restricted to simple human poses only. In case of CP-VTON, the target warp shows inappropriate deformation in both sleeve and torso regions. In the torso, this is occurring due to the flexibility of TPS which is in a way confusing the geometric matching network of CP-VTON. In Sec. 3.3.4 of the Chapter. 3 we have shown the feature maps of CP-VTON's geometric matching network (GMN) and added the explanation in claim of the poor learning of features by CP-VTON. ACGPN addresses this by imposing a second-order difference

Table 4.1: Quantitative evaluation on different datasets.

Dataset	Methods	FID↓	SSIM↑
DeepFashion - I	Ours (ch2)	**33.67**	-
	Ours (ch3)	64.35	-
	Ours	43.46	-
DeepFashion - II	Ours (ch2)	**26.40**	0.86
	Ours (ch3)	63.16	0.75
	Ours	31.16	**0.91**
MPV - I	MGVTON	44.44	-
	CP-VTON	28.38	-
	Ours (ch2)	22.59	-
	Ours (ch3)	24.11	-
	ACGPN	19.17	-
	Ours	**16.30**	-
MPV - II	MGVTON	42.76	-
	CP-VTON	32.56	-
	Ours (ch2)	25.10	-
	Ours (ch3)	26.43	-
	ACGPN	24.04	-
	Ours	**22.99**	-
MPV - III	MGVTON	38.58	0.77
	CP-VTON	25.78	0.81
	Ours (ch2)	24.88	0.82
	Ours (ch3)	22.53	0.86
	ACGPN	21.46	0.87
	Ours	**15.77**	**0.89**
MPV - front	MGVTON	35.70	0.76
	CP-VTON	21.03	0.74
	Ours (ch2)	12.06	0.89
	Ours (ch3)	14.34	0.80
	ACGPN	14.35	0.88
	Ours	**9.45**	**0.93**

constraint, that controls the differences in slopes between neighboring intervals for each axis of the predicted mesh grid. The method proposed in Chapter 3 produces visually better results compared to that of Chapter 2 and CP-VTON but still, the sleeves are not warped correctly. This is because the warping methods of these approaches are based on TPS whose bending constraint restricts higher change in slopes in between consecutive axes of the predicted grid. CP-VTON and ACGPN have tried to compensate for the limitations of their warping method in their try-on stage. Notice, the blurry

Table 4.2: Quantitative analysis of the proposed mask prediction network (MPN) and the semantic generation module of ACGPN. Scores are rounded to the fourth decimal place.

Dataset	Method	SSIM
MPV-front	ACGPN	0.9475
	Ours	**0.9507**
MPV-III	ACGPN	0.9421
	Ours	**0.9557**

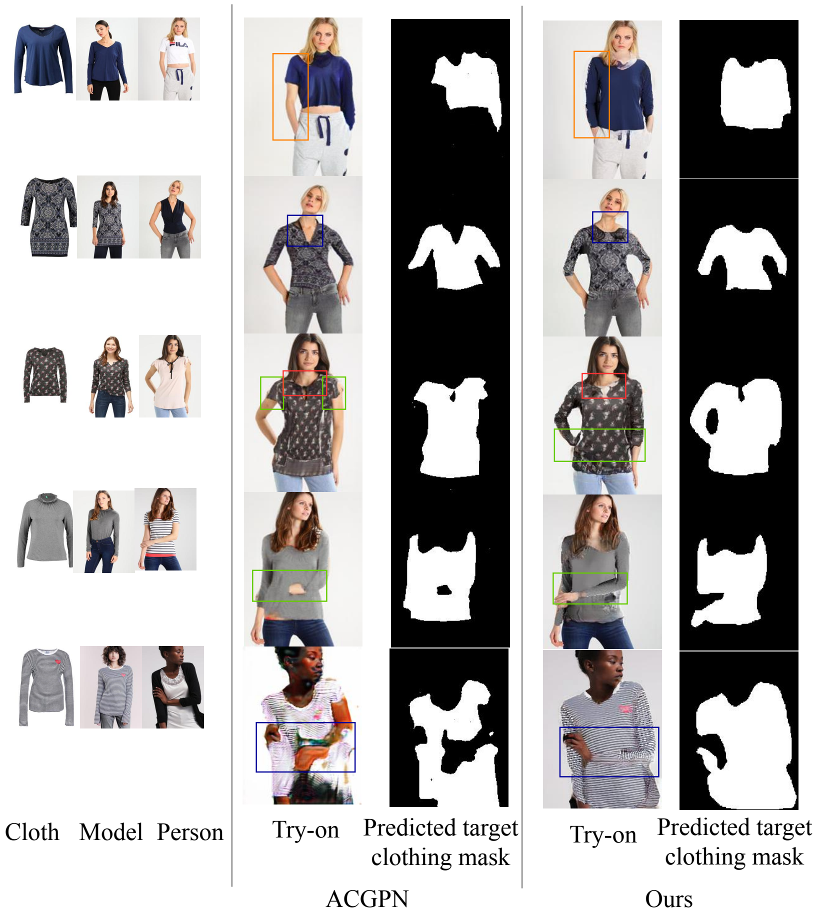

Cloth
Model
Person
Try-on
Predicted target clothing mask
Try-on
Predicted target clothing mask
ACGPN
Ours

Figure 4.15: Qualitative comparison of the performance of the proposed MPN with that of the semantic generation module (SGM) of ACGPN. We have highlighted the areas to be noticed in each of the results. Observe in most of the cases SGM does not preserve the features of the source clothing in the try-on result, instead inherits the features from the old cloth of the person (i.e., before try-on) which is not expected.

areas in their predicted warps which are mostly because those regions are being filled with relevant details by the try-on synthesis module of the respective methods which are not strong enough to fill the texture or color accurately. In comparison to others, our results show better-predicted warps. Similar to Chapter. 2 and Chapter. 3, employing the idea of landmarks correspondence in warping produces correct transformation in the torso region. However, unlike the other compared methods, we can handle the bending of the sleeves, and also our part-by-part warping approach is able to handle the cases of overlap between the sleeve and the torso. For instance, see the results in the top 3 rows of Fig. 4.16.

Some additional qualitative comparisons of our results with that of the compared methods are presented in Figs. 4.17, 4.18. We have also highlighted the significant details for better understanding. We provide some additional results of ours in Fig. 4.19 on various model-person pose and clothing features combinations. Overall, analyzing the results reveal that our approach is more apt in handling pose and feature variability in VTON.

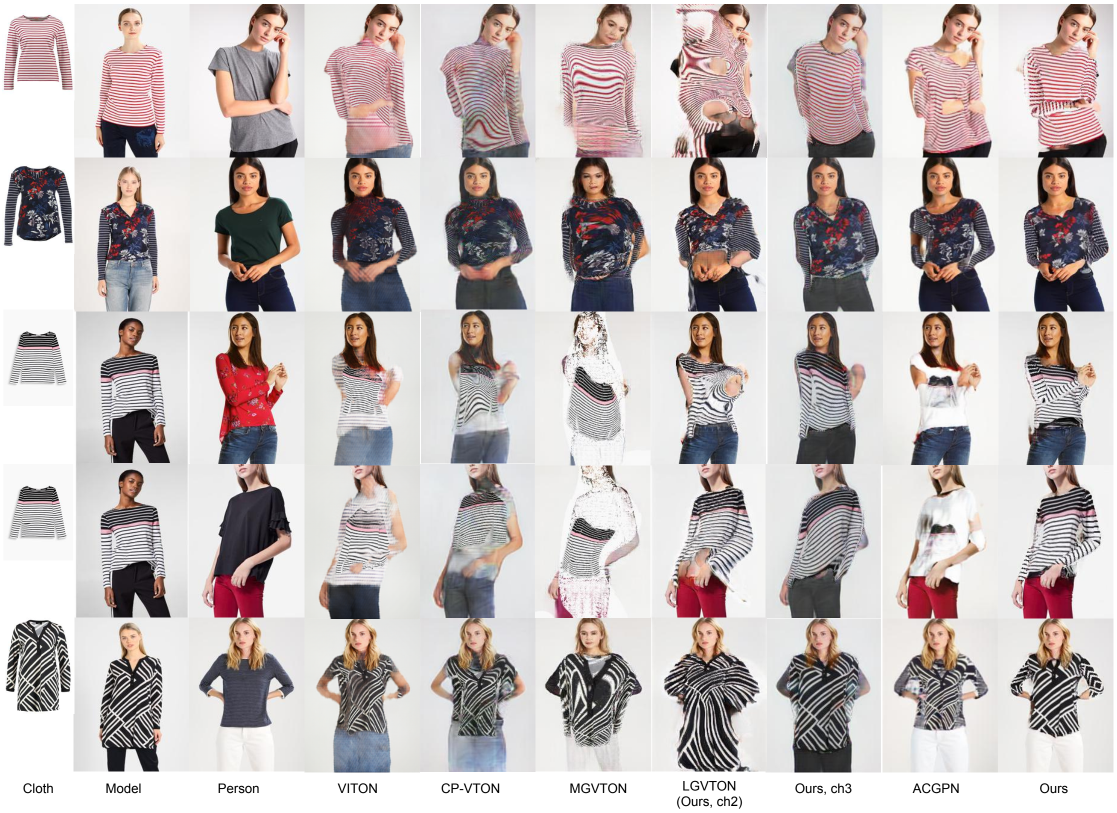


Figure 4.16: Comparative study on critical hand postures.

Ablation Study. We analyze the significance of employing MPN and its human parsing branch in our approach. For the former case, we compute the results without using the mask predicted by MPN.

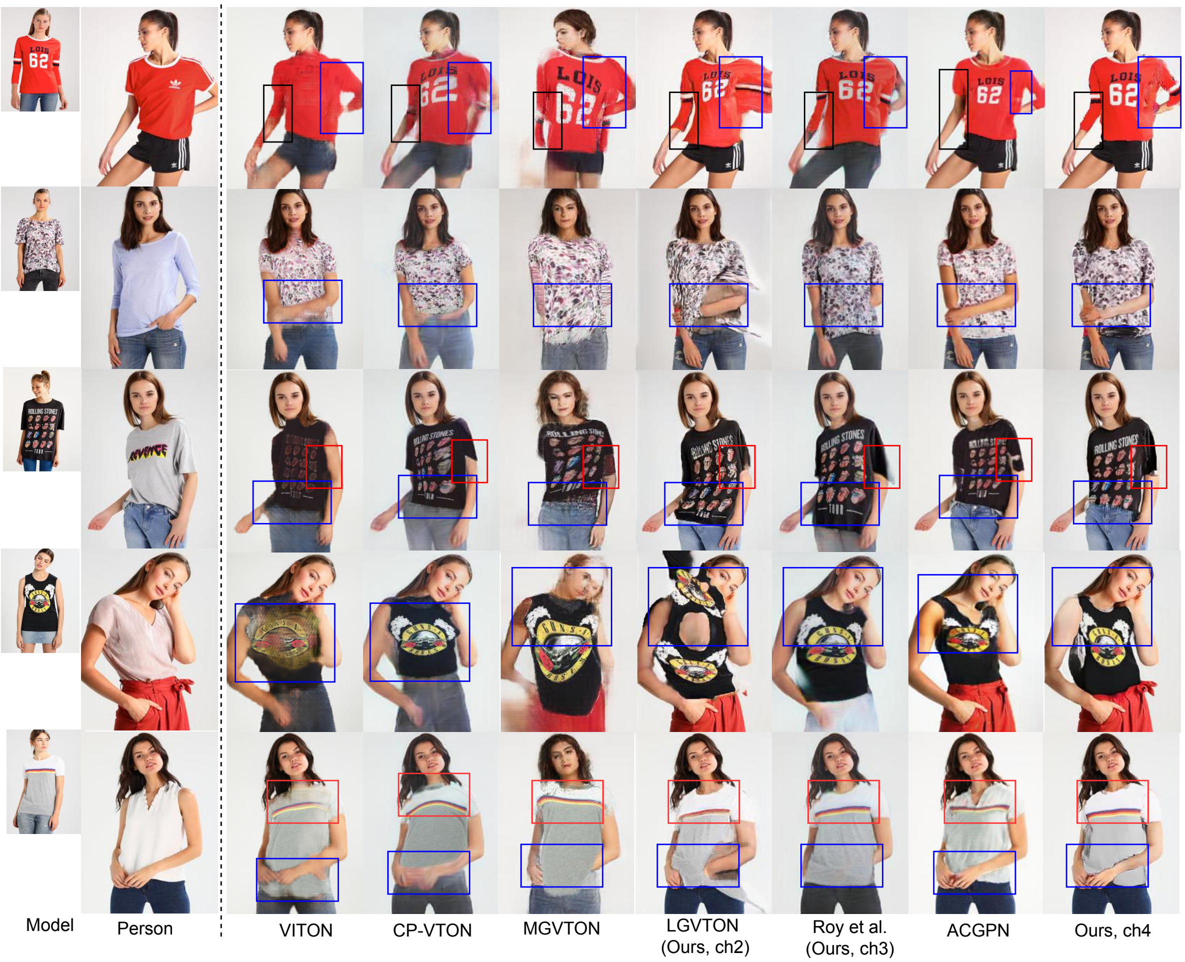


Figure 4.17: Comparative study on MPV dataset. Significant details are zoomed-in and shown right after each output.

For that, we replace the target clothing mask predicted by MPN with the mask of the predicted warp and compute the final result. For the latter case, we trained an instance of MPN without the parsing branch and compute the final results using that. SSIM and FID scores on the results of the two cases are given in Table. 4.3. Compared to the w/o MPN instance, our method secures a better score, which can also be verified from the visual comparison given in Fig. 4.20, e.g., see the previously occluded parts are being inpainted in target in our results. Compared to the without parsing instance, our results secure a better FID score as shown in Table. 4.3. A visual comparison of the results from the MPV-I set, portrayed in Fig. 4.20, shows improvement in some of the predicted feature details e.g., near the collar and in the adjacent region of upper and lower body clothes we see improved details.

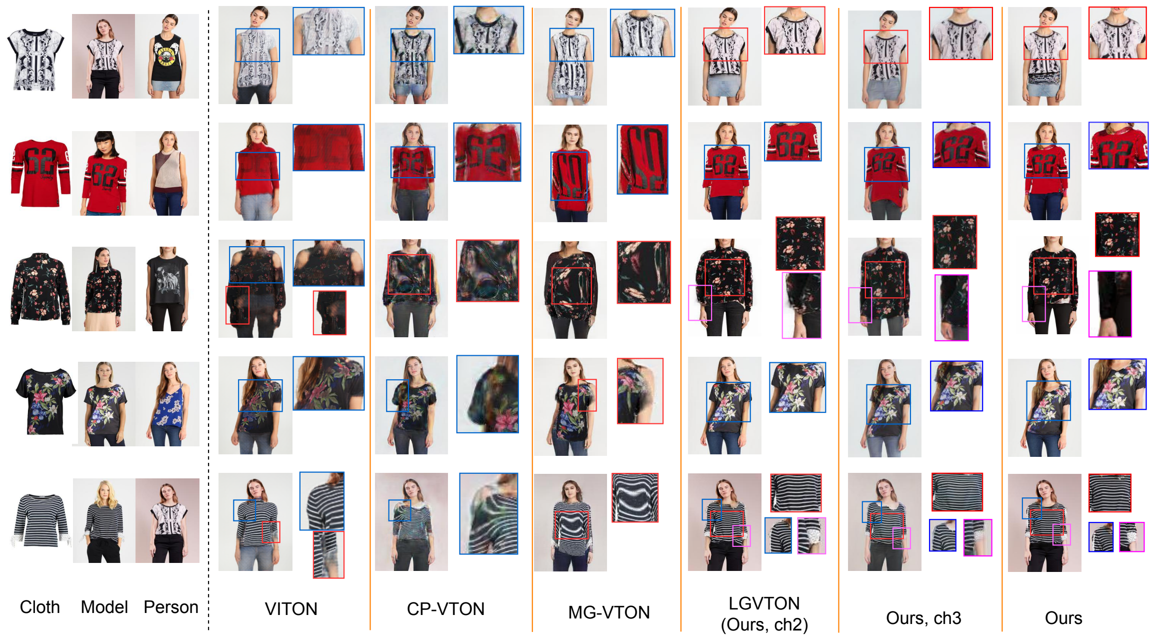


Figure 4.18: Comparative study on MPV dataset. Significant details are zoomed-in and shown right after each output.

Table 4.3: On the significance of MPN and MPN without parsing branch in the synthesis stage of our method.

	MPV-III		MPV-I	
Method	SSIM↑	FID↓	SSIM↑	FID↓
Ours (w/o MPN)	0.89	16.71	-	16.69
Ours (w/o parsing in MPN)	0.89	16.65	-	16.66
Ours	0.89	**15.77**	-	**16.30**

4.4.3 Number of parameters

A cost comparison in terms of the number of parameters is presented in Table.4.4. Observe that the no. of parameters in our method (column ‘P_1’) is 37.33% of that of CP-VTON and 10.76% of that of ACGPN and 6.72% of that of MGVTON. However, our cost of computation of inputs (column ‘Total P_2’) is higher especially due to employing the part-parsing inputs. In terms of the total number of parameters we are comparable to ACGPN and better compared to MGVTON. In terms of number of parameters though this method costs higher compared to our previous two approaches but we get significant improvement in performance in this approach. Note that, we do not report the execution

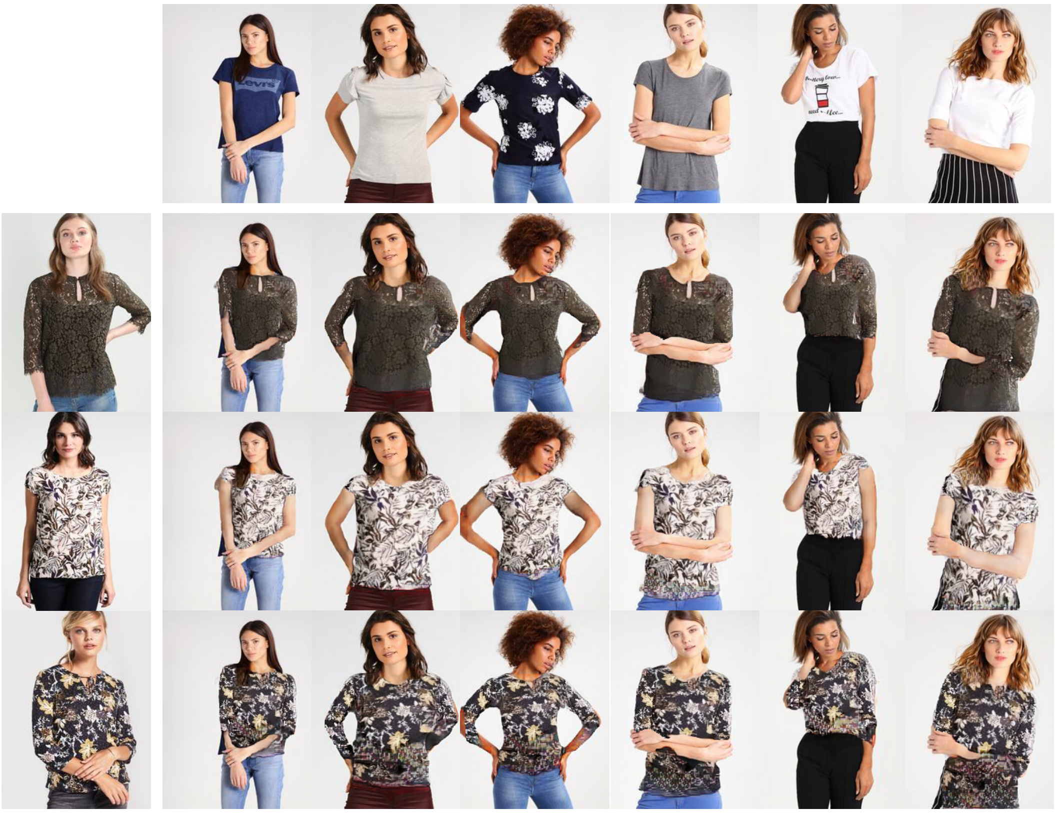


Figure 4.19: Our results on different pose and clothing feature combinations of model and person. The 1st row and column shows the model and person images respectively.

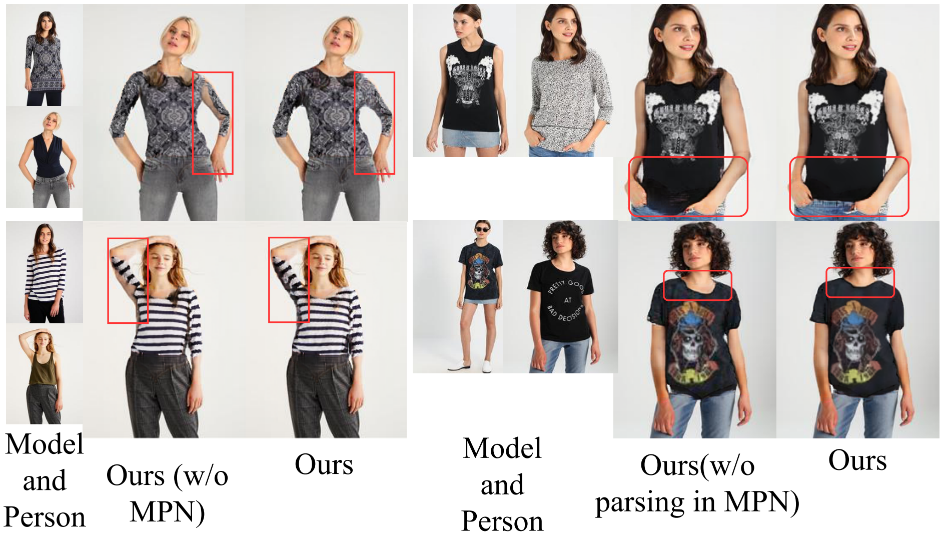


Figure 4.20: illustration of the significance of MPN and MPN without parsing branch in the synthesis stage.

times as it is highly sensitive to the underlying deep learning framework and different methods we have compared with are implemented using different deep learning frameworks. Moreover, the time of computation correlates well with the number of parameters.

Table 4.4: Comparing the number of parameters (in millions) of different methods. The total number of parameters of the proposed method is given in column P_1. The parameters required to compute different inputs are specified in the columns under P_2. The sum of the values reported in P_1 and P_2 are reported in the column Total. We used the following abbreviations for the different input estimation methods, HP - Human parsing, PE - Pose estimation, DPE - Densepose estimation, PPE - part-parsing estimation. The best and the second-best results are highlighted with bold notation and blue color respectively.

Methods	P_1	$\frac{\text{Ours} \times 100}{\text{Methods}}$	P_2					Total	$\frac{\text{Ours} \times 100}{\text{Methods}}$
			HP	PE	DE	PPE	Total P_2		
VITON	29.34	51.40%	75.65	52.31	-	-	127.96	157.30	186.04%
CP-VTON	40.40	37.33%	75.65	52.31	-	-	127.96	168.36	173.82%
MGVTON	224.46	**6.72%**	75.65	52.31	-	-	127.96	352.42	**83.04%**
LGVTON (Ours,ch2)	2.45	615.5%	75.65	52.31	59.73	-	187.69	190.14	153.91%
Ours, ch3	31.16	48.40%	75.65	52.31	59.73	-	187.19	218.35	134.02%
ACGPN	140.07	10.76%	75.65	52.31	-	-	127.96	268.03	109.18%
Ours	15.08	100%	75.65	52.31	59.73	89.87	277.56	292.64	100%

4.5 Discussion

This work proposes a novel model-to-person virtual try-on approach that attempts to propose a solution that is robust in terms of pose variation. More specifically, we attempt to solve three existing issues of VTON: overlap, folding, and occlusion. Previous methods have shown limited performance when the target warp requires significant bending or overlap. Bending in the target warp occurs due to the flexibility in human arm movements. Most of the existing warping-based methods use thin plate spline (TPS) transformation. However, the formulation of TPS has some limitations as follows. (1) TPS cannot model the cases of overlap in the clothing, since it is a 2D transform, (2) the bending constraint of TPS limits its applicability in case the target warp require significant bending, e.g., in case of arm bending of the model/target person that results in bending/straightening of the sleeves. Instead of employing one transformation function to warp the whole clothing, in this chapter, we follow a part-based approach, where the source person is segmented into semantically meaningful parts and each part is warped independently. So, to accommodate the cases involving significant bend, we propose a feature-based warping method being inspired from a traditional warping approach called field transform. This proposed method overcomes several issues of the previous clothes warping methods. Our warping method is guided by pose key points and follows constraints of human body movements. This results in more accurate warps. We also propose two learning-based modules that aid in improving the fit of the warp and synthesize a seamless output while taking care of the occluded part of the source clothing.

This method has some limitations too. Our method is based on landmark correspondences and, therefore, it is unable to work when the landmarks are not reliably detected. In addition, employing a part-based approach also imposes the additional cost of computing the part-parsing of the model and person. Also, the results of this method can be improved with better inpainting techniques.

Chapter 5

Conclusion

5.1 Overview

In this era of e-commerce, buying clothes online is becoming more and more popular. However, a major bottleneck of this online service is the absence of an appropriate trial room so that the buyer can try the dress before placing the order. The concept of ‘virtual try on’ has cropped up to fulfill this requirement. In this thesis, the problem of image-based virtual try-on (VTON) has been explored with an emphasis on the human structural guidance-based approaches to the warping of clothes. The kind of VTON problem that has been often explored in literature extensively is the cloth-to-person (C2P) scenario, which assumes a separate outfit image is available. However, we observe that the images of outfits are generally available in the form of a person (called model) wearing them. Under this situation, this work proposes VTON approaches for the model-to-person (M2P) scenario, where the image of clothing is transferred from the image of the model to that of the person. This M2P problem, compared to C2P, is much more difficult for two reasons: (1) in C2P the clothing is assumed to be in a position similar to the anatomical position of human, i.e., in an upright position, directly facing the observer; (2) since the clothing is displayed in the upright front pose, there is no occluded parts in the frontal part of the clothing. Both of these simplistic situations are not, in general, true for M2P problems.

The VTON problem is usually attempted in multiple stages because of the problem’s complexity. The warping stage works to transform the source clothing in the shape and pose of the target person; the synthesis stage works on combining the aligned clothing with the person’s image to predict the final try-on output. The synthesis stage needs to handle the cases of occlusion by own body parts. The main contributions of this thesis are related to the entire process with emphasis on the first stage i.e., the warping stage. Our solutions, as already mentioned, are based on the observation that clothing can be transferred from the model to the person according to the way the body shape and pose change from the model to the person. Other than warping, we propose other learning-based modules for improving or refining the fit of the predicted warped clothing addressing the limitations of the warping stages and synthesis modules to predict the final output.

5.2 Summary

In Chapter. 2 we present a self-supervised structural key points-based approach to VTON. Our method contains three modules. Among these modules, the first module utilizes the correspondence between the estimated landmark sets of the model's image with that of the person in order to predict the target warp related to the source clothing. It is observed that the use of structural key points i.e., landmarks for computing the target warp enables the warping function to be defined in terms of the geometry of the human body and the clothing more or less satisfactorily, because the landmarks related to humans are the anatomical key points i.e., the location of the adjoining positions of the different bones. This can well approximate the pose but is not always sufficient for representing the body shape of a human. This is reflected in the predicted warp where, though the alignment of the warped clothing is good, improvement is needed near the edges. We call this problem a warping glitch, and to refine the fit of the predicted warp, we propose a Mask Generator Module (MGM). This helps our final module, i.e., the image synthesizer module in combining the predicted warp with the person's image to synthesize a photo-realistic final try-on output. Our ablation study establishes the necessity and efficacy of each of these modules. It is worth noteworthy that this work suggests that the landmarks can be used successfully in the domain of virtual try-on. The results of this method also show that compared to learning the warping function based on learned features from the images, the landmark-based approach predicts better output. This is because of the structural consistency which is implicitly included as a constraint by using the landmark-based correspondence method for defining the warping function. However, this method is limited to simple human poses only. In the case of poses with bent or folded arms, this method often fails to predict a feasible warp. Here the computation of the warp is dependent on the whole set of landmarks. But for short-sleeved clothes, consideration of those landmarks that do not fall in the region of the source clothing makes the pose complex, and in the case of folds and bends the result degrades.

To deal with the limitations associated with the above-mentioned method, in Chapter. 3 we explored a neural network-based feature learning approach. The proficiency of deep neural networks in learning features specific to the task has been extensively explored in literature. In this chapter, we trained a special type of convolutional neural network called geometric matching network (GMN) to learn the features from the model and the person images and thereby establish the matching between those features to predict the desired TPS warping function. A similar network has been trained by multiple benchmark methods. However, due to a lack of human body geometry-related constraints embedded in the warping transformation, the predicted warps often showed irregular textures. Now it is observed that the clothing can be transferred from the model to the person by using the body shape and pose changes from the model to the person. In order to instill this idea into our feature learning and matching process, we provide the dense human pose representations of the model and of the person as input to the network instead of providing the model and the person images directly. Rigorous experiments establish the significance of our method compared to others. However, this method has a few limitations too. While the method can handle bending and folding of arms or pose variability between the model and the person better compared to other methods, in many cases, it fails to achieve photo-realistic results. This problem becomes severe in the case of long-sleeved outfits. Because due to crossed arm postures, the overlap between different clothing parts might happen. The existing approaches including the proposed approaches till this part of the thesis, especially the warping-based methods employing TPS transform, cannot tackle such cases. The reason is, image-based warping methods mostly reorganize the in-plane points. On the other hand, the overlap between different parts indicates some 3D displacement of the grid points. Hence, TPS and other existing ones cannot model such cases.

In Chapter 2 and Chapter 3, we have dealt with both explicitly specified features e.g., landmarks and learned features. Based on the computed feature matching, the TPS transformation function for computing the target warp is predicted. However, TPS transformation, as pointed out earlier, cannot handle the cases of overlap, because overlap among clothing parts (e.g., sleeves overlaps on the torso) occur due to the movement of arms in 3D. The arrangement of landmarks resulting from the 3D arm movement can not be modeled by TPS. To address these issues, in Chapter. 4 we attempt an approach where the clothing from the source person (i.e., model) is segmented into semantically meaningful parts and each part is warped independently to the shape of the person. It is observed that this idea of part-based warping solved the overlap issue. We applied TPS Transformation to the torso part of the clothing as it undergoes restricted deformation, and to handle bending and folding of the sleeves we proposed a hand-crafted feature-based warping method where we employed constraints related to the current problem context. Our idea is inspired by the traditional field warping method proposed by Beier et al. (Beier and Neely, 1992). In addition, we propose two learning-based modules: a synthesizer network and a mask prediction network for refining the fit of the predicted warp and predicting the final try-on output. Rigorous experiments executed on benchmark datasets show that this method achieves state-of-the-art performance producing photo-realistic, robust VTON solutions without requiring any paired training data.

5.3 Future Scope of Work

The main contribution of this thesis lies in establishing the significance of human body shape guidance in the context of warping the source clothing. However, the landmark-based warping approaches proposed in Chapters. 2 and 4 suffer from a basic limitation. These methods fail when the landmarks can not be reliably detected. For the side view images where a part of the human is occluded from view, the human landmark estimation method Cao et al. (2017) fails to predict the location of the landmarks. While with the correct prediction for the occluded areas the method in Chapter 2 can not predict a satisfactory solution, however, our method in Chapter 4 can handle this. But we can not expect the estimated landmarks in the occluded region will be correct always. Therefore, estimating correct landmarks and the ability to handle occluded regions can be a feasible area of research in the future.

In Chapter. 4 we overcome the limitations of TPS transform in addressing the cases of overlaps due to sleeves of the clothing, which may occur if a person is posing with his hand folded or crossed. We tackle this by employing a part-based warping approach. However, this imposes the additional cost of computing the human part segmentation of the model and the person. In the future, efforts can be made to explore other possible solutions than employing the part segmentation-based solution. Moreover, another possible research direction could be to explore whether the part segmentation of the model's clothing can be predicted from the human part-based labels available from the densepose Alp Güler et al. (2018) representation. Though one challenge in that direction is that densepose represents part-based labels of pixels corresponding to the human body only. However, in general, clothing may be loosely fitted. In that case the contour of the clothing may not be withing the human body contour. Hence for the pixels which are outside the human body area, part-labels need to be predicted for getting part-segmentation of the clothing. Although that is another challenging problem, but a problem worth exploring.

The solutions proposed in this thesis are based on extracting the clothing from the model's image followed by transforming it according to the shape and pose of the candidate person. However, the source clothing might have wrinkles and folds that generally happen during wearing the outfit. The removal of these folds and wrinkles before transforming the clothing may produce a more practical and visually appealing solution. Since VTON is all about enhancing the user's experience, therefore, an attempt can be made to look into this direction in the future.

Apart from these future scopes of work, our method in Chapter. 4 can open a new research problem. Here we used the part segmentation of the model and the person. Combining the different segmented parts of the model's and the person's clothing can be done to synthesize a new clothing image for try-on. This can open the windows to e-commerce based tailor-made product purchasing. For example, a user may want to order an item of clothing with the torso of the clothing from one of the model's clothing and the sleeves from some other clothing. This can be a good solution for fashion designers as well, who generally run lots of trials before designing a new dress.

It has been observed that it is difficult to design one end-to-end network for the try-on task. This is mostly because VTON is a critical task and it puts a burden on one module to learn the whole idea of try-on. In addition, the problem of lack of paired training data makes it more difficult. While the method proposed in Lewis et al. (2021) has shown this could be done, this method's execution time is high and also not reliable as it only works for those images which have had a good representation in the feature space. Our method in Chapter. 4 achieves state-of-the-art performance. Concerning this, the predicted target to source pixel correspondences computed by our revised field transform method may be used as ground-truth for training a neural network for an end-to-end VTON system. This may save execution time because of the reduced number of parameters due to a reduction in the number of modules.

Another possible research direction could be revising the formulation of the TPS function itself to relax the bending constraint while maintaining structural constraints of the human body and the clothing. In terms of execution time, VTON is supposed to be a real time system, where the prospective buyer needs not spend much time to try on a particular clothing from a huge collection. So we have to employ inexpensive methods. Keeping this issue in mind we have proposed a warping method inspired from the Beier and Neely method for warping Beier and Neely (1992); which is a relatively old method of warping but still shows reasonably well performance in the current problem context. However, more modern techniques may be employed in future.

Appendix A

Supplementary Materials for Chapter 2

Here, we first discuss the formulation of Thin-plate spline (TPS) transformation. Thereafter, we provide the implementation details of our network architecture. In addition, we give some results of LGV-TON on DeepFashion Liu et al. (2016a) and MPV dataset Dong et al. (2019a) in Figs. A.2, A.3, A.4, and A.5.

A.1 Thin Plate Spline (TPS) definition

In general, given the n pairs of source and target control points, an interpolation function may be constructed by minimizing the data error given by sum of squared difference as follows.

$$\tau[\omega] = \sum_{j=1}^{N} \|\omega(x_j, y_j) - (u_j, v_j)\|_2^2, \tag{A.1}$$

where $\{(x_j, y_j)\}$, $\{(u_j, v_j)\}$, $j = 1, \cdots, N$ represents two different sets of data points. Now finding a solution to this interpolations is an ill-posed problem because we are trying to to interpolate continuous function from sparse supplied data. In other words, depending on the form of the transformation function $\omega(..)$ there may be various solutions.

TPS transform (Grimson, 1981; Duchon, 1977) impose some quality criterion, such as smoothness, that the interpolation function should satisfy. Following the thin plate spline transform $\tau[\omega]$ is considered as a energy functional defined as

$$\tau[\omega] = P(\omega) + S(\omega), \tag{A.2}$$

where

$$P(\omega) = \sum_{j=1}^{N} \|\omega(x_i, y_i) - (u_i, v_i)\|_2^2,$$

$$S(\omega) = \lambda \iint_{\mathbb{R}^2} [\omega_{xx}^2 + 2\omega_{xy}^2 + \omega_{yy}^2] dx\, dy.$$

Here $P(\omega)$, as before, is the penalty function due to deviation of data points. $S(\omega)$ is the stabilizing functional, called the second-order Sobolev semi-norm, whose minimization ensures smoothness of the function ω. λ is the regularization parameter ($\in [0,1]$), controlling the relative importance between a close fit (first term) and the smoothness (second term) of the interpolated function.

A closed-form solution of (1.2) as proposed by Wahba (1990) is given by

$$(u, v) = \omega(x, y) = \mathbf{a}_0 + \mathbf{a}_1 x + \mathbf{a}_2 y + \sum_{j=1}^{N} \mathbf{c}_j \nu(\|(x,y) - (x_j, y_j)\|_2), \tag{A.3}$$

where $\mathbf{a}_0$, $\mathbf{a}_1$, $\mathbf{a}_2$, $\{\mathbf{c}_j : j = 1, 2, \cdots, n\}$ are vector parameters with dimension equal to the dimension of the control points, which is 2 in our case. The radial basis kernel used in TPS is $\nu(p)$ = (p^2 ln p). Note that, we use the bold notation for representing vectors. Rewriting the Equation. A.3 with scalar parameters instead of vector we get the following two equations (since the dimension of control points in 2 in our case),

$$u = a_0^x + a_1^x x + a_2^x y + \sum_{j=1}^{N} c_j^x \nu(\|(x,y) - (x_j, y_j)\|_2), \tag{A.4}$$

$$v = a_0^y + a_1^y x + a_2^y y + \sum_{j=1}^{N} c_j^y \nu(\|(x,y) - (x_j, y_j)\|_2), \tag{A.5}$$

where $\mathbf{a}_0 = (a_0^x, a_0^y)$, $\mathbf{a}_1 = (a_1^x, a_1^y)$, $\mathbf{a}_2 = (a_2^x, a_2^y)$, and, $\mathbf{c}_j = (c_j^x, c_j^y)$.

Originally the idea of TPS was formulated to model the bending of a thin metal plate. In such formulation $S(\omega)$ is analogous to measuring the bending energy in the plate. For instance, as $S(.)$ approaches zero, the plate becomes increasingly planar, i.e., less and less bending. Bending causes displacement of the plate in the orthogonal direction to the lie of the plate. However, an orthogonal displacement might also cause displacement of the x and y coordinates of the plane of the plate. Hence instead of considering the orthogonal displacement, it can be considered as applied directly to the x and y coordinates of the plate. This way the formulation of TPS was formulated to be used to solve an interpolation problem Bookstein (1989).

A.2 Implementation Details

A.2.1 Pose Guided Warping Module (PGWM)

The dense neural network in fashion landmark predictor network $\mathscr{F}$ contains 6 consecutive dense layers with 900, 800, 600, 500, 250, 100 nodes respectively, each having activation function tanh. Finally, the output layer has 12 nodes (for 6 fashion landmarks each with 2 coordinate values x and y) and sigmoid as the activation function.

A.2.2 Mask generator module (MGM)

The architecture of MGM is that of an hourglass network Newell et al. (2016). Generally, the processing of clothes requires identifying the different parts of them to establishing a semantic understanding of their structure. A well-known network that suits this requirement is the hourglass network Newell et al. (2016).

An hourglass network is a CNN (Convolutional Neural Network) that captures features at various scales and is effective for analyzing spatial relationships among different parts of the input. Multiple of these hourglass networks can be stacked together with intermediate supervision for making it deeper. However, for our purpose, the stack size of 1 is found to be sufficient. An overview of the architecture of one hourglass network is given in Fig. A.1. The network is called hourglass due to its top-down, bottom-up architecture, which matches with the shape of an hourglass. It uses convolution

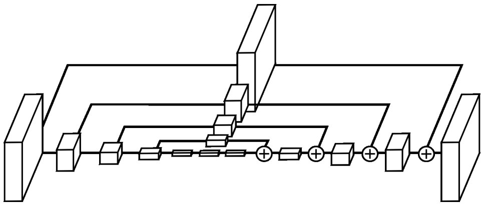


Figure A.1: Overview of an hourglass network (taken from the original paper Newell et al. (2016)), where each block represents a residual module.

and max pooling layers to downscale features to a low resolution. After that, the top-down sequence of upsampling begins, and at each resolution, the corresponding features from downsampling parts are added as skip connections. However, it applies more convolutions on the features of skip connections and then does element-wise addition of the two sets of features.

A.2.3 Image Synthesizer Module(ISM)

The hourglass network in G (the generator network of ISM) has a stack size of 1. The convolution layer generating I_m in G has a kernel of size 1×1 with L_1 regularization and sigmoid activation function. For training ISM, we alternate between 3 steps of generator training and 1 step of discriminator training. It is trained with the same settings of Adam as our PGWM. For DSSIM loss, the kernel size taken is 3×3. The discriminator D is a patchGAN discriminator Isola et al. (2017). Instead of classifying the whole image as real or fake, it classifies each patch of the image, where the patch size is much smaller than the input image size. Hence pixels separated by more than a patch diameter gets modeled independently. This makes it work like a texture/style loss as discussed in Isola et al. (2017), which helps to keep better texture in the final output image. Existing works have shown the efficacy of patchGAN Isola et al. (2017),Xian et al. (2018) in image-based problems. For human parsing, we used the human parsing network proposed by Gong et al. (2017) pretrained on the LIP dataset Gong et al. (2017). The dataset contains 19 part labels, 6 labels for body parts, and 13 for clothing categories. During our quantitative analysis we used the models of CP-VTON, VTNFP, MG-VTON pretrained on MPV Dong et al. (2019a) dataset. For VITON we used the VITON dataset pretrained model weights provided in its official implementations. Some more results of this work on DeepFashion and MPV datasets are shown in Figs. A.2, A.3, A.4, and A.5.

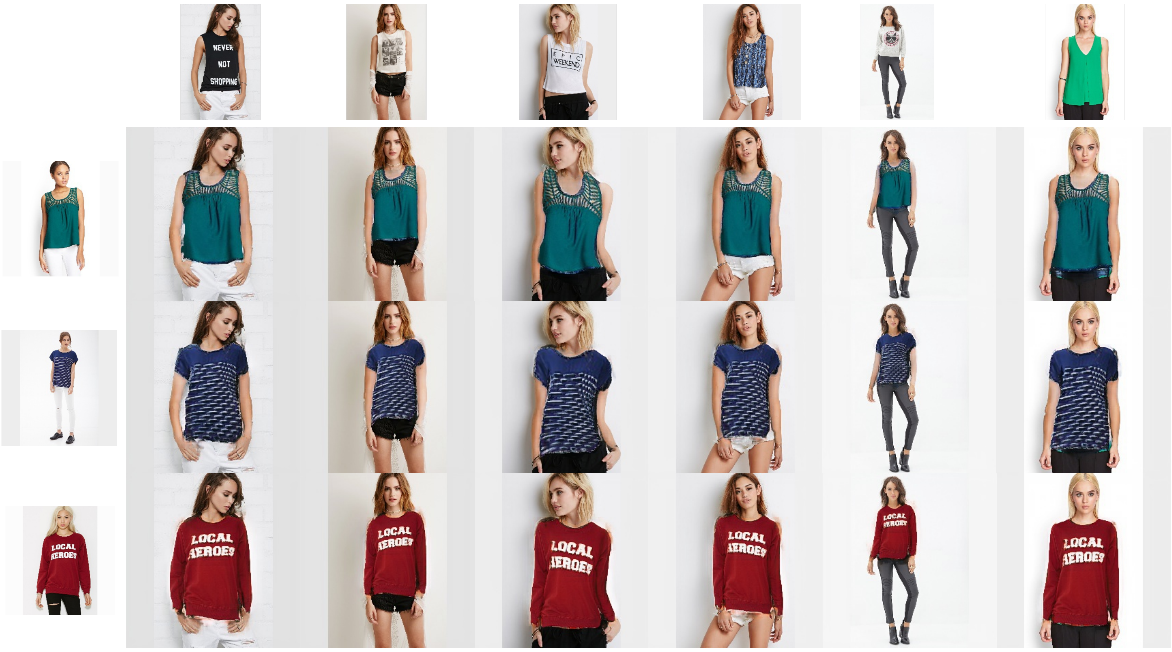


Figure A.2: Our results on DeepFashion Liu et al. (2016a) dataset. Image at position (i,j) represents the result of VTON when person at j^{th} column wears the clothes of the model at the i^{th} row.

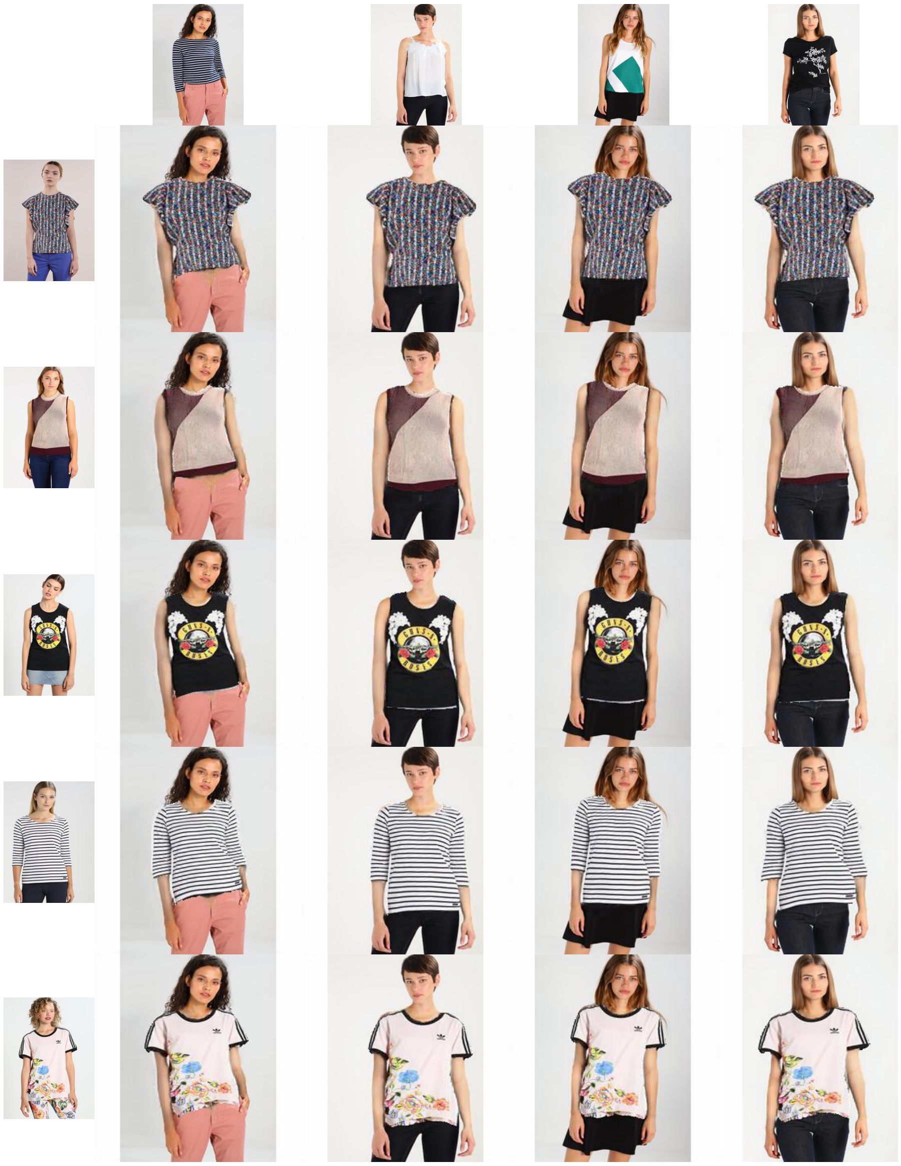


Figure A.3: Our results on MPV Dong et al. (2019a) dataset. Image at position (i,j) represents the result of VTON when person at j^{th} column wears the clothes of the model at the i^{th} row.

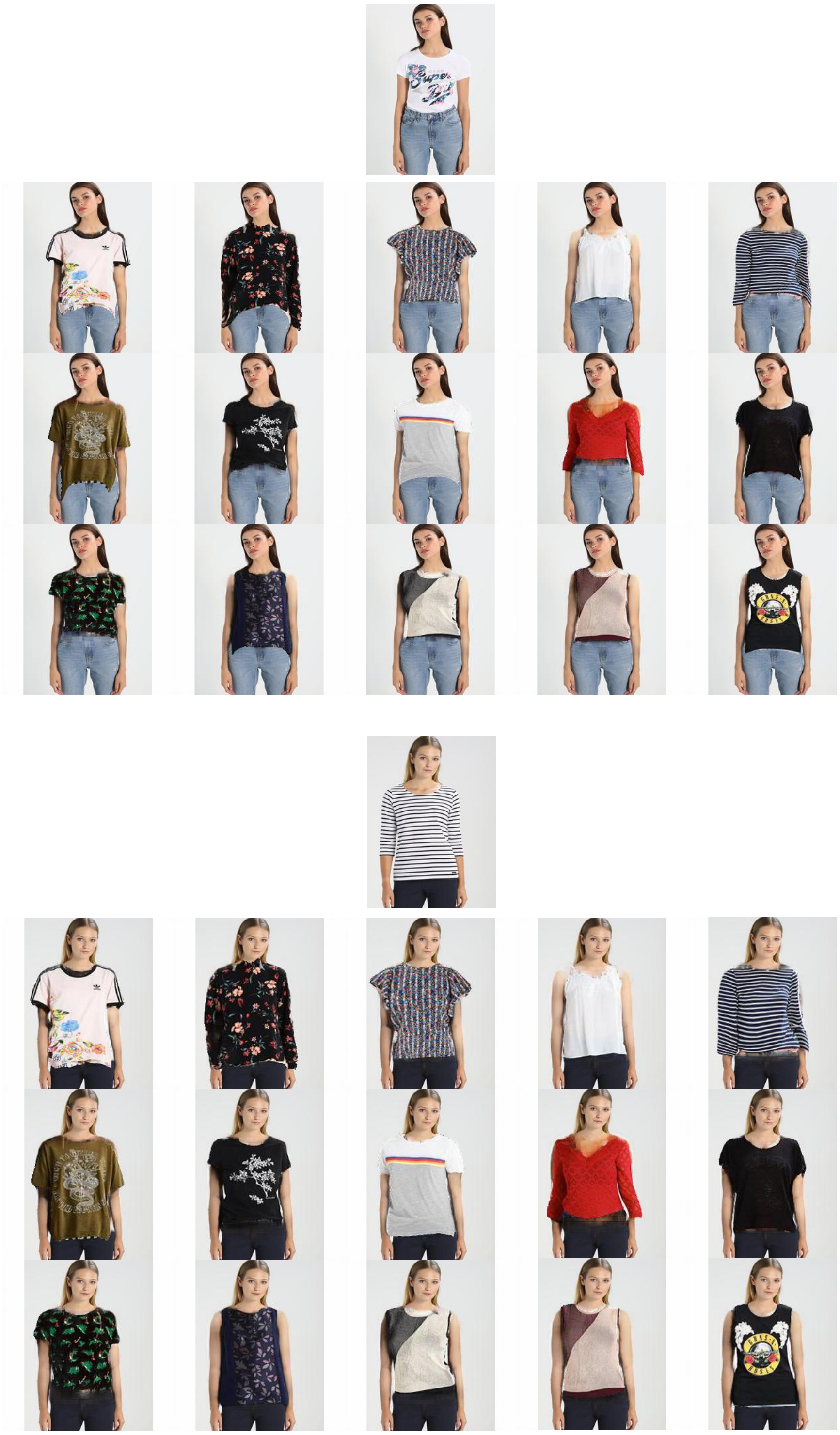


Figure A.4: Our results on MPV dataset for different model and person combinations.

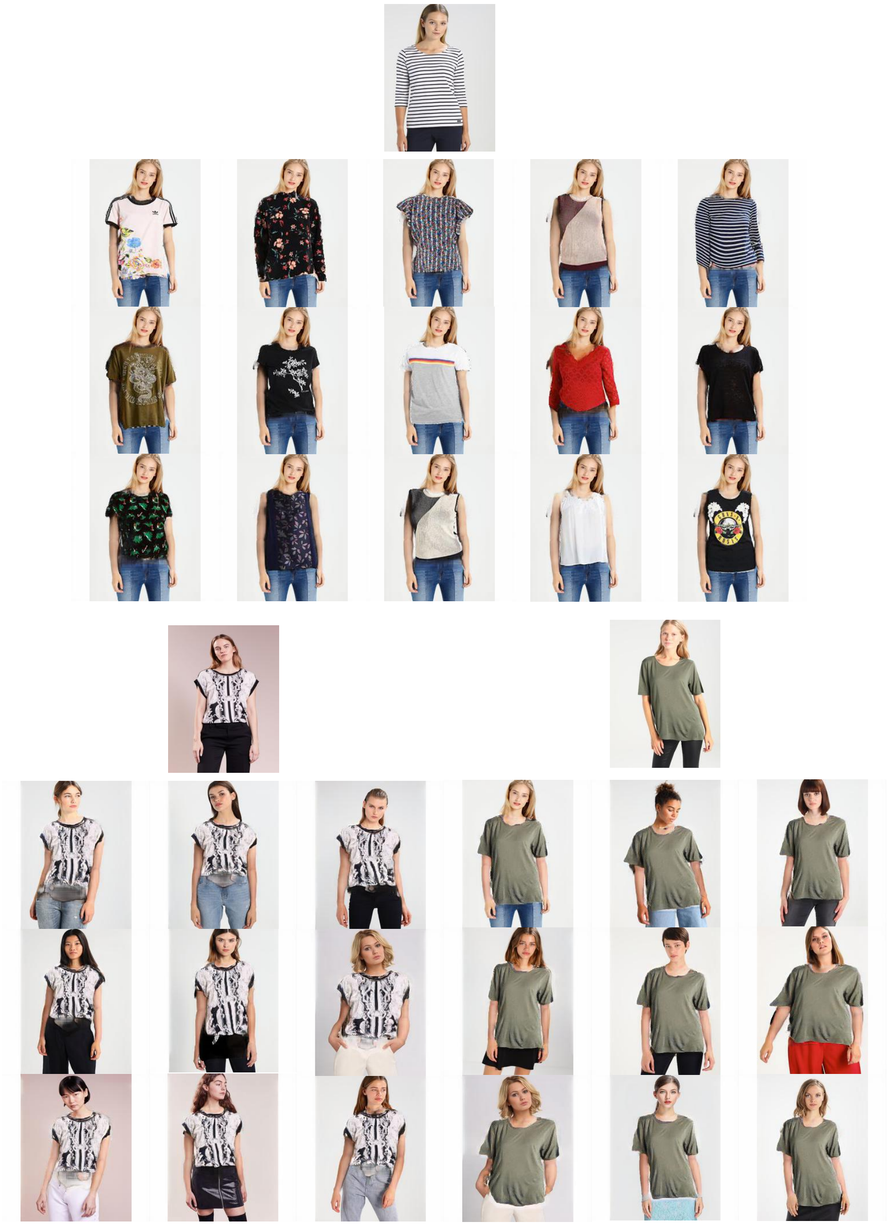


Figure A.5: Our results on MPV dataset for different model and person combinations.

Appendix B

Supplementary Materials for Chapter 4

B.1 On identifying the ambiguous regions of the target warp

This discussion is a continuation from the last paragraph of Sec. 4.3.3 of the main chapter. Previously we proposed the two conditions to identify ambiguous location X in the target grid,

- $|\phi_1| > \frac{|\phi'|}{2}$ when assumption 1 holds at X i.e., $|\phi_1| < |\phi_2|$ (case 1).
- $|\phi_2| > \frac{|\phi'|}{2}$ when assumption 2 holds at X i.e., $|\phi_2| < |\phi_1|$ (case 2).

Here we explain how our formulation aids in identifying the ambiguous regions of the target warp. The condition 1 says, a target pixel X is ambiguous if $|\phi_1| < |\phi_2|$ and $|\phi_1| > \frac{|\phi'|}{2}$. Now let us understand the reason.

It is mentioned that at X, $|\phi_1| < |\phi_2|$.

Therefore, based on our assumption 1, in such case,

$$\phi'_1 = \phi_1 \tag{B.1}$$

Note that the absolute operator is disregarded because as previously mentioned the angles are computed so that the signs are same.

We know, $\phi'_2 = \phi' - \phi'_1$. Therefore, using Eq. (B.1) we get,

$$\phi'_2 = \phi' - \phi_1. \tag{B.2}$$

Considering, $|\phi_1| > \frac{|\phi'|}{2}$ we get,

$$\begin{aligned}
& 2|\phi_1| > |\phi'| \\
\implies & |\phi_1| > |\phi'| - |\phi_1| \\
\implies & |\phi_1| > |\phi'_2| (\text{using Eq.(B.2)}) \\
\implies & |\phi'_1| > |\phi'_2| (\text{using (B.1)}).
\end{aligned} \tag{B.3}$$

Therefore at X, $|\phi_1| < |\phi_2|$ (given) and for the corresponding X', $|\phi'_1| > |\phi'_2|$ (using (B.3)). Hence, the relation between the angles is reversed in target and source respectively. This physically means, a pixel X closer to the upper arm in the target (without loss of generality) maps to a pixel X' that is closer to the lower arm in the source. However, based on our assumption the relationship should have been the same. Therefore, our assumption does not hold here. Similar justification also holds for condition 2. Therefore, we are able to detect the ambiguous regions by checking where our assumptions fail. We chose not to interpolate the values in such regions and let the pixel values be predicted through inpainting by the proposed ISN.

B.2 Training Details

Our approach has two trainable networks - mask prediction network (MPN) and image synthesizer network (ISN). Below we discuss the training data and loss functions related to each of this network.

MPN is trained on same person multi-view images, which means each training model and person image pair contains images of the same person wearing the same clothing from two different viewpoints. Some sample images of such type are shown in Fig. B.1. Sample training inputs to the main branch and corresponding ground truths of this network are shown in Fig. B.1. We have used the MSE loss function for training both the branches. However, an additional loss, called *count loss* as mentioned in Roy et al. (2021) is used in the main branch. The loss weights for count loss and MSE loss are 10^{-4} and 1.0 respectively. Count loss is simply the MSE loss between the count of pixel values labeled as one in the output and the corresponding ground truth.

We train ISN using self-supervision because paired training data is not available in our case. To replicate the probable missing regions occurring during testing we used a random mask dataset Liu et al. (2018), which is used to randomly remove parts of the clothing from the input training images. A

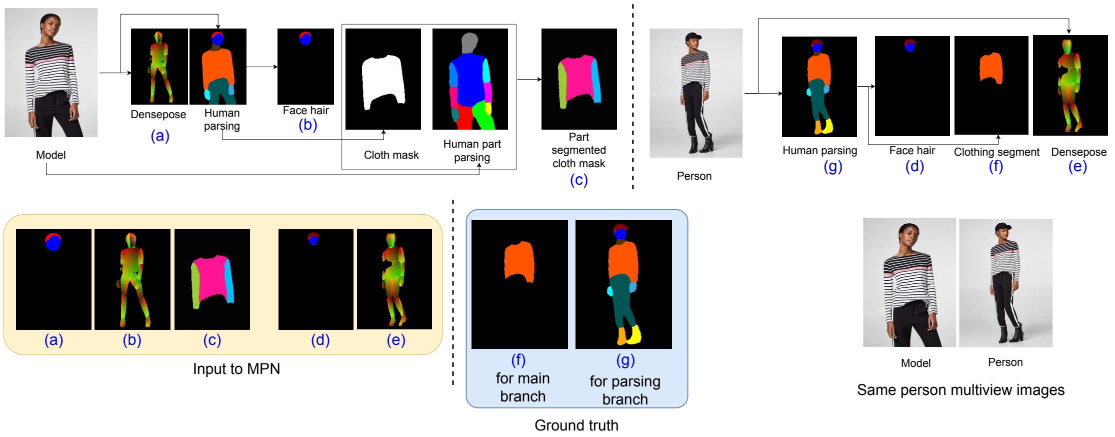


Figure B.1: Sample inputs and ground truths for training MPN.

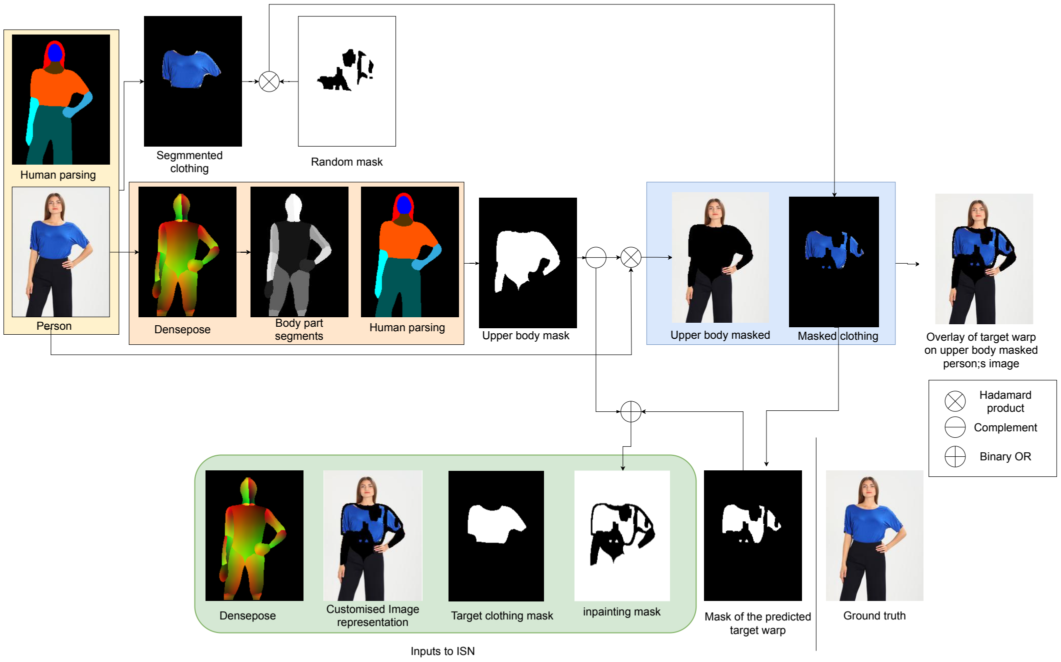


Figure B.2: Sample inputs and ground truths for the self-supervised training of ISN.

sample input data and the corresponding ground truth are shown in Fig. B.2. The reason to use random masks is to make the network learn to fill missing regions of clothing. Note that using the random mask in such cases is a very common practice in inpainting literature. Since the input set of images is prepared to replicate the test image scenarios, therefore, our training strategy is self-supervised. We train this network using inpainting loss proposed in Liu et al. (2018). This is a weighted combination of MSE loss, total variation loss, perceptual loss Johnson et al. (2016), and style loss Liu et al. (2018). The weights are kept similar to that reported in Liu et al. (2018). Additionally, with this loss, we have

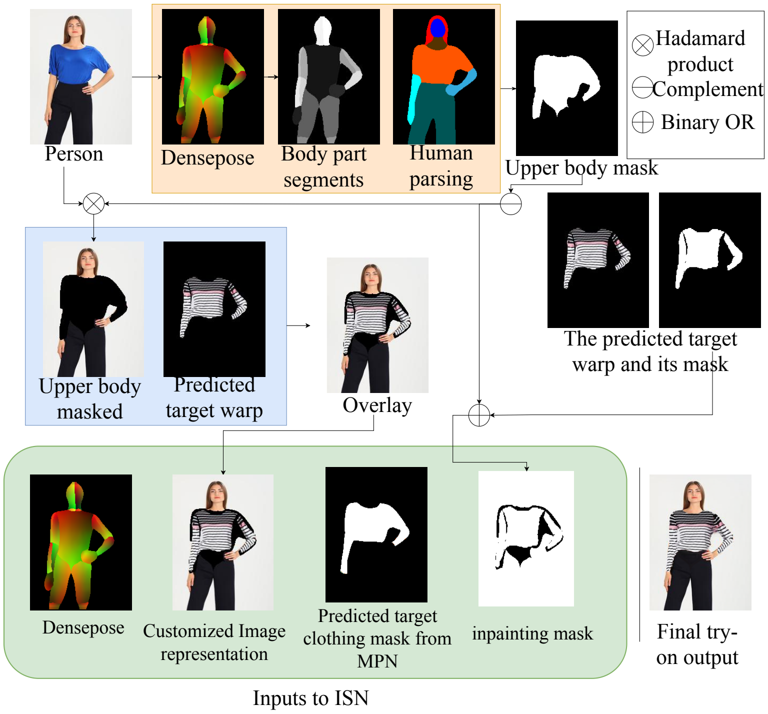


Figure B.3: Illustration of the inputs during test-phase to the ISN.

used DSSIM loss with loss weights 1.0. DSSIM is related to the image similarity metric (SSIM) Wang et al. (2004) by the formulae (1 - SSIM)/2.

B.3 Implementation Details

Based on our datasets the loss values of ISN and MPN networks converged around 12 and 10 epochs respectively. We trained ISN on the MPV-front dataset. We collected a set of same person multiview image pairs containing 58,968 samples for training the MPN network. The details of these datasets are given the main chapter. We used the Adam optimizer with standard settings for training both these networks. The architecture of MPN is already shown in the main chapter and that of ISN is the same as the hourglass network Newell et al. (2016) with the number of input layers changed according to our need.

B.4 Results

Some results of our method on a different model and person combinations have been given in Fig.B.4.

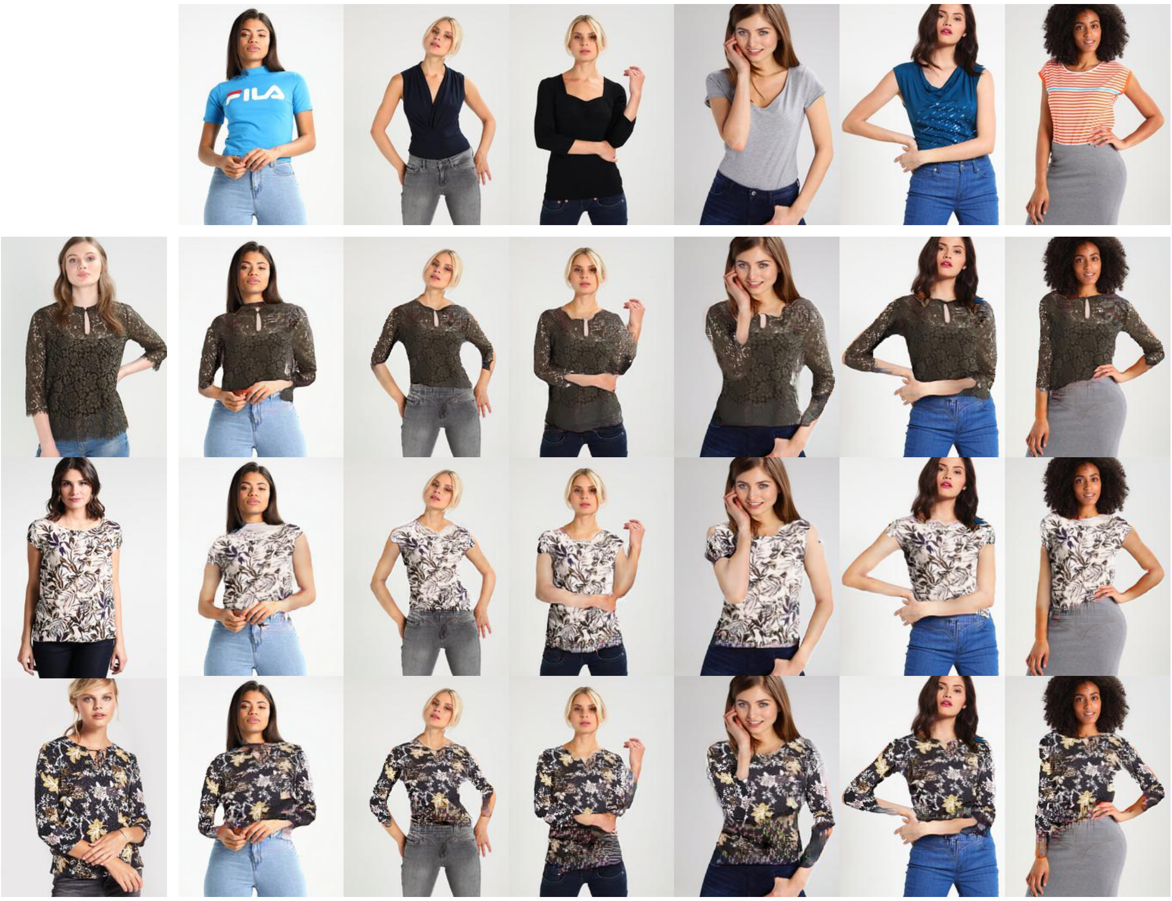


Figure B.4: VTON results of our method on different model and person combinations. The top most row shows the person images and the left most column shows the model images.

List of Publications by the Author Related to The Thesis

D. Roy, S. Santra, and B. Chanda. Incorporating human body shape guidance for cloth warping in model to person virtual try-on problems. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2020.

D. Roy, S. Santra, and B. Chanda. Lgvton: a landmark guided approach for model to person virtual try-on. *Multimedia Tools and Applications*, Jan 2022a. ISSN 1573-7721. doi: 10.1007/s11042-021-11647-9.

D. Roy, S. Santra, D. Mukherjee, and B. Chanda. Significance of skeleton-based features in virtual try-on. *arXiv preprint arXiv:2208.08076*, 2022b.

List of Other Publications by the Author

D. Roy, S. Santra, and B. Chanda. Incorporating human body shape guidance for cloth warping in model to person virtual try-on problems. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2020.

M. Wadhwani, D. Kundu, D. Chakraborty, and B. Chanda. Text extraction and restoration of old handwritten documents. In *Digital Techniques for Heritage Presentation and Preservation*, pages 109–132. Springer, 2021.

References

Z. Al-Halah, R. Stiefelhagen, and K. Grauman. Fashion forward: Forecasting visual style in fashion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 388–397, 2017.

R. Alp Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 3, 6, 11, 18, 25, 43, 65, 70, 92

D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. *ACM Trans. Graph*, 24:408–416, 2005.

Authors. The frobnicatable foo filter, 2006. ECCV06 submission ID 324. Supplied as additional material `eccv06.pdf`.

T. Beier and S. Neely. Feature-based image metamorphosis. *ACM SIGGRAPH computer graphics*, 26(2):35–42, 1992. 3, 14, 66, 67, 68, 92, 94

S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. In *Advances in neural information processing systems*, pages 831–837, 2001. 44

S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE transactions on pattern analysis and machine intelligence*, 24(4):509–522, 2002. 4

K. Birner. One click to empowerment?: Opportunities and challenges for labour in the global value chain of e-commerce. *International Journal of Labour Research*, 7(1/2):55, 2015.

F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016a. 2, 6

F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016b. 21, 24

F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989. 10, 64, 66, 97

A. Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019. 35

L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool. Apparel classification with style. In *Asian conference on computer vision*, pages 321–335. Springer, 2012.

Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 18, 20, 43, 92

S.-Y. Chen, K.-W. Tsoi, and Y.-Y. Chuang. Deep virtual try-on with clothes transform. In *International Computer Symposium*, pages 207–214. Springer, 2018. 3

W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3d pose estimation. In *3D Vision (3DV)*, 2016.

W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, and J. Liu. Fashion meets computer vision: A survey. *ACM Computing Surveys (CSUR)*, 54(4):1–41, 2021. 3

A. Chopra, R. Jain, M. Hemani, and B. Krishnamurthy. Zflow: Gated appearance flow-based virtual try-on with 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5433–5442, 2021. 5, 6

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 49

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 33, 34

F. Devernay and O. Faugeras. Straight lines have to be straight. *MVA*, 13:14–24, 2001.

G. Donato and S. Belongie. Approximate thin plate spline mappings. In *European conference on computer vision*, pages 21–31. Springer, 2002. 23

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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 9026–9035, 2019a. x, 2, 3, 5, 7, 8, 16, 17, 18, 28, 31, 32, 33, 35, 50, 51, 53, 55, 63, 64, 79, 95, 98, 100

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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 1161–1170, 2019b. 16

D. Dowson and B. Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982. 33

J. Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer, 1977. 3, 4, 8, 19, 23, 95

Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, and P. Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8485–8493, 2021. 5, 6

K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 932–940, 2017. 18, 19, 29, 43, 69, 70, 98

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 34, 40, 65

W. Grimson. *From images to surfaces: A computational study of the human early visual system*, volume 4. MIT press Cambridge, MA, 1981. 3, 4, 8, 95

M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the IEEE international conference on computer vision*, pages 3343–3351, 2015.

X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7543–7552, 2018. 2, 3, 4, 7, 8, 16, 17, 18, 31, 32, 33, 35, 38, 53, 63

X. Han, X. Hu, W. Huang, and M. R. Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10471–10480, 2019. 5, 6

R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.

S. He, Y.-Z. Song, and T. Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3470–3479, 2022. 5, 6

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 33, 53

C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, and W.-H. Cheng. Fit-me: Image-based virtual try-on with arbitrary poses. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4694–4698. IEEE, 2019a. 2, 3, 8, 16

C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, J. Liu, and W.-H. Cheng. Fashionon: Semantic-guided image-based virtual try-on with detailed human and clothing information. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 275–283, 2019b. 2, 3, 8, 16

Y. Hu, X. Yi, and L. S. Davis. Collaborative fashion recommendation: A functional tensor factorization approach. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 129–138. ACM, 2015.

J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE international conference on computer vision*, pages 1062–1070, 2015.

S. U. Islam, A. Glover, R. J. MacFarlane, N. Mehta, and M. Waseem. The anatomy and biomechanics of the elbow. *The Open Orthopaedics Journal*, 14(1), 2020. 73

P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 26, 98

T. Issenhuth, J. Mary, and C. Calauzènes. End-to-end learning of geometric deformations of feature maps for virtual try-on. *arXiv preprint arXiv:1906.01347*, 2019. 8, 17

H. Jae Lee, R. Lee, M. Kang, M. Cho, and G. Park. La-viton: A network for looking-attractive virtual try-on. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 8, 17

S. Jandial, A. Chopra, K. Ayush, M. Hemani, B. Krishnamurthy, and A. 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*, pages 2182–2190, 2020. 2, 3, 5, 8

N. Jetchev and U. Bergmann. The conditional analogy gan: Swapping fashion articles on people images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2287–2292, 2017. 3, 16, 65

J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 26, 27, 105

Y. Kalantidis, L. Kennedy, and L.-J. Li. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 105–112. ACM, 2013.

T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 6, 7

A. Kendall and R. Cipolla. Geometric loss functions for camera pose regression with deep learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

S. Kubo, Y. Iwasawa, and Y. Matsuo. Generative adversarial network-based virtual try-on with clothing region. 2018.

S. Kubo, Y. Iwasawa, M. Suzuki, and Y. Matsuo. Uvton: Uv mapping to consider the 3d structure of a human in image-based virtual try-on network. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

A. Lerios, C. D. Garfinkle, and M. Levoy. Feature-based volume metamorphosis. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 449–456, 1995. 68

K. M. Lewis, S. Varadharajan, and I. Kemelmacher-Shlizerman. Tryongan: body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG)*, 40(4):1–10, 2021. 3, 7, 65, 93

X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, and S. Yan. Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval. *IEEE Transactions on Multimedia*, 18(6): 1175–1186, 2016.

K. Lin, L. Wang, K. Luo, Y. Chen, Z. Liu, and M.-T. Sun. Cross-domain complementary learning using pose for multi-person part segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. ix, 70, 71, 72

G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018. 79, 104, 105

T. Liu, J. Zhang, X. Nie, Y. Wei, S. Wei, Y. Zhao, and J. Feng. Spatial-aware texture transformer for high-fidelity garment transfer. *IEEE Transactions on Image Processing*, 30:7499–7510, 2021. 65

W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5904–5913, 2019a. 5

Y. Liu, W. Chen, L. Liu, and M. S. Lew. Swapgan: A multistage generative approach for person-to-person fashion style transfer. *IEEE Transactions on Multimedia*, 21(9):2209–2222, 2019b.

Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016a. x, 28, 32, 79, 95, 99

Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang. Fashion landmark detection in the wild. In *European Conference on Computer Vision*, pages 229–245. Springer, 2016b.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 49

A. B. Mabrouk and E. Zagrouba. Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Systems with Applications*, 91:480–491, 2018.

N. Magnenat-Thalmann, B. Kevelham, P. Volino, M. Kasap, and E. Lyard. 3d web-based virtual try on of physically simulated clothes. *Computer-Aided Design and Applications*, 8(2):163–174, 2011.

J. A. Maintz and M. A. Viergever. A survey of medical image registration. *Medical image analysis*, 2(1):1–36, 1998. 19, 64, 72

F. Malagelada, M. Dalmau-Pastor, J. Vega, and P. Golano. Elbow anatomy. *Sports injuries: prevention, diagnosis, treatment and rehabilitation*, 2:527–53, 2014. ix, 72

V. Masteikaitě, V. Sacevičienė, and V. Čironienė. Compressed loop method for the bending behaviour of coated and laminated fabrics analysis. *Journal of Industrial Textiles*, 43(3):350–365, 2014. ix, 72

Y. Men, Y. Mao, Y. Jiang, W.-Y. Ma, and Z. Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020. 3, 7, 65

M. R. Minar and H. Ahn. Cloth-vton: Clothing three-dimensional reconstruction for hybrid image-based virtual try-on. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 26

A. Neuberger, E. Borenstein, B. Hilleli, E. Oks, and S. Alpert. Image based virtual try-on network from unpaired data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 35

A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. x, 26, 97, 106

T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 5

G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4):73, 2017.

A. H. Raffiee and M. Sollami. Garmentgan: Photo-realistic adversarial fashion transfer. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3923–3930. IEEE, 2021. 2, 3, 5, 8, 50, 51

A. Raj, P. Sangkloy, H. Chang, J. Hays, D. Ceylan, and J. Lu. Swapnet: Image based garment transfer. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*, pages 679–695, 2018.

I. Rocco, R. Arandjelovic, and J. Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6148–6157, 2017. 4, 8, 17, 49, 50, 52

D. Roy, S. Santra, and B. Chanda. Incorporating human body shape guidance for cloth warping in model to person virtual try-on problems. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2020. 2, 3, 8, 63, 64

D. Roy, D. Mukherjee, and B. Chanda. An unsupervised approach towards varying human skin tone using generative adversarial networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10681–10688. IEEE, 2021. 104

D. Roy, S. Santra, and B. Chanda. Lgvton: a landmark guided approach for model to person virtual try-on. *Multimedia Tools and Applications*, Jan 2022. ISSN 1573-7721. doi: 10.1007/s11042-021-11647-9. 2, 3, 8, 63, 64

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016. 33, 34

M. Sekine, K. Sugita, F. Perbet, B. Stenger, and M. Nishiyama. Virtual fitting by single-shot body shape estimation. In *Int. Conf. on 3D Body Scanning Technologies*, pages 406–413. Citeseer, 2014. 1

T. Shai, V. Dimitri, and U. Michael. Polyharmonic smoothing splines and the multidimensional wiener filtering of fractal-like signals. In *IEEE TRANSACTIONS ON IMAGE PROCESSING*. IEEE, 2006.

Y. Shigeki, F. Okura, I. Mitsugami, and Y. Yagi. Estimating 3d human shape under clothing from a single rgb image. *IPSJ Transactions on Computer Vision and Applications*, 10(1):16, 2018. 24

D. Song, T. Li, Z. Mao, and A.-A. Liu. Sp-viton: shape-preserving image-based virtual try-on network. *Multimedia Tools and Applications*, pages 1–13, 2019. 16, 18

R. Sprengel, K. Rohr, and H. S. Stiehl. Thin-plate spline approximation for image registration. In *Proceedings of 18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 3, pages 1190–1191. IEEE, 1996. 16, 23, 52

F. Sun, J. Guo, Z. Su, and C. Gao. Image-based virtual try-on network with structural coherence. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 519–523. IEEE, 2019. 16

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 33, 34

Y. R. Tsai. An efficient and accurate camera calibration technique for 3D machine vision. In *Proc. CVPR*, 1986.

G. Wahba. *Spline models for observational data*, volume 59. Siam, 1990. 9, 23, 96

B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 589–604, 2018a. 2, 3, 4, 7, 8, 16, 17, 18, 31, 32, 35, 49, 53, 55, 56, 63, 64, 65

W. Wang, Y. Xu, J. Shen, and S.-C. Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4271–4280, 2018b.

X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018c. 2, 3, 4, 8

Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 27, 33, 52, 53, 106

Z. Wu, G. Lin, Q. Tao, and J. Cai. M2e-try on net: Fashion from model to everyone. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 293–301. ACM, 2019. 6, 17, 31, 35, 53, 65

W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu, and J. Hays. Texturegan: Controlling deep image synthesis with texture patches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2018. 27, 98

Z. Xie, Z. Huang, F. Zhao, H. Dong, M. Kampffmeyer, and X. Liang. Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan. *Advances in Neural Information Processing Systems*, 34:2598–2610, 2021. 5, 35

H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. 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. 2, 3, 4, 7, 8, 16, 17, 18, 63, 64

G. Yildirim, N. Jetchev, R. Vollgraf, and U. Bergmann. Generating high-resolution fashion model images wearing custom outfits. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3, 7, 65

R. Yu, X. Wang, and X. Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10511–10520, 2019. 2, 3, 4, 7, 8, 16, 17, 31, 32, 50, 51, 63, 64

M. Zanfir, A.-I. Popa, A. Zanfir, and C. Sminchisescu. Human appearance transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018a.

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, 2018b.

W. Zeng, M. Zhao, Y. Gao, and Z. Zhang. Tilegan: category-oriented attention-based high-quality tiled clothes generation from dressed person. *NEURAL COMPUTING & APPLICATIONS*, 2020. 5

B. Zhao, J. Feng, X. Wu, and S. Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1520–1528, 2017.

R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, 2013.

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, 2019a. 16

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, 2019b. 3

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3