Envisioning the Future of Fashion: The Creation And Application Of Diverse Body Pose Datasets for Real-World Virtual Try-On

WITS
UNIVERSITY

Molefe Molefe
1858893

Supervised by Prof. Richard Klein

August 7, 2024

A dissertation submitted to the Faculty of Science, University of Witwatersrand, in fulfilment of the requirements for the degree of Master of Science.

Ethical Clearance Number: CSAM-2022-09

Abstract

Fashion presents an opportunity for research methods to unite machine learning concepts with e-commerce to meet the growing demands of consumers. A recent development in intelligent fashion research envisions how individuals might appear in different clothes based on their selection, a process known as "virtual try-on". Our research introduces a novel dataset that ensures multi-view consistency, facilitating the effective warping and synthesis of clothing onto individuals from any given perspective or pose. This addresses a significant shortfall in existing datasets, which struggle to recognise various views, thus limiting the versatility of virtual try-on. By fine-tuning state-of-the-art architectures on our dataset, we expand the utility of virtual try-on, making them more adaptable and robust across a diverse range of scenarios. A noteworthy additional advantage of our dataset is its capacity to facilitate 3D scene reconstruction. This capability arises from utilising a sparse collection of images captured from multiple angles, which, while primarily aimed at enriching 2D virtual try-on, inadvertently supports the simulation of 3D environments. This enhancement not only broadens the practical applications of virtual try-on in the real-world but also advances the field by demonstrating a novel application of deep learning within the fashion industry, enabling more realistic and comprehensive virtual try-on experiences. Therefore, our work heralds a novel dataset and approach for virtually synthesising clothing in an accessible way for real-world scenarios.

Acknowledgements

"If I can see further, it is only because I stand on the shoulders of giants" - Sir. Isaac Newton.

This quote represents the tulmuteous journey that I embarked during my pursuit of this esteemed degree at the prestigous university, The University of the Witwatersrand. The intent of our research was to innovate meaningful technology that will usher a new era of how we interact with technology. Through this journey I have been fortunate to have the support of many individuals who have been instrumental in my academic and personal growth. These are the individuals who have been the giants on whose shoulders I have stood. Their names are too numerous to mention, but I would like to express my gratitude to them for their unwavering support and encouragement.

To my supervisor, every momemt spent with you has been a learning experience. Your guidance and support has been invaluable. I am grateful for the opportunity to have worked with you. Prof. Richard Klein gave an introductory lecture to *Introduction To Data Structures and Algorithms*, and in this very lecture I decided, not knowing anything about them at all, that I would pursue a master's degree in Computer Science. I am grateful for the opportunity to have been taught by you.

To my father, who tragically passed away recently during this program, your untimely death made me suffer greatly and delve into deep despair, I miss you everyday. I am grateful for the sacrifices you made to ensure that I had the best education. I am grateful for the values you instilled in me, and I hope that I have made you proud. To my mother, who has been my pillar of strength, I am grateful for your unwavering support and encouragement. To my wife, I am eternally grateful for your love and support. You have been my rock during this journey.

This dissertation is a testament that nothing can stand in the way of the course of destiny that God has placed in your life, not even death. To my Lord and saviour Jesus Christ, I am honoured to have had this wonderful opportunity.

Declaration

I, Molefe Reabetsoe-Phenyo Mothlabane Molefe, hereby declare the contents of this research dissertation to be my own work. This dissertation is submitted for the degree of Master of Science in Computer Science at the University of the Witwatersrand. This work has not been submitted to any other university, or for any other degree.

Signature: 

Date: 2024/08/07

Contents

List of Figures

List of Tables

Nomenclature

In this section, we explain the symbols and terms utilised throughout this dissertation. We train various virtual try-on architectures that each utilise their own notations. To unionise a standardised representation, we have structured the nomenclature into distinct tables.

Firstly, we describe the notation that labels the inputs of each stage that help us understand the overall outline of virtual try-on.

Introduction to Symbols and Notations

We commence by detailing the symbols denoting the inputs at various stages of virtual try-on, offering insight into the system's overall framework. This foundational knowledge is pivotal for comprehending the successive stages of virtual garment fitting.

Table 1: Symbols that help us understand the inputs to the network.

Symbol	Description
	Inputs of Virtual Try-On
I_p	Individual in the reference image
$(.)_p$	The subscript p denotes the individual in the reference image
$\otimes$	Represents channel-wise concatenation
c	Clothing item of the individual in the reference image
c_{un}	Unpaired clothing item represents a clothing that the individual is not wearing
c_2	Second clothing item whereby there already exists a separate clothing item in the pipeline
c_m	Binary mask of the clothing item that the individual in the reference image is wearing
c_e	Warped version of clothing item c extracted from the individual I_p that is used to represent the ground truth for warping networks. This is usually obtained by cropping the clothing item from the individual in the reference image I_p.
c_{e_m}	Binary mask of the warped clothing item representing the predicted output
c_t	Target clothing(in-shop) item representing the ground truth. In a paired setting, $c_t = c$ such that the separate target clothing item is the same clothing item that the individual in I_p is wearing.

The table below outlines symbols crucial for interpreting parser-based information fed into the network, enhancing the virtual try-on process. It includes representations and

data forms that strip away clothing characteristics to focus on body shapes, poses, and segmentations essential for accurate garment fitting

Table 2: Symbols that help us understand the parser-based information to the network.

Symbol	Description
I_{car}	Clothing agnostic representation of the individual in the reference image. This means that the region of interest, in most cases this refers to the items corresponding to the torso for upper-body virtual try-on, is removed from the individual in the reference image I_p
S_{car}	Clothing agnostic segmentation of the individual in the reference image. This means that the region of interest, are given segmentation labels
S	Segmentation label
P_D	Densepose representation of the individual in the reference image
P_P	2D pose estimation of the individual in the reference image

The table below delineates symbols pivotal for understanding the outputs generated by virtual try-on networks. It encompasses various forms of output, including warped clothing items, appearance flows, and segmentation maps, each contributing to the final visualisation of the digital try-on experience, showcasing how garments adjust to and appear on a virtual model.

Table 3: Symbols that help us understand the outputs of the networks.

Symbol	Description
Outputs of Virtual Try-On	
c_w	Predicted warped clothing item representing the predicted output
c_f	Predicted apperance flow that is used to warp a clothing item. Typically, they refer to apperance flows that represent a set of 2D coordinate vectors, also known as dense correspondences, that map clothing pixels to where they should be present in the warped clothing image.
c_{w_m}	Binary mask of the predicted warped clothing item representing the predicted output
$\hat{I}_p$	Individual from the reference image wearing the target image. This notation is used in the ideal scenario that we have $(I_p, c_t) \rightarrow I_p$ such that I_p is the same individual from I_p but wearing c_t.
$\hat{I}_o$	Final result that is composed of the warped clothing item c_w and the rendered individual $\hat{I}_p$. Often used interchangeably with $\hat{I}_p$
$\hat{S}_p, \hat{S}$	Segmentation map of the individual wearing the warped clothing c_w.
$\hat{S}_{raw}$	Segmentation map generated by the last layer of the feature fusion block

The table below presents symbols fundamental to understanding the loss functions used in the optimisation of virtual try-on systems. These symbols represent key mathematical concepts and elements, such as coefficients, operations, and specific types of data, that are crucial for the formulation and calculation of loss functions in the model's training process.

Table 4: Symbols that are commonly used in loss functions.

Symbol	Description
λ	Coefficient that behaves as a hyperparameter for any loss term
$\odot$	Element-wise multiplication
g_t	Ground truth
$\phi_i(I)$	denotes the feature map of Image I of the i-th layer in the visual perception network

The table below explains the symbols that represent the outcomes of *High-Resolution Virtual Try-On with Misalignment and Occlusion-Handling* designed to seamlessly integrate clothing onto individuals.

Table 5: Symbols related to High-Resolution Virtual Try-On with Misalignment and Occlusion-Handling[1]

Symbol	Description
$W(c_m, F_{f_4})$	Warped clothing item representing the predicted output of the condition generator
$\hat{S}$	Segmentation map containing the warped clothing segmentation labels predicted by the condition generator
$\hat{S_{raw}}$	Segmentation map generated by the last layer of the feature fusion block

The table below highlights symbols central to the understanding of parser-free virtual try-on networks. It encapsulates various outputs and methodologies, including the differentiated roles of teacher and student networks within the learning process.

Table 6: Symbols related to Parser-Free networks

Symbol	Description
g	Warped output of the parser-free network
s_I	output of the student network
f_i^t	denotes the i-th point on the flow maps of i-th scale
$\mathcal{P}$	is the generalised Charbonnier loss function
t_{pi}	denote the feature maps at the $i-th$ scale extracted from $p*$ within the Teacher ntework and the synthetic image t within the Student
t	try-on result of the Teacher
s	try-on result of the Student

Chapter 1

Introduction

Fashion not only mirrors how individuals choose to represent themselves through their attire but also serves as a stylistic expression of their character. Historically, clothing was predominantly handwoven, serving the primary function of appropriate dressing. However, the advent of the Industrial Revolution catalysed innovations that allowed individuals to express their creativity in their appearance more freely. Consequently, the fashion industry has emerged as a significant entity within the global economy, encompassing areas such as textile production, manufacturing, design, marketing, retailing, and advertising [Ross 2022].

In today's fast-paced world, the demand for convenience, personalisation, and sustainability in the fashion sector is more pronounced than ever. With the rise of online shopping and mixed reality, there's a growing need for technologies that enable customers to virtually try on clothing before making a purchase. Integrating artificial intelligence with virtual try-on offers an immersive experience that allows consumers to visualise themselves in various clothing remotely. To implement this, researchers [Han *et al.* 2018; Gupta *et al.* 2024; Song *et al.* 2023] have developed sophisticated computer vision techniques to address specific challenges, including virtual try-on, visual search, and forecasting clothing trends to enhance the online shopping experience. The application of virtual try-on within mobile applications represents a significant and rapidly growing market segment with profound implications for the fashion and retail industries.

Virtual try-on involves a process where computer vision algorithms digitally superimpose clothing onto images of people, aiming to simulate how these clothes would look in real life. This process involves two main steps. First, warping algorithms, such as thin-plate spline transformations, adjust the clothing's shape to fit the person's pose in the image. Next, generative networks are trained to overlay the adjusted clothing onto the person. Although previous studies [Liu *et al.* 2016a; Han *et al.* 2018; Choi *et al.* 2021] have developed fashion datasets to help improve this technology, we introduce a novel dataset that includes a broader range of body positions and diverse perspectives, enhancing the technology's ability to handle different real-world situations. We illustrate the shortcomings of existing datasets and introduce a pilot study that integrates a 3D scene reconstruction with virtual try-on using our dataset, making it possible to create more realistic 3D clothing try-ons from 2D images. This breakthrough allows researchers to extend the utility of virtual try-on capabilities beyond frontal-facing images.

This dissertation focuses on adapting virtual try-on technologies to meet real-world demands effectively. While existing technologies [Santesteban *et al.* 2022; Guan *et al.* 2012; Pons-Moll *et al.* 2017] allow users to visualise clothing onto themselves, they often rely on complex 3D computer graphics techniques. These methods not only require specialised skills but are also costly to implement. Our research endeavors to introduce a virtual try-on solution that addresses these real-world challenges by circumventing the need for resource-intensive 3D modeling techniques. Fundamentally, our research primarily focuses on two problems critical for introducing diverse body orientations and camera positions for the task of virtual try-on:

- **Ineffective datasets**: Currently, researchers tackle 2D virtual try-on using datasets consisting of paired images of clothing items and individuals facing the camera [Han *et al.* 2018; Dong *et al.* 2019; Choi *et al.* 2021]. However, these datasets are not suitable for 3D scene reconstruction, limiting the applicability of virtual try-on technologies to a fewer array of poses and positions [Song *et al.* 2023].
- **3D virtual try-on**: Although 3D virtual try-on techniques are reliant on computer graphics pipelines to render scenes that contain 3D shape representations for both humans and clothing, our approach mitigates this dependency. By providing 2D image frames of varying angles, we enable 3D scene reconstruction using neural radiance fields as a by-product. Our primary focus remains on addressing the limitation of frontal-facing datasets, thus allowing clothing items to be superimposed from various angles, beyond just frontal-facing images.

Figure 1.1 highlights the significant contribution of our novel dataset by comparing the virtual try-on's ability to warp clothing items. Conventional datasets fail to recognise a change in posture and angle for the same garment. By introducing the **PoseCam** (**POSE**-Adaptive **CAM**era-Angle) dataset, we aim to extend the capability of virtual try-on systems to foster a more dynamic and realistic interaction with virtual garments, accommodating an extensive array of poses and body movements.

Figure 1.1: Here, we compare the qualitative performance of training state-of-the-art warping architectures on the VITON-HD [Choi *et al.* 2021] dataset and our novel dataset. The figure demonstrates the significant improvement that our novel dataset contributes to the task of warping through more diverse training examples.

In virtual try-on research, the capability to handle a variety of poses and viewing angles is paramount to providing an immersive and realistic user experience. Figure 1.2 depicts how traditional datasets [Han *et al.* 2018; Dong *et al.* 2019; Choi *et al.* 2021] often fall short in representing the complex interplay of clothing with diverse human postures and viewpoints. This limitation becomes apparent when the architectures attempt to synthesise warped clothing items on individuals, leading to inaccuracies and, at times, unrealistic garment fittings when subjects shift their poses or change their

orientation relative to the camera. Current datasets limit virtual try-on from achieving consistent and accurate warping of clothing items across different perspectives of the same individual. Consequently, this inadequacy directly impacts the performance of the composition network tasked with synthesising the clothing item onto the avatar, resulting in misaligned and unconvincing garment overlays that detract from the realism and usability of virtual try-on applications.

Figure 1.2: Results without the **PoseCam** dataset: This shows the model(using DM_VTON) generated composition with less visual consistency compared to **PoseCam** finetuning. The image quality and transformation fidelity are less refined, indicating inferior learning from the original training checkpoints.

In this work, highlighted in Figure 1.3, we demonstrate the outperformance of state-of-the-art virtual try-on architectures when trained on our novel dataset. Our novel dataset targets this specific challenge by enriching the training examples with multi-angle representations. This approach allows the virtual try-on system to adapt to varying angles, ensuring that the clothing items retain their realistic drape and fit regardless of the subject's orientation relative to the camera.

Figure 1.3: Results with **PoseCam** dataset: This shows the result of the model(DM_VTON) generated output of the individual wearing the jacket. The clarity and consistency in the visual quality illustrate the effectiveness of finetuning on the **PoseCam** dataset.

By fine-tuning our network on this comprehensive dataset, we have successfully addressed the problem of angle-specific rigidity that often plagues virtual try-on systems trained on traditional datasets.

Our work broadens the scope to include diverse body orientations, enhancing the applicability of virtual try-on to a wider array of poses and positions. This adaptation significantly augments the utility of virtual try-on technologies for real-world applications. As such, our contributions are as follows:

- Introduction of a novel **virtual try-on dataset** comprising of image pairs with diverse camera angles. The dataset also supports 3D scene reconstruction, facilitating a more immersive and diverse fitting experience.
- A novel **inference framework** that allows for the integration of neural radiance fields with virtual try-on to simulate 3D virtual try-on using 2D images.

Given our main contributions, we limit our methodology to introducing a novel dataset without alterations to the state-of-the-art architectures used to produce the experiment results. Furthermore, we posit a potential application of the dataset to integrate 2D virtual try-on with 3D scene reconstruction to simulate real-world virtual try-on.

As such, the dissertation is structured as follows: We begin with Chapter 2, where we delve into the intersection of fashion and computer vision to enable applications from outfit recommendations to virtual clothing generation. The discussion focuses on virtual try-on, which lets individuals visualise themselves in different outfits digitally, eliminating physical trials. We outline the virtual try-on pipeline, including the warping and composition stages, and detail the generative algorithms required to adjust and synthesise clothing onto an individual's body. Additionally, we emphasise the need for diverse datasets to train models effectively, discussing existing datasets and the need to expand these to improve virtual try-on performance and accessibility. This chapter aims to provide a concise overview of computer vision's role in fashion, particularly in developing and implementing virtual try-on for real-world scenarios, highlighting the technical challenges and the importance of robust datasets.

In Chapter 3, we provide a detailed examination of various datasets pivotal to the development and enhancement of virtual try-on, articulating their contributions, limitations, and how a novel dataset could propel the field forward.

Our contribution is detailed in Chapter 4, delving into the detailed steps required to collect the **PoseCam** dataset.

Rather than modifying existing architectures, our strategy focuses on applying state-of-the-art virtual try-on architectures in a new context. Using these techniques, Chapter 5 explains how we set up the experiments to facilitate the training and evaluation procedure for optimal performance. The results are expressed in Chapter 6, where we present qualitative and quantitative analyses of finetuning the **PoseCam** dataset for virtual try-on.

To demonstrate the extensibility of our dataset, Chapter 7 introduces a pilot study regarding how the **PoseCam** dataset allows us to reconstruct 3D virtual try-on scenes using novel view rendering techniques for real-world virtual try-on.

Finally, we conclude the dissertation in Chapter 8 by providing ideas for future work, inspiring continued innovation and exploration in this dynamic field, and reflect on the evolution of fashion technology and our part to play in advancing artificial intelligence in the fashion industry. [1]

[1] Research code and implementation is available at https://github.com/mo-rsa24/Fashion-NeRF

Chapter 2

Background

The integration of fashion and computer vision technology represents a significant advancement in enhancing consumer interaction and personalisation in the apparel industry. Three key techniques in fashion research are fashion detection, fashion analysis, and fashion synthesis. Fashion detection involves the identification of clothing items and human body parts within images, enabling the extraction of garment attributes and styles. This includes techniques such as fashion parsing, which use pixel-level classification within images through semantic segmentation to associate clothing items with specific body parts. Another aspect of fashion detection is item retrieval [Kalantidis *et al.* 2013], which involves extracting specific garments from images to find similar items from a gallery. A common technique segments images into superpixels that are used to predict clothing labels [Yamaguchi *et al.* 2012].

Fashion analysis refers to the process of extracting and interpreting detailed attributes and features of fashion items, such as colour, texture, style, and fit, to gain insights into trends, consumer preferences, and design elements. It encompasses the annotation and examination of fashion images to discern style trends and attributes, including popularity prediction [Massip *et al.* 2018], which leverages a vision-based approach to determine how various factors influence the appeal and subsequent success of outfit images online. Kalantidis *et al.* [2013] analyse fashion model popularity trends by studying the correlation between social media metrics (post frequency, engagement levels, and visual appeal) and the rising popularity of fashion models in the industry.

Finally, fashion synthesis refers to the process of using generative models and style transfer techniques to create new and diverse visual representations of fashion items. This technique allows for the transformation and reimagining of existing images. Virtual try-on [Han *et al.* 2018], a fashion synthesis technique, leverages generative algorithms to enable users to virtually try on clothing items without physically wearing them. Collectively, these techniques illustrate the pivotal role of computer vision in revolutionising fashion retail, showcasing the potential for increased efficiency, customer satisfaction, and environmental sustainability [Cheng *et al.* 2021].

2.1 Models For Image-Based Virtual Try-On

The core objective of virtual try-on systems is to accurately render clothing items on a digital avatar such that they appear realistic and adhere to the complex interplay of fab-

rics, body dynamics, and lighting. Achieving this level of realism requires sophisticated image synthesis and manipulation techniques powered by deep learning.

Image-based virtual try-on, as demonstrated in Figure 2.1, is a specialised task within the broader field of fashion synthesis. This process involves overlaying a target piece of clothing onto a person depicted in a reference image. The goal is to realistically adapt the target clothing to the body pose and shape of the individual in the reference image, resulting in a seamless garment transfer. The pipeline begins with identifying the key inputs: a reference image and a target clothing image. The reference image is a photo of a person, serving as the base onto which the target clothing will be applied. The target clothing image is a separate image of the clothing item that needs to be virtually tried on by the person in the reference image. The output of this process is a synthesised image showing the reference person wearing the target clothing, with the garment realistically warped and integrated according to the person's pose and body shape.

Figure 2.1: Synthensise target clothing onto an individual, given the reference image and the in-shop clothing item separately [Choi *et al.* 2021].

Virtual try-on is complex, with numerous challenges such as the diverse range of clothing styles, body shapes, and poses, among many others. To successfully perform virtual try-on, several challenges need to be addressed. First, clothing warping techniques must effectively deform the target clothing image to match the contours and pose of the reference person. Next, the garment transfer process must seamlessly blend the warped clothing into the reference image, requiring precise adjustments for lighting, shading, and texture to maintain realism. Resolution constraints must be overcome to ensure both the reference and clothing images are of high quality, achieving a visually appealing result. Additionally, occlusion handling is crucial to address parts of the body or background that might obscure sections of the clothing, ensuring these issues are managed smoothly. Ensuring consistency in style and colour is also vital, as the clothing's design, texture, and hue must match the reference image to maintain a cohesive appearance. Finally, providing user interactivity is essential, allowing users to adjust parameters such as clothing size, orientation, and placement to better fit their preferences.

Deep learning models, especially Generative Adversarial Networks (GANs) and Conditional GANs (cGANs), have emerged as pivotal tools for tackling this challenge [Han *et al.* 2018; Raj *et al.* 2018]. These networks are required to learn how to generate images that capture the unique attributes of the clothing and adapt seamlessly to the

individual's body, pose and the surrounding environment. In this context, the generative model attempts to produce images that are visually indistinguishable from real photographs of individuals wearing the specified clothing items. The challenge lies in training the generative model to capture the fine-grained details of clothing, adapt it to different body shapes and poses, and ensure that the resulting images are perceptually realistic [Choi *et al.* 2021].

Ultimately, the goal is to create a generative model that can generate diverse and photorealistic images of virtual individuals wearing various clothing items, advancing the field of generative image synthesis within the context of virtual fashion try-on. This necessitates addressing limitations prevalent in existing datasets to enhance character preservation such that an individual's representation is maintained independent of diverse body poses, camera orientation, and occlusion (clothing regions obstructed by hair, arms, or obstacle) [Han *et al.* 2018; Shim *et al.* 2023; Choi *et al.* 2021].

2.1.1 Virtual Try-On Pipeline Overview

Virtual try-on necessitates a generative pipeline, depicted in Figure 2.2, that requires image pairs comprised of a clothing item I_c and an individual I_p. The goal is to generate an image I_p where the clothing item is seamlessly superimposed onto the individual. To achieve this, the pipeline is divided into two focused stages, each addressing a specific component of the challenge. To explain the models required for virtual try-on in this section, we discuss the details of the input representations extracted from the dataset in Section 2.2. Thus, the remainder of this section discusses, without detail for the required dataset, a high-level overview of the virtual try-on pipeline.

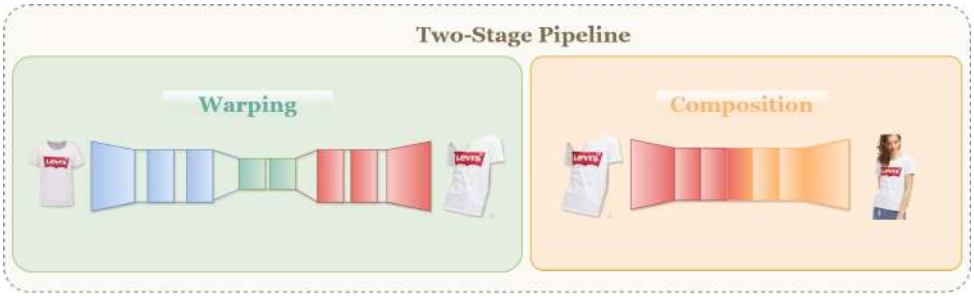


Figure 2.2: This figure represents the virtual try-on two-stage pipeline that takes as input an image pair, comprised of an individual I_p and their corresponding clothing item I_c, and synthesises the warped garment on the individual's region of interest.

Figure 2.3 illustrates the initial phase, known as the warping stage, of the pipeline that is concerned with adjusting the clothing image to fit the individual's body shape and pose accurately. The inputs for this stage include the clothing image with its corresponding mask and the individual's descriptive body features, represented as the concatenation of landmarks and segmentation labels acquired using off-the-shelf networks, generating a person representation. These person representations are used to construct clothing-agnostic representations that guide the network to localise regions of interest for image composition. Techniques such as Thin Plate Splines (TPS), Spatial Transformer Networks (STN), and FlowNet are employed to deform the clothing image, ensuring it conforms to the individual's body contours. The output of this stage is a warped clothing image that aligns with the individual's pose and physique. Following the cloth warp-

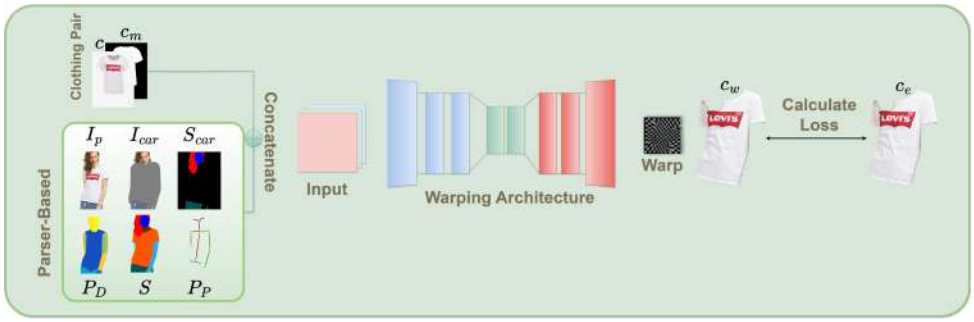


Figure 2.3: This figure demonstrates the first stage in virtual try-on that warps garments to match an individual's pose. The warping network takes as input the clothing image with its corresponding mask and the individual's descriptive body features, known as the person representation, to generate a warped clothing image.

ing phase, the composition stage depicted in Figure 2.4, synthesises the final image by merging the warped clothing image with the individual's image. This stage utilises interpolation methods or generative networks to integrate the clothing onto the individual's body in a realistic and natural manner. The resulting image should preserve the original context of the individual's surroundings, with modifications limited to the area where the new clothing is placed. The division of the virtual try-on pipeline into dis-

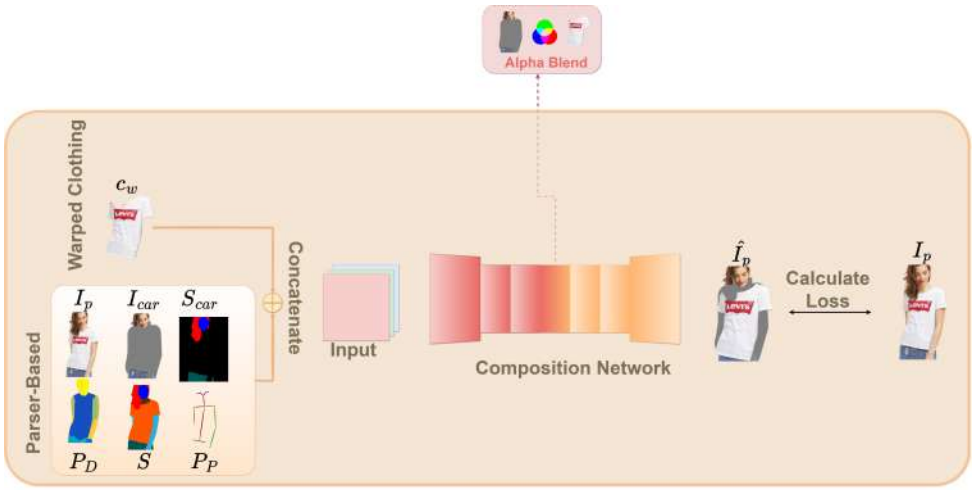


Figure 2.4: Following Figure 2.3, the second stage in virtual try-on utilises the warped clothing item and superimposes it on the individual using a generative network.

tinct stages—cloth warping and try-on synthesis—is crucial for several reasons. Firstly, it simplifies the overall complexity of the task by breaking it down into manageable components. This separation allows for targeted optimisations and adjustments within individual stages without necessitating a complete overhaul of the entire pipeline. Additionally, it accommodates the use of specialised techniques and technologies that are best suited to the specific challenges presented by each stage.

In an ideal scenario, the experimental setup for training a virtual try-on network involves the reconstruction of an individual wearing specific attire based on an input tuple (I_p, c) to generate the output image $\hat{I}_p$. This process entails training separate networks to warp the clothing item c to align with the pose of the reference image I_p, thereby

synthesising the target image $\hat{I}_p$. As depicted in Figure 2.5, this scenario exemplifies virtual try-on as a supervised image reconstruction task.

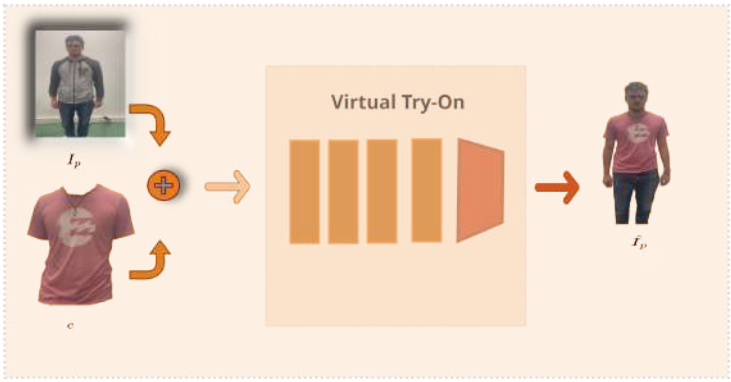


Figure 2.5: Given input pair (I_p, c), we train a virtual try-on network to reconstruct an image of the individual I_p wearing c. In this instance, the underlying assumption is that we have access to the ground truth $\hat{I}_p$ representing the same individual from I_p but wearing c.

Figure 2.6 illustrates a practical scenario for virtual try-on, unlike in a traditional paired setting of supervised learning, where the input image I_p also serves as the ground truth. This necessitates the network's capability to generalise across unpaired image-clothing combinations. Consequently, virtual try-on datasets typically comprise of image pairs of individuals and their corresponding clothing items, predominantly sourced from fashion websites through web scraping. These datasets often feature images with subjects directly facing the camera, which limits the scope of virtual try-on techniques to synthesising clothing primarily for frontal views. If the clothing item is unavailable during training, we extract the clothing item from the individual's image using segmentation masks. As a result, the individual I_p and the clothing item c are concatenated channel-wise to serve as input to the virtual try-on network to reconstruct the original image I_p.

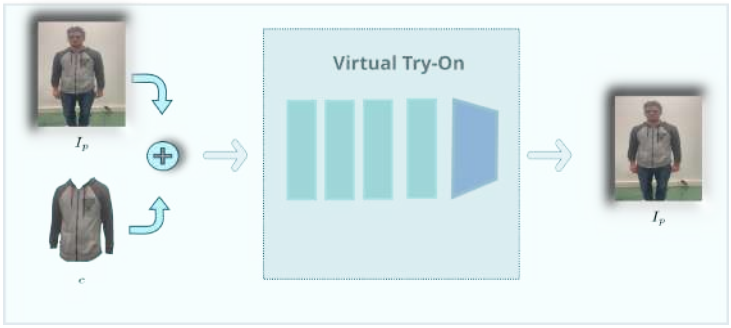


Figure 2.6: This figure demonstrates how reframing the task, virtual try-on imitates self-supervised learning. Due to the scarcity of datasets featuring images of individuals unpaired clothing items, virtual try-on techniques synthesise warped clothing items onto individuals and treat the input individual as the ground truth.

We have demonstrated the reframing of virtual try-on to imitate a self-supervised representation task such that the network reconstructs the ground truth from the paired

input images. As such, Figure 2.7 demonstrates how virtual try-on scenarios can be categorised under two settings: the paired setting, where the ground truth is available, and the unpaired setting, where the ground truth is not available. The paired setting, which mimics self-supervised learning, scenario is the most common in virtual try-on, as it is challenging to obtain images of individuals with both paired and unpaired clothing items.

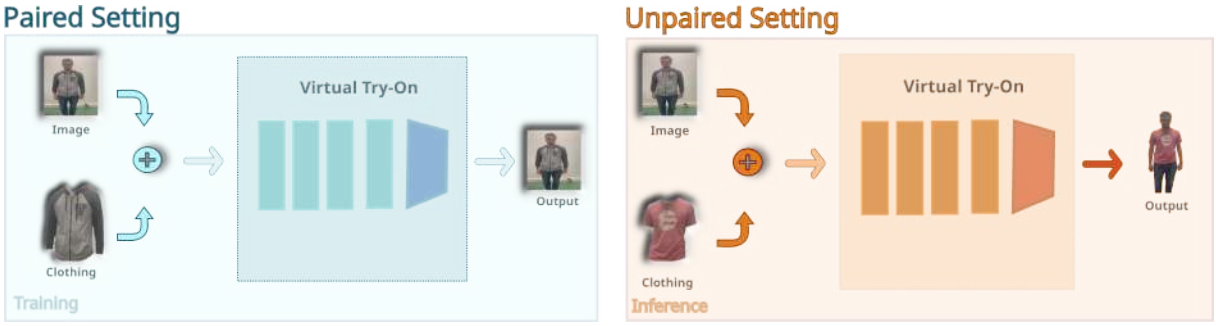


Figure 2.7: Here, we demonstrate the difference between paired and unpaired settings. The difference lies in the whether the clothing item the individual is wearing in the reference image is unavailable. Due to the scarcity of datasets featuring images of individuals with and without paired clothing items, the virtual try-on task imitates a self-supervised learning problem by reconstructing the ground truth(the try-on image) from the paired input images (the clothing and the individual).

2.1.2 How Virtual Try-On Warps Clothing For Image Composition

In the realm of computer graphics and image processing, warping refers to the technique of modifying an image or polygon to adapt to a new shape or form. This process typically involves a range of transformations, including rotation, scaling, translation, and skewing, to achieve the desired alteration in appearance or orientation. In a more advanced form, warping can involve the use of complex mathematical models to ensure that the transformed image maintains a realistic perspective and proportion when applied to a new surface or in a new context. In Figure 2.8, we consider a high-level overview of each virtual try-on stage - neglecting the complex input representation and architectural descriptions required to train each network.

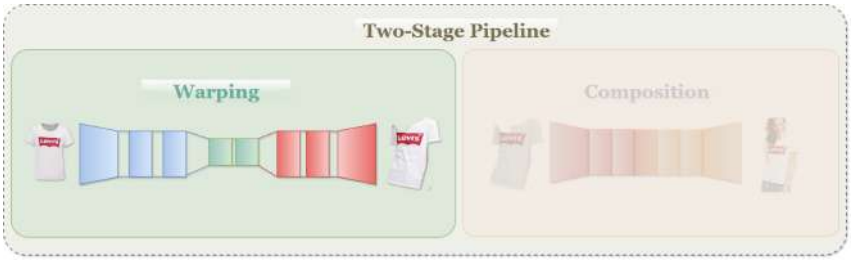


Figure 2.8: In this figure, we simplify the input representation to primarily focus on the methodolgy that allows warping to transform clothing items. Warping algorithms comprise of three techniques: Thin Plate Spline Transformation, Spatial Transformer Networks, and Flow Estimation, each of which transforms clothing items to match the pose of the individual.

Specifically, in virtual try-on systems, warping is a critical step that allows clothing items to fit onto the body of an avatar or person in a realistic manner. This process takes into account the pose and proportions of the individual, adjusting the clothing item so that it drapes naturally over the avatar's body, following the contours and natural lines of movement. The goal is to simulate how the fabric would behave in the real world, taking into consideration factors like stretching, compression, and folding. It manipulates pixels to distort the perception of the objects. Consequently, image warping serves as a foundational prerequisite for the synthesis of clothing items on individuals.

Thin-Plate Spline Transformation

Thin-Plate Splines (TPS) interpolation is a mathematical method used in computer graphics and image processing to model smooth surface deformations. This technique is particularly effective when working with scattered data points, known as a set of control points. The fundamental principle of TPS is to interpolate a surface that not only passes through these control points but also minimises the amount of bending or warping, hence the analogy to a thin metal plate that has been bent to fit through the points.

The process of TPS, illustrated in Figure 2.9, involves calculating a smooth, minimal-energy surface based on the positions of the control points. Unlike simpler interpolation methods that might produce jagged or unrealistic transformations, TPS ensures that the resulting deformed surface maintains a natural and coherent appearance. This is achieved by minimising a mathematical measure of the surface's curvature, ensuring the deformation doesn't introduce unnecessary complexity or distortion.

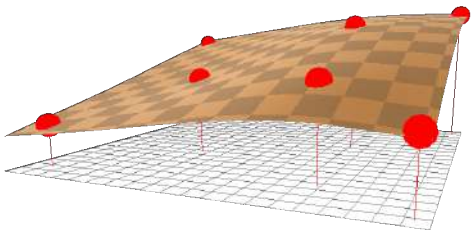


Figure 2.9: A least bent surface illustrating a thin plate spline passing through a set of control points. Each control point is represented by a red circle on the bent surface with a defined height above the grid.

The interpolation function $f(x, y)$ can be represented as:

$$f(x,y) = a_1 + a_x x + a_y y + \sum_{i=1}^{p} w_i U\left(\|(x_i, y_i) - (x, y)\|^2\right) \tag{2.1}$$

where the terms are categorised in the following manner:

1. $\mathbf{p}$: is the number of control points.
2. $(\mathbf{X}, \mathbf{Y})$: are the grid representation.
3. $\boldsymbol{U}(||(\boldsymbol{x_i}, \boldsymbol{y_i}) - (\boldsymbol{x}, \boldsymbol{y})||)$: represents the distance between a control point and a position (x, y) such that:
   (a) $\mathbf{U}(r) = r^2 logr$: is the radial basis function.
4. $\boldsymbol{w_i}$: denotes the coefficient of each control point.

5. $\boldsymbol{a_1}, \boldsymbol{a_x}, \boldsymbol{a_y}$: are the coefficients of the representing the linear plane approximating all control points.

In the context of image processing, TPS can be used to map pixels from an original image to a new, deformed image. This mapping takes into account the geometric transformations needed to move pixels from their original positions to new positions defined by the deformation, as depicted in the following equation [Bookstein 1989].

$$f : (x, y) \longrightarrow (x', y'), \quad x, y, x', y' \in \mathbb{R} \tag{2.2}$$

where (x, y) describe the pixel location capable of representing control points with (x', y') representing coordinates in the deformed image. Given a 2D image, the thin-plate spline warping algorithm interpolates a smooth function over a discrete set of control points (pixel locations) to sample for displacements. This means that for every pixel (x, y) in image A, we need to know where it will be located in image B at position $(x_i + \Delta x_i, y_i + \Delta y_i)$. In this case, we solve for two smooth functions that sample Δx_i and Δy_i that represent the displacements in the x and y directions, respectively. Deforming an image using thin-plate splines, as shown in Figure 2.10, modifies Equation 2.1 to interpolate a surface height for each displacement such that [Lombaert 2002]:

$$f_{x'}(x, y) = a_1 + a_x x + a_y y + \sum_{i=1}^{p} w_i U\left(\|(x_i, y_i) - (x, y)\|^2\right) \tag{2.3}$$

$$f_{y'}(x, y) = a_1 + a_x x + a_y y + \sum_{i=1}^{p} w_i U\left(\|(x_i, y_i) - (x, y)\|^2\right) \tag{2.4}$$

The above equation is subject to the presupposition that coordinates (x, y) are independent of each other. This means that the displacement of a pixel in the x-direction renders the displacement of the y-direction as nought and vice-versa.

Figure 2.10: Deforming an image using thin plates [Lombaert 2002].

Figure 2.11 illustrates the deformation of an image using thin-plate splines to interpolate two surfaces in each direction to sample for displacements. The control points are displaced to create a smile in the original image such that the displacement of each pixel is determined by the height of the surface at the pixel's location. For example, the surface for the smile corner coordinates illustrates a large displacement in the y-direction and a minor displacement in the x-direction.

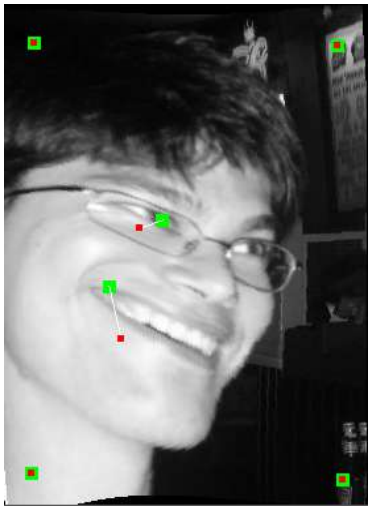


Figure 2.11: Interpolating a surface over the smile coordinates (x, y) to determine the displacement $(\Delta x, \Delta y)$ on the next image [Lombaert 2002].

Spatial Transformer Networks

Convolutional Neural Networks [LeCun and Bengio 1998] are a class of deep neural networks that extract features from the input space using convolution operators. To improve their localisation, Jaderberg *et al.* [2015] introduce a learnable (differentiable) attachable module, the *Spatial Transformer*, that spatially manipulates feature maps of convolutional neural networks without affecting optimisation. In virtual try-on systems, spatial transformer networks (STNs) are used to warp clothing items to fit the body of an individual. In virtual try-on systems, spatial transformer networks (STNs) utilise their input transformation capabilities to warp clothing items so they fit an individual's body. This involves the STNs learning specific transformation parameters, such as scaling, rotation, and warping, required to align the clothing with the person's pose. By accurately adjusting these parameters, the virtual try-on system ensures that the clothing appears to fit the person naturally, providing a realistic and precise representation of how the clothing would look in real life.

In the field of image classification, Spatial Transformer Networks (STNs) are employed to transform and scale-normalise regions of an image to enhance classification accuracy [Lin and Lucey 2017]. Additionally, STNs are used for co-localisation, a process that involves identifying and localising objects within different images that belong to the same class or share a similar semantic representation, even when these images contain varying information. Furthermore, STNs are applied in spatial attention mechanisms to increase computational efficiency by performing attention tasks on low-resolution inputs.

Figure 2.12 illustrates a spatial transformer network that can be categorised in the following manner:

- **Localisation Network**: takes the input feature map (for example, the convolutional output from an input image) and outputs the parameters of the desired spatial transformation. This network can be a simple feed-forward network that learns to predict transformation parameters, such as scale, rotation, and translation, that align the input image with a target pose or configuration.

- **Grid Generator**: once the transformation parameters are determined, the grid generator creates a coordinate grid for the input image. For each pixel in the output feature map, the grid generator defines where in the input image it should sample from.
- **Sampler**: the sampler uses the grid to perform a backward mapping, interpolating pixel values from the input image to produce the transformed output. This interpolation can be done through various methods, such as bilinear interpolation, which ensures smoothness in the output image.

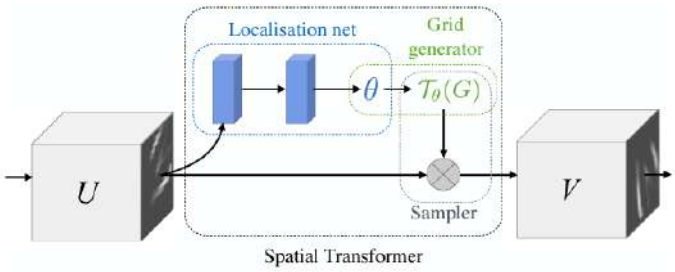


Figure 2.12: The localisation network regresses transformation parameters θ from the input feature map U to produce the warped output feature map V using the regular spatial grid G (transformed to sampling grid $\tau_\theta(G)$).

The size of θ varies on the transformation, e.g θ is 6-dimensional for an affine transformation (allowing rotation, scaling, translation and shearing).

$$\begin{pmatrix} x_i^s \\ y_i^s \\ z \end{pmatrix} = \mathcal{T}_\theta(G_i) = \mathrm{A}_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{2.5}$$

where$(\frac{x_i^t}{z}, \frac{y_i^t}{z})$ are the homogenous target coordinates of the regular grid in the output feature space, (x_i^s, y_i^s) are the source coordinates in the input feature map that define the sample points, and A_θ is the affine transformation matrix.

Figure 2.13 illustrates separate instances of warping showing how an input image, that passes through the convolutional layers uses the localisation network $f_{loc}(U)$ to learn the thin-plate spline parameters θ. Then, a sampling kernel computes the output pixels for each location (x_i, y_i) to produce output feature map $V \in \mathbb{R}^{H' \times W' \times C}$ which is comprised of pixels that lie on a grid $G = \{G_i\}$ such that each pixel has a location $G_i = (x_i^t, y_i^t)$.

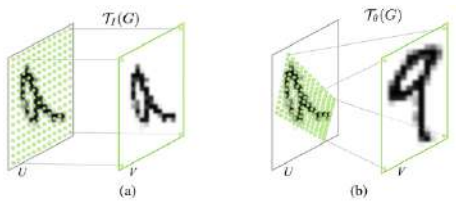


Figure 2.13: Separate instances of parameterised sampling grid applied to image U to produce output V. (a) The sampling grid is the regular grid $G = \tau_I(G)$,where I is the identity transformation parameters.(b) The sampling grid is the result of warping the regular grid with an affine transformation $T_\theta(G)$

As such, the transformation applied on the feature map can take any parameterised differentiable form (plane projective transformation, thin plate spline) that allows gradients to be propagated back from the sampled points $\tau_\theta(G_i)$ to the localisation network output θ.

Flow Estimation is a fundamental topic in computer vision and fluid dynamics, focusing on the prediction or measurement of movement within a given scene or fluid environment. This topic branches into numerous applications, including optical flow in image sequences for computer vision and flow dynamics in fluid mechanics. In the context of virtual try-on applications, appearance flows, exemplified in Figure 2.14, are particularly relevant. They refer to the mapping or transformation of clothing items onto the target body in images or videos, enabling realistic fitting simulations [Ranjan and Black 2017; Ilg *et al.* 2017; Hui *et al.* 2018].

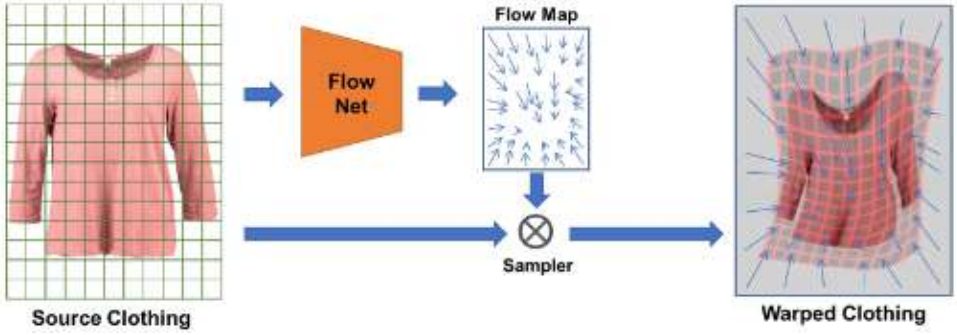


Figure 2.14: This figure [Song *et al.* 2023] illustrates how flow estimation warps clothing regions according to an estimated flow that is attained by predicting 2D coordinate vectors indicating which pixels in the source image could be used to reconstruct the target view [Ilg *et al.* 2017].

Appearance flows in computer vision relate to mapping or transforming textures, patterns, or entire images from one domain or shape to another, which is crucial in applications like virtual try-on. Mathematically, this process involves calculating a vector field, denoted as the appearance flow F, which describes how each pixel in the source image I_s should move to fit onto the target domain I_p. Specifically, the appearance flow is represented as 2-D coordinate vectors specifying which pixels in the input view could be used to reconstruct the target view. The function F maps every pixel position (x, y) in I_s to a new position (x', y') in I_p, formally expressed as:

$$F : (x, y) \mapsto (x', y') \tag{2.6}$$

This mapping ensures the source's texture and appearance are realistically overlaid onto the target, accommodating for perspective, deformation, and shape changes.

In computer vision terms, appearance flows are integral for tasks involving image warping, morphing, or transformation. The process can be outlined in a few key steps. Initially, feature matching is performed to identify corresponding key points in both the source and target images, crucial for applications like virtual try-on where points on clothing items must align with specific body parts. Subsequently, flow estimation calculates the appearance flow, determining the necessary pixel movements to align the source with the target. This can be achieved through optical flow algorithms or learning-based approaches that directly predict the flow field.

Following flow estimation, image warping applies the calculated pixel movement to adjust the source image, ensuring the clothing item realistically conforms to the body's contours and movements. Finally, blending techniques might be applied to ensure seamless integration of the warped source image with the target, adjusting for differences in lighting, shading, and texture. Through these steps, appearance flows enable the realistic transfer of visual information between domains, supporting applications from virtual try-on to face swapping by mathematically modeling and computationally processing pixel movements for coherent and visually plausible transformations.

Composition, depicted in Figure 2.15, fuses warped clothing items on individuals using alpha blending as an initial step in an image compositing technique. This foundational process is crucial for the preliminary integration of clothing with the person's figure. Following this, existing approaches harness the capabilities of a U-Net architecture [Ronneberger *et al.* 2015] to refine and enhance this basic composition, aiming to synthesise high-resolution images that more convincingly blend clothing with individuals. However, these architectures, despite their advanced potential for refining the alpha-blended composition, often undergo unstable training which leads to variations in synthesis quality [Wang *et al.* 2018b].

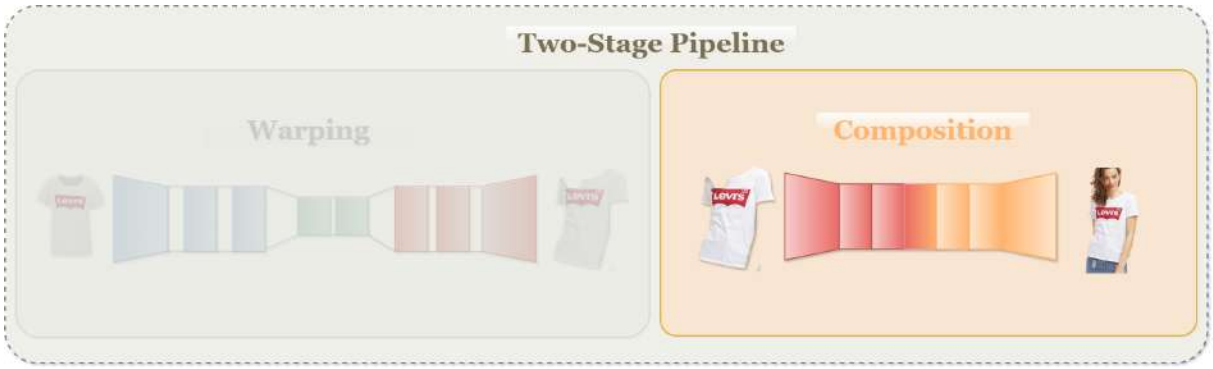


Figure 2.15: Composition represents the final stage in virtual try-on, where the warped clothing is composited onto the individual using a generative network.

To synthesise clothing onto an individual, a virtual try-on network requires a person representation to guide the composition process to identify regions of interest.

The try-on module has a naive solution that attempts to augment the warped clothing directly onto the human subject. This leads to an unnatural appearance surrounding the boundary regions of clothes such that body parts are occluded. Thus, Wang *et al.* [2018a] concatenates the person representation with the warped clothing as an input to the UNet image generator that simultaneously renders a person with their composition mask. The image generator attempts to simultaneously composit the clothing item with the segmentation map and the warped clothing onto the individual. For this reason, Han *et al.* [2018] learn how to composite warped clothing using a refinement network.

Generative modeling represents a frontier in artificial intelligence that focuses on creating models capable of generating new data instances that resemble a given dataset. These models have become pivotal in numerous applications, from image and video synthesis to natural language processing, and even to the creation of music and art. The essence of generative modeling lies in its ability to learn the underlying distribution of a dataset, enabling the generation of new data points that, while not identical, share

the intrinsic characteristics of the original data. Generative adversarial networks are a deep learning technique that learns to mimic data distributions by training competing networks [Goodfellow *et al.* 2014]. These networks consist of a generator network that captures a data distribution and a discriminator network that estimates whether the generated sample belongs to the training data. GANs are often used to generate images and videos in applications such as removing artefacts; super-resolution; pose transfer and image translation. In virtual try-on, GANs are used to generate high-quality images of individuals wearing target clothing.

Figure 2.16 illustrates the architecture mainly consisting of opposing deep neural networks that engage in a zero-sum game such that the generator minimises the objective against the adversarial discriminator to produce data similar to the training data. Adversarial training allows the generator to learn a data distribution similar to the training data. Thus, interpolating between the latent space maps random noise to synthetic samples that mimic the training data.

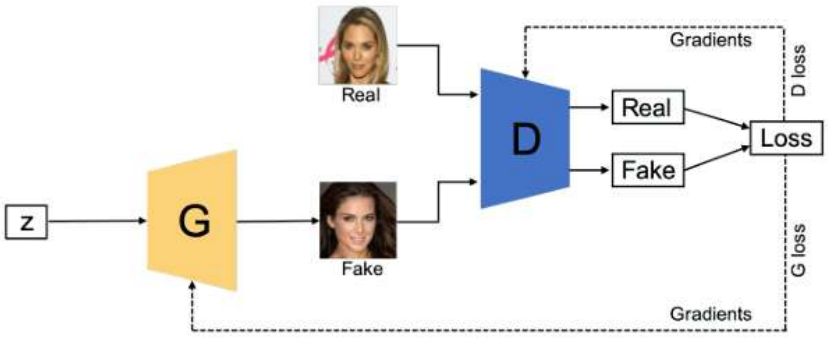


Figure 2.16: In the diagram, there are two main components: a generator (G) and a discriminator (D). The generator creates fake images from a random noise vector (Z), while the discriminator evaluates both real and fake images. The discriminator's goal is to distinguish between real images and the fakes produced by the generator. This is indicated by the two pathways leading to the discriminator: one from the real image and one from the fake image generated by G [Jabbar *et al.* 2021].

The GAN objective function is essentially a minimax game where the generator aims to generate data that is indistinguishable from real data, while the discriminator strives to correctly differentiate between real and generated data. This results in a dynamic equilibrium where both networks improve iteratively.

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \tag{2.7}$$

where:

1. G: The generator in a GAN framework, which creates fake data samples from random noise.
2. D: The discriminator in a GAN framework, which tries to distinguish between real data samples and fake ones created by G.
3. $V(D,G)$: The value function for a GAN that the generator tries to minimize and the discriminator tries to maximize.
4. x: Real data samples from the data distribution p_{data}.
5. z: Random noise used as input for G to generate fake data samples.

6. $D(x)$: The discriminator's estimate of the probability that x is a real data sample rather than a fake one.
7. $G(z)$: The data sample generated by G from random noise z.
8. $\mathbb{E}_{x \sim p_{data}(x)}$: The expected value over all real data samples x following the data distribution p_{data}.
9. $\mathbb{E}_{z \sim p_z(z)}$: The expected value over all random noise samples z following the noise distribution p_z.

A conditional Generative Adversarial Network (cGAN) is a specialised variant of the standard Generative Adversarial Network (GAN) framework, where both the generator and discriminator networks are provided with additional conditioning information during training and generation. Unlike traditional GANs that generate data solely from random noise, cGANs use this extra conditioning input, often in the form of labels or specific data, to guide and control the generation process.

This conditioning ensures that the generated data aligns with the provided information, making CGANs invaluable for tasks like image-to-image translation, super-resolution, style transfer, data augmentation, and attribute-based image generation. In essence, cGANs offer a more directed and controlled approach to generative tasks, enhancing their utility across various domains, including computer vision, image processing, and artistic expression.

In a virtual try-on system, the cGAN is conditioned on an image of a person and the image or description of a clothing item. The generator of the cGAN synthesises an image in which the clothing item is realistically superimposed on the person, considering factors like pose, lighting, and fabric drape. Meanwhile, the discriminator assesses the authenticity of the generated image, comparing it against real images of people wearing clothes. The system iteratively refines the generator's output, learning to adjust for the nuances of different body shapes, clothing styles, and the interactions between them.

Image-to-image translation, depicted in Figure 2.17, leverage cGAN to transform an input image from one domain to another while preserving important visual characteristics. In this context, a "domain" refers to a distinct category or style of images.

In the real-world, before-and-after images of a person wearing specific garments are unavailable. In theory, virtual try-on translates images between two unpaired domains to superimpose clothing items onto individuals. However, a practical scenario of virtual try-on trains a neural network to reconstruct the image of a person given separate images of the person and their corresponding clothing item. Given this setup, virtual try-on techniques utilise a paired setting to synthesise warped clothing items on individuals. In this instance, the generative networks required for image synthesis leverage image-to-image translation with conditional adversarial networks [Isola *et al.* 2017].

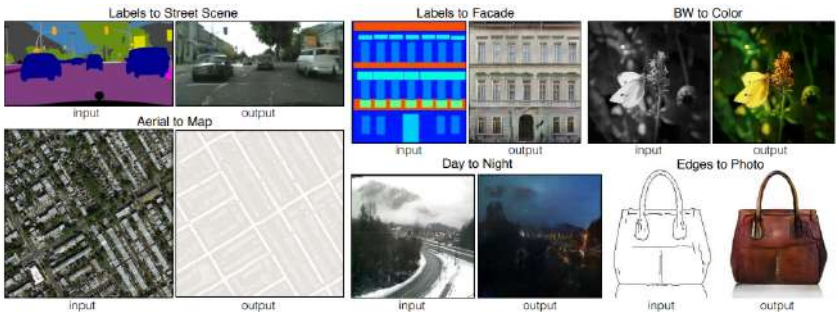


Figure 2.17: Learn to translate an image from one domain into another [Isola *et al.* 2017].

2.1.3 State Of The Art Architectures For Image-Based Virtual Try-On

Our primary contribution is limited to the introduction of a novel dataset for virtual try-on, and as such, we do not propose a new architecture for virtual try-on. However, we provide a comprehensive overview of the state-of-the-art architectures for virtual try-on, which is essential for understanding the context in which our dataset is used.

We train 6 virtual try-on architectures each comprising of warping and composition tasks in their respective pipelines. Their warping approaches are classified into two different techniques: Spatial transformer network that learns the parameters of a thin plate spline transformation and appearance flows, commonly known as FlowNet. Each of these techniques transform the spatial position of pixels to warp the clothing items. Table 2.1 below categorises each method according to their warping method and parser-type.

Method	**Warping**
HR_VITON[1]	FlowNet
CP_VTON[2]	STN
CP_VTON+[3]	STN

Table 2.1: Comparison of different virtual try-on methods showcasing the warping techniques and parser dependencies used by each approach.

High-Resolution Virtual Try-On with Misalignment and Occlusion-Handling

HR-VITON, depicted in Figure 2.18, addresses the challenge of generating high-resolution virtual try-on images that retain fine details and realistic textures [Choi *et al.* 2021]. The core of its technique involves a coarse-to-fine strategy where the initial warping is performed at a lower resolution to ensure the global fit and placement of the clothing item on the avatar. This is followed by a refinement stage, where the details of the clothing texture and the fit are enhanced at a higher resolution. Their framework addresses misalignment and occlusion issues through a unified try-on condition generator. This generator performs warping and segmentation map generation simultaneously, ensuring alignment and natural handling of occlusions.

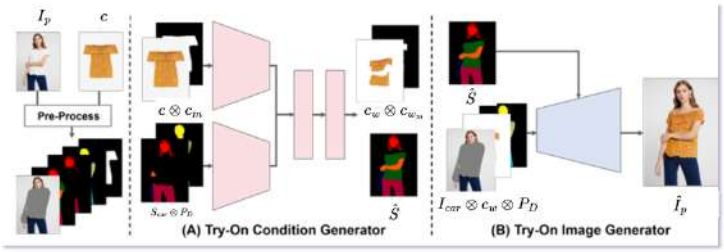


Figure 2.18: Choi *et al.* [2021] introduce a novel high-resolution virtual try-on framework that addresses misalignment and occlusion issues through a unified try-on condition generator.

Traditional methods [Han *et al.* 2018] tackle this by warping the clothing to conform to the person's body shape, followed by generating a segmentation map to distinguish the item from the person's figure. However, these stages often operate in isolation, without sharing critical information, leading to misalignment, where the warped clothing doesn't match the segmentation map, and pixel-squeezing artifacts, especially around areas occluded by body parts. This misalignment significantly degrades the perceptual quality of the final try-on image, an issue exacerbated at higher resolutions.

Figure 2.19: Choi *et al.* [2021] attempt to remove misaligned regions, highlighted by red circles, caused by pixel squezing, depicted by the green circles, artifacts to improve the warping quality of the condition generator, improving from VITON-HD [Choi *et al.* 2021]. Pixel-squeezing artifacts distort, i.e the arm appears to bulge, regions of space surrounding areas of the body parts that occlude the clothing item.

Lee *et al.* [2022] addresses the challenges depicted in Figure 2.19 by introducing a novel try-on condition generator that unifies the warping and segmentation map generation stages to reduce the effect of misaligned regions. Through a feature fusion block, this module facilitates information exchange between the stages, thereby eliminating misalignment and pixel-squeezing artifacts. This integrated approach ensures that the clothing item and segmentation map are perfectly aligned, removing the misalignment entirely and handling occlusions naturally.

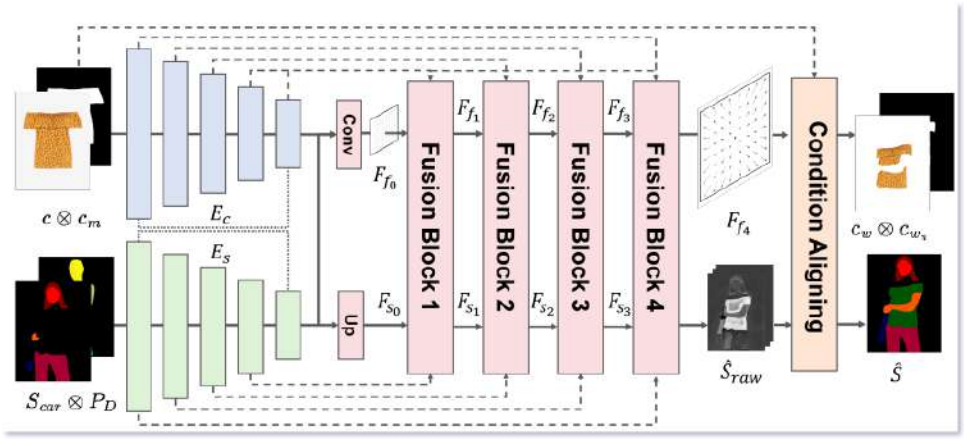


Figure 2.20: The condition generator is comprised of feature fusion blocks that fuse different feature pyramids obtained from the encoders to predict the segmentation map and appearance flow for warping the clothing image [Choi *et al.* 2021].

The condition generator is comprised of feature fusion blocks, that fuse different feature pyramids obtained from the encoders to predict the segmentation map and appearance flow for warping the clothing image. There are two pathways in the feature fusion block, depicted in Figure 2.20, that exchange information to jointly estimate the appearance flow and segmentation map. To correct any misalignment in the segmentation map $\hat{S}_p$, condition aligning is employed. As a result, the condition generator outputs a warped clothing image c_w with its corresponding mask c_{wm}, and separately, $\hat{S_{raw}}$.

These overall objective function for end-to-end training is defined as:

$$\mathcal{L}_{TOCG} = \lambda_{CE}\mathcal{L}_{CE} + \mathcal{L}_{cGAN} + \lambda_{L1}\mathcal{L}_{L1} + \mathcal{L}_{VGG} + \lambda_{TV}\mathcal{L}_{TV} \tag{2.8}$$

where λ_{CE}, λ_{L1}, and λ_{TV} are hyper-parameters controlling the importance between different losses that are defined by:

- **Pixel-Wise Cross-Entropy Loss** (L_{CE}): Cross-entroy between the segmentation maps to learn $\hat{S}_p$:

$$L_{CE} = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log \hat{y}_n + (1-y_n)\log\left(1-\hat{y}_n\right)\right]$$

- **Specialised** $L1$ **Loss** ($\mathcal{L}_{L1}$): is a key component that comprises intermediate warping predictions $W(c_m, F_{f_i})$ extracted from the feature fusion blocks. The relative importance of each term is determined by w_i, formulated as:

$$\mathcal{L}_{L1} = \sum_{i=0}^{3} w_i \cdot \left\| W\left(c_m, F_{f_i}\right) - c_{e_m}\right\|_1 + \left\|c_{w_m} - c_{e_m}\right\|_1$$

- **Least Squares GAN Loss** ($\mathcal{L}_{cGAN}$) [Mao *et al.* 2017]: conditional GAN loss between $\hat{S}$ and S

$$\min_G V_{LS}(G) = \frac{1}{2}\mathbb{E}_{\mathrm{z}\sim p_{\mathrm{z}}(\mathrm{z})}\left[(D(G(\mathrm{z})) - c)^2\right]$$

- **Specialised Perceptual Loss** ($\mathcal{L}_{VGG}$): aims to guide the warping process, with the importance of each term again determined by w_i. This loss is described by the

following equation:

$$\mathcal{L}_{VGG} = \sum_{i=0}^{3} w_i \cdot \phi\left(W\left(c, F_{f_i}\right), I_c\right) + \phi\left(c_w, c\right)$$

- **Total Variation Loss** (L_{TV}): focuses on the last appearance flow in the feature block, learning the flow estimation at smaller representations. It is quantified by:

$$L_{TV} = ||\nabla F_{f_4}||_1$$

Lee *et al.* [2022] generate the try-on image $\hat{I}_p$ through the fusion of a clothing-agnostic image I_{car}, a warped clothing image c_w, and a pose map P_D, all of which are guided by the refined segmentation map $\hat{S}$. The image generator, depicted in Figure 2.18 (B), employs a series of residual blocks and upsampling layers to enhance the output quality. They utilise SPADE [Park *et al.* 2019] normalisation within the residual blocks, modulated by the segmentation map $\hat{S}$. The inputs (I_{car}, c_w, P_D) are resized and concatenated to the activation preceding each residual block.

The objective function for the try-on image generator is constructed using a combination of losses, similar to those applied in SPADE [Park *et al.* 2019] and pix2pixHD [Wang *et al.* 2018b]. It comprises three distinct types of losses: the conditional adversarial loss, the perceptual loss, and the feature matching loss [Wang *et al.* 2018b] defined as follows:

$$\mathcal{L}_{TOIG} = \mathcal{L}_{cGAN} + \lambda_{VGG}\mathcal{L}_{VGG} + \lambda_{FM}\mathcal{L}_{FM} \tag{2.9}$$

where the λ_{VGG} and λ_{FM} hyperparameters control the relative importance between the loss terms, and the feature matching loss is defined as follows:

- **Feature Matching Loss** ($\mathcal{L}_{FM}(G, D_k)$) [Wang *et al.* 2018b]: is based on the discriminator D that stabilises the training of the generator G. Here, D_k represents the ith-layer feature extractor of the discriminator, T is the total number of layers, and N_i denotes the number of elements in each layer. The loss is given by:

$$\mathcal{L}_{FM}(G, D_k) = \mathbb{E}_{(\mathbf{s},\mathbf{x})} \sum_{i=1}^{T} \frac{1}{N_i} \left[\left\| D_k^{(i)}(\mathbf{s}, \mathbf{x}) - D_k^{(i)}(\mathbf{s}, G(\mathbf{s})) \right\|_1 \right]$$

Clothing Shape and Texture Preserving Image-Based Virtual Try-On

Wang *et al.* [2018a], illustrated in Figure 2.21, argue that previous architectures [Long *et al.* 2015; Ronneberger *et al.* 2015; He *et al.* 2016] handle large spatial deformation poorly, leading to misalignment artifacts and compromised textural details in the clothing. To address this, they introduce a novel Geometric Matching Module (GMM) that aligns the person representation with the clothing item for warping. This module is trained from scratch, uses real clothing for ground truths, and directly supervises the warping output with the ground truth through a pixel-wise L1 loss, marking an advancement over previous methods.

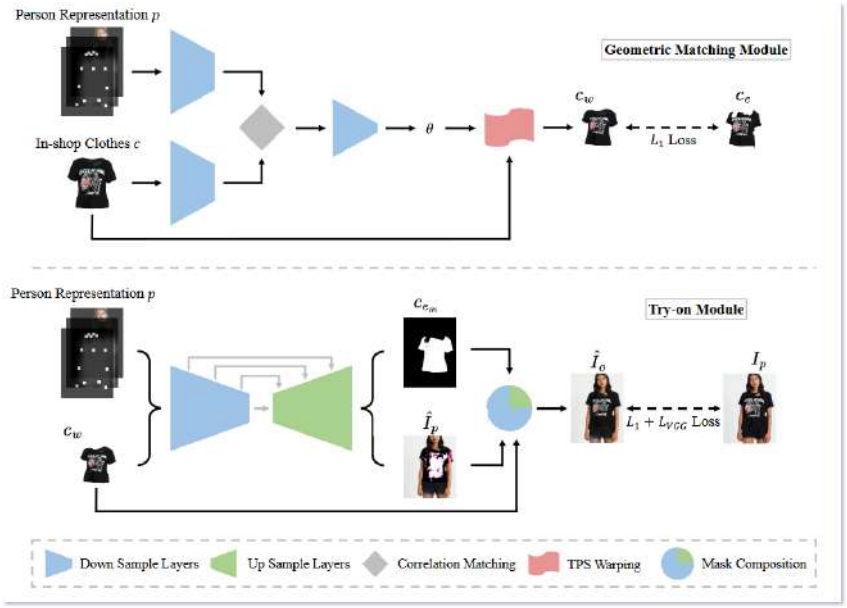


Figure 2.21: Complete architecture that performs warping followed by image composition to perform virtual try-on. The upper part of the pipeline starts with two inputs, the person's body representation and the in-shop clothing image. The geometric matching module yields a set of warping parameters that use TPS to warp the clothing image. The lower part illustrates the try-on phase where the warped clothing is combined with the person's representation to synthesise the clothing onto the individual [Wang *et al.* 2018a].

Wang *et al.* [2018a] introduce a geometric matching module, depicted in Figure 2.22, that preserves clothing characteristics and obtain significant alignment by learning the parameters of a thin plate transformation for warping clothing. The warped clothing correspond to the natural deformations that are prevalent in the reference image.

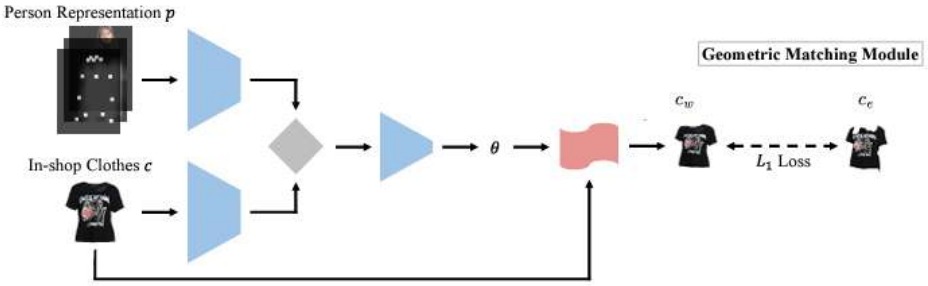


Figure 2.22: CP-VTON introduce the geometric matching module that aligns the person representation p with the clothing item c for warping [Wang *et al.* 2018a].

The authors use a preprocessed person representation that the network requires in tandem with the input pair of the person and their clothing item. This allows the network to generalise the warping process for any clothing to prepare for the composition task. Thus, the person representation (preserving face, hair and body shape) consists of the following:

- **Pose heatmap:** 18 channel feature map encoding the human pose such that each white rectangle (11x11) represents a keypoint.

- **Body shape:** 1 Channel image denoting the blurred binary mask of the person
- **RGB image:** Facial and hair region of the human subject

Furthermore, the paper encodes the correlation between the person pose and the in-shop clothing item using extracted features of a correlation module. This correlation map, representing the encoding of the pairwise similarity between the two feature maps (clothing and person representations), is passed to a regressor that predicts the warped control points for the thin-plate spline stage.

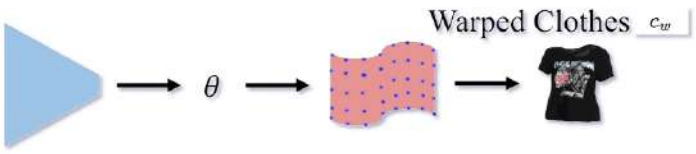


Figure 2.23: Regressor during the thin-plate stage that predicts the warped control points [Wang *et al.* 2018a].

The control points are used to solve for a smooth thin-plate spline transformation that warps the in-shop clothing item to match the target clothing onto the human subject. The loss is minimised using the pixel-wise L1 loss between the warped clothing and the corresponding target clothes.

$$\mathcal{L}_{GMM}(\theta) = L_1(c_w, c_e) = L_1(T_\theta(c), c_e)$$

where T_θ is a TPS transformation module that warps clothing c to c_w such that $c_w = T_\theta(c)$.

Wang *et al.* [2018a] introduce a novel Geometric Matching Module (GMM) utilised in a virtual try-on system for transforming target clothes into a warped version aligned with a person's image. This network is comprised of four main components: two neural networks for feature extraction, a correlation layer, a regressor for spatial transformation parameters, and a Thin-Plate Spline (TPS) module for image warping.

The composition stage (try-on module) trains a U-Net based conditional image generation network that aligns clothing with the target body shape. The try-on image synthesis process is refined by incorporating a composition mask to blend the warped clothing with the target person's image, creating the final try-on result. The system is trained with sample triplets to reduce the difference between the synthesised image and the actual image using a combination of L1 loss and VGG perceptual loss.

The VGG loss measures feature map discrepancies at different convolutional layers of a pre-trained VGG network, emphasising the preservation of perceptual characteristics. To further ensure fidelity in the try-on image, the composition mask is optimised to select the warped clothing accurately, applying an L1 regularisation on the mask itself. This method aims to maintain the distinct features of the clothing while seamlessly integrating it with the person's image. This methods' loss function is defined as follows:

$$\mathcal{L}_{TOM} = \lambda_{L1} L_1(\hat{I}_o, I_p) + \lambda_{VGG} \mathcal{L}_{VGG}(\hat{I}_o, I_p) + \lambda_{mask} L_1(1, M) \tag{2.10}$$

Clothing Shape and Texture Preserving Image-Based Virtual Try-On+

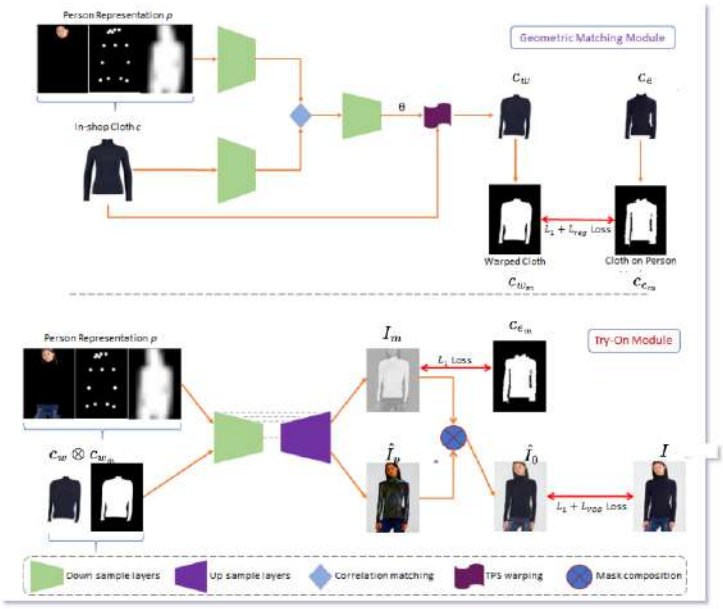


Figure 2.24: The CP-VTON+ enhances the CP-VTON framework. Key upgrades include a refined alignment strategy that reduces the misalignment between the warped clothing and the person's body. During the synthesis phase, the enhanced composition mask and try-on module interact to integrate the clothing onto the person, correcting for any potential artifacts and ensuring a smooth transition between the clothing and the skin [Minar *et al.* 2020].

Minar *et al.* [2020], depicted in Figure 2.24, introduce a novel virtual try-on architecture that builds upon the CP_VTON architecture. This system, referred to as CP-VTON+, added a automatic refinement for segmentation that aligns the person representation with the clothing item for warping.

2.1.4 From Parser-Based To Parser-Free Architectures: Virtual Try-On Without Additional Information

Parser-free networks have garnered attention because they simplify virtual try-on by eliminating the need for a preliminary data preprocessing. They learn to generate the desired output without needing explicit, detailed person representations. To achieve this, they levereage knowledge distilllation techniques to teach student networks to replicate outputs from teacher networks. In this instance, parser-based networks behave as teacher networks that require parser-based information to perform virtual try-on. Then, the student networks learn to reproduce the reconstructed image of the individual wearing target clothing without requiring the parser-based information. This approach offers greater flexibility and speed by circumventing the integration of intermediate representations. Consequently, parser-free systems are particularly advantageous when rapid try-on results are desired, or where preprocessing data may be incomplete or unavailable. Table 2.2 provides different virtual try-on methods, disguishing the warping techniques and parser types used by each approach.

Method	Warping	Parser Type
HR_VITON[1]	FlowNet	Parser-Based
CP_VTON[2]	STN	Parser-Based
CP_VTON+[3]	STN	Parser-Based
PF_AFN[4]	FlowNet	Parser-Free
FS_VTON[5]	FlowNet	Parser-Free
DM_VTON[6]	FlowNet	Parser-Free

Table 2.2: In this table, we incorporate parser-free networks that have been developed to simplify the virtual try-on pipeline by reducing the reliance on precise pose estimation and human parsing segmentation maps. The table also includes parser-based networks that require human parsing information to perform virtual try-on.

Parser-free network requires training a two-stage parser-based virtual try-on pipeline that leverages parser-based information. Then, we feed paser-based outputs as inputs to parser-free networks to perform virtual try-on using only person-clothing image pairs. As a result, each parser-free virtual try-on pipeline trains 4 networks: 2 parser-based networks (warping and composition) and 2 parser-free networks (warping and composition).

Parser-Free Virtual Try-on via Distilling Appearance Flows

Ge *et al.* [2021] introduce a "teacher-tutor-student" scheme, known as a parser-free appearance flow network, that distill an appearance flow (learn to warp clothing items without parser-based information) and synthesise a try-on image. Parser-free virtual try-on represents a shift towards simplifying the pipeline by reducing the reliance on precise pose estimation and human parsing segmentation maps.

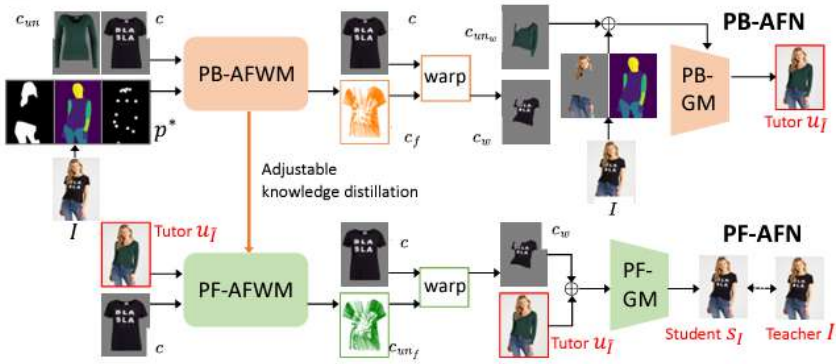


Figure 2.25: The pipeline begins with training parser-based warping networks to produce warped clothing and try-on images respectively. The resultant outputs are used as tutors in parser-free counterparts to produce the try-on result. This alleviates the requirement for parser-based information during inference [Ge *et al.* 2021].

To accomplish this, they proposed a network that is trained using a knowledge distillation technique that leverages the output of a parser-based network to guide the synthesis of garments. More specifically, as depicted in Figure 2.25, the parser-based network generates a fake person image, which is then used as an input for a parser-free "student" model. This model then undergoes supervision by comparing it against

the original, real person image. The goal is for the student model to emulate the real images closely without parser-based information.

The "teacher-tutor-student" scheme utilises a tutor phase to train a two-stage virtual try-on pipeline that produces a warped clothing item and its corresponding try-on image. This tutor phase, considered as parser-based, often produces unsatisfactory segmentation maps affecting the outcome of virtual try-on. To tackle this, parser-free appearance flow networks utilise parser-based information as supervised labels to improve the quality of virtual try-on. Instead, the network leverages tutor knowledge to perform virtual try-on using only the individual and their corresponding clothing item. In this case, the outputs of the parser-based networks represent tutor knowledge that guide the student network to distill appearance flows (warped clothing) and the try-on image. The teacher knowledge refers to the person representation extracted, previously utilised by parser-based networks as supervised labels, during preprocessing to construct the parser-based information. In summary, the knowledge bases and networks can be categorised in the following manner:

1. **Teacher Knowledge:** The training dataset comprising of the following: image of the individual I_p, the clothing item c, the extracted clothing item c_e, person representation extracted during preprocessing $P*$. The teacher knowledge is used to supervise the student network.
2. **Tutor Knowledge:** The output of the parser-based network, which includes the appearance flow (warped clothing) c_f and the try-on image $\hat{I}_p$.
   (a) **Tutor Network:** The parser-based network that produces the warped clothing c_w and the try-on image $\hat{I}_p$.
3. **Student Network:** The parser-free network that learns to distill the appearance flows c_f and synthesise the try-on image $\hat{I}_p$ without parser-based information.

Parser-based generative networks take several inputs, including the warped clothing, human pose estimation, and the preserved region on the human body. It uses these inputs to generate the try-on image. On the other hand, Parser-free generative networks use a different set of inputs, including the extracted clothing item c_e and the input image I_p to distill appearance flows c_f and produce the try-on image $\hat{I}_p$. Both of these generative modules employ a Res-UNet architecture, which is an extension of the UNet architecture with the addition of residual connections. This architecture is chosen for its ability to maintain the intricate details of the warped clothing and generate realistic virtual try-on results.

The optimised loss function $\mathcal{L}$, for both warping and composition, that contains is represented as follows:

$$\mathcal{L} = \lambda_{L1} L_1(s_t, I) + \lambda_{VGG} L_{VGG}(s_t, I) + \lambda_{sec} \mathcal{L}_{sec} \tag{2.11}$$

where s_t is the output of the student network and I represents both c_e and I_p such that:

$$s_t = \begin{cases} c_w, \text{ if we are training a warping network} \\ \hat{I}_p, \text{ if we are training a composition network} \end{cases}$$

$$I = \begin{cases} c_e, \text{ if we are training a warping network} \\ I_p, \text{ if we are training a composition network} \end{cases}$$

and the second-order smooth constraint $\mathcal{L}_{sec}$ is defined as follows:

- **Second-Order Smooth Constraint** ($\mathcal{L}_{sec}$): This constraint is designed to encourage the co-linearity of neighbouring appearance flows, which is crucial for pre-

serving the natural flow and drape of garments on the virtual models. The loss is given by.

$$\mathcal{L}_{sec} = \sum_{i=1}^{N} \sum_{t} \sum_{\pi \in \mathcal{N}_l} \mathcal{P}\left(f_i^{t-\pi} + f_i^{t+\pi} - 2f_i^t\right)$$

Style-Based Global Appearance Flow for Virtual Try-On

He *et al.* [2022], illustrated in Figure 2.26, introduce a StyleGAN [Karras *et al.* 2019] based architecture for global appearance flow estimation that circumvents local appearance flows susceptible to misalignment and occlusion artifacts. The network leverages a global style vector to encode the image-context, thus allowing the appearance flow to consider a larger receptive field. As a result, the network's global context makes warping robust for various scenarios and minimises misalignment for appearance flow estimation. The proposed model's central innovation is the style-based global appearance flow estimation module, which diverges from conventional methods that predict appearance flow based on local feature correspondence.

Flow-style virtual try-on, inspired by parser-free appearance flow networks, leverages a parser-based pipeline as a tutor to distill appearance flows and try-on images for parser-free networks. Additionally, they incorporate feature extractors from the person I_p and an unpaired clothing item c_{un} to encode in a style vector from the lowest resolution feature maps to refine local garment deformation. The warping module learns an appearance flow c_f using the style vector and the feature maps to warp the clothing item. The following describes the notation of the FS_VTON architecture:

1. $\mathcal{F}$: A parser-free pipeline including both warping and composition networks.
   (a) $\mathcal{E}_p$: Feature extraction for the person image in a parser-free setting.
   (b) $\mathcal{E}_g$: Feature extraction for the garment (clothing) in a parser-free setting.
2. $\mathcal{F}^{\mathcal{PB}}$: A parser-based pipeline including both warping and composition networks.
   (a) $\mathcal{E}_p^{PB}$: Feature extraction for the person image in a parser-based setting.
   (b) $\mathcal{E}_g^{PB}$: Feature extraction for the garment (clothing) in a parser-based setting.

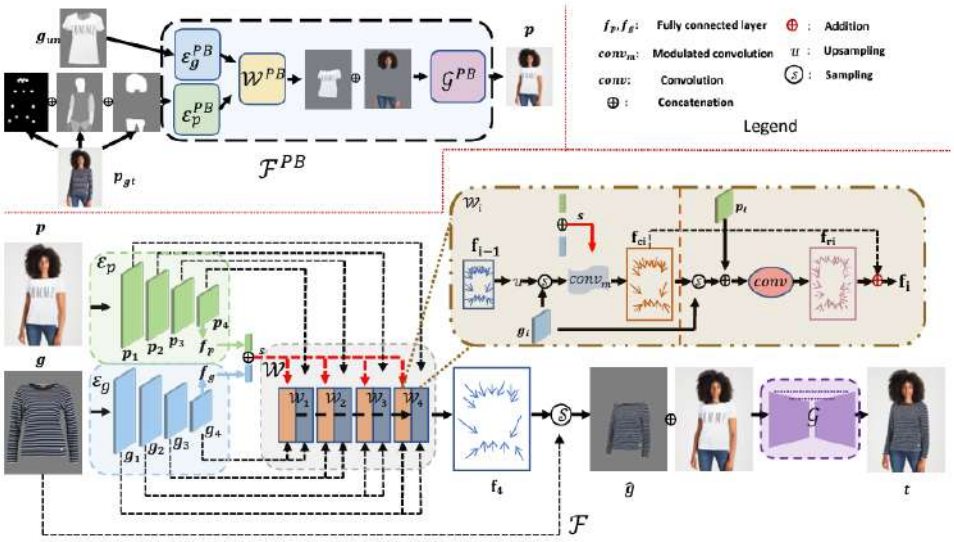


Figure 2.26: Ge *et al.* [2021] overcome the limitation of adopting local appearacen flows to warp clothing items by proposing a global appearance flow estimation model.

For feature extraction, two convolutional encoders are applied to the person and garment images to produce feature maps, which are used by the warping module to predict the appearance flow. The innovation lies in the style-based global appearance flow estimation, which employs a StyleGAN-based architecture to encode the whole-image context. This global context is crucial for handling complex garment deformations. A flow refinement module is introduced to refine the appearance flow at a local level, ensuring accurate alignment of the garment with the person's image.

The warping module utilises a StyleGAN architecture to derive a global style vector that stacks warping blocks composed of a style-based appearance flow and a local correspondence. The global style vector that is required for the warping module is obtained by is extracted from each of the last layers of each of the feature extraction blocks. Style modulation [Karras *et al.* 2019], applied in a StyleGAN architecture, predicts a coarse appearance flow in a style-based appearance flow prediction layer that represents a global style vector that has global receptive field but cannot estimate local deformations. To address this, a local correspondence-based appearance flow refinement layer is introduced to refine the appearance flow at a local level. The refinement layer is designed to capture local deformations and ensure accurate alignment of the garment with the person's image. This process produces outputs $c_{w_{un}}$ and $\hat{I}_{p_{un}}$ where the individual is wearing an unpaired clothing item. Similar to PF_VTON, the overall objective is:

$$\mathcal{L} = \lambda_{VGG} L_{VGG}(t, p_{gt}) + \lambda_{L_g} L_g + \lambda_R \mathcal{L}_R + \lambda_D \mathcal{L}_D \tag{2.12}$$

where each component is defined as follows:

- **Warping Module Loss** (L_g): warping model $\mathcal{W}$ where m_g describes p_{gt} 's clothing mask.

$$L_g = ||\hat{g} - m_g \cdot p_{gt}||$$

- **Smoothness Loss** (L_R): is employed to ensure the smoothness of the generated images, particularly in the appearance flows. It helps maintain the continuity and coherency of the textures and patterns in the generated clothing items on the virtual models.

$$L_R = \sum_i \|\nabla f_i\|$$

- **Distillation or Generalised Charbonnier Loss** (L_D): describes the distillation loss (L_D) is tailored to guide the learning process of the person encoder $\mathcal{E}_p$ within the framework $\mathcal{F}$. This is focused on the output feature map from the i^{th} block in the person encoder $\mathcal{E}_p^{PB}$, and it is represented by.

$$L_D = \sum_i \left\| p_i^{PB} - p_i \right\|$$

Distilled Mobile Real-time Virtual Try-On

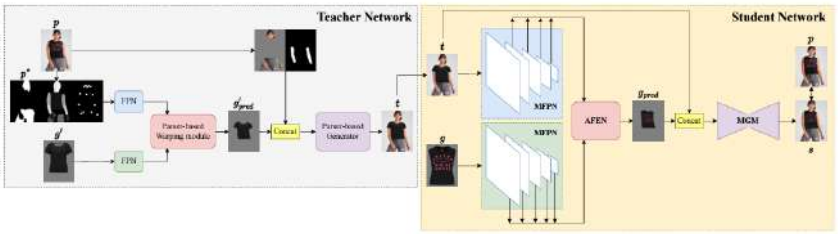


Figure 2.27: Nguyen-Ngoc *et al.* [2023] leverage knowledge distillation to train a lightweight Student network to generate high-quality try-on images without the need for human parsing.

Nguyen-Ngoc *et al.* [2023], depicted in Figure 2.27, introduce the Distilled Mobile Real-time Virtual Try-On (DM-VTON) framework, an approach designed to enhance the efficiency, inference speed and simplicity of virtual try-on applications. DM-VTON employs a knowledge distillation learning strategy, utilising a robust Teacher network to supervise and refine the performance of a Student network. This supervision eliminates the need for human parsing, leveraging their Mobile Feature Pyramid Network (MFPN) and Mobile Generative Module (MGM), both of which are built upon the lightweight MobileNet architecture [Sandler *et al.* 2018]. These components ensure the Student network achieves high throughput and minimal memory usage while maintaining output quality comparable to state-of-the-art (SOTA) methods. To address the issue of limited pose variation in common Virtual Try-On (VTON) datasets, which leads to model overfitting and diminished performance in real-world scenarios, they propose the Virtual Try-on-guided Pose for Data Synthesis (VTP-DS).

The warping module, inspired by FS_VTON, incorporates two feature pyramid networks (FPN), which are instrumental in extracting features from both the human representation and the garment image. For the purpose of transforming the garment's shape, the Teacher network employs a style-based network capable of estimating global appearance flow through modulated convolution. This innovative process begins with the network predicting an initial, coarse appearance flow using a global style vector, which is then finely adjusted at a local level to ensure precise alignment of the garment with the target person's shape and posture. This dual-step refinement process enhances the Teacher network's ability to maintain detail and address significant alignment challenges effectively. The loss function is demonstrated as follows:

$$\mathcal{L} = \lambda_{L_1} L_1 + \lambda_{VGG} L_{VGG} + \lambda_{\text{sec}}\, L_{\text{sec}} + \lambda_{\text{dis}}\, L_{dis}, \tag{2.13}$$

With the generative module, we also apply L1 loss perceptual loss between the synthesised image and the ground truth image to supervise the training process of MGM:

The generator within the Teacher network is designed around an encoder-decoder architecture complemented by skip connections. The Mobile Generative Module is used to facilitate the synthesis of the complete try-on image, combining the warped garment and the image of the target individual. This module represents a blends UNet and MobileNetV2 to guide the creation of the try-on image and significantly lessen the computational load and minimise the size of the model. The loss is defined as follows:

$$\mathcal{L} = \lambda_{L_1} L_1 + \lambda_{VGG} L_{VGG} \tag{2.14}$$

2.2 Constructing Parsed-Based Information For Virtual Try-On

Virtual try-on systems rely heavily on pre-processed data such as pose estimation, human parsing, and clothing-agnostic representations, to accurately synthesise clothing onto individuals. They involve an extensive preparatory phase where the body pose is estimated, the human body is parsed into distinct regions, and a clothing-agnostic representation is generated. This preprocessing allows the network to understand the spatial context of the clothing and the body, ensuring that the garment conforms to the contours and posture of the individual. These systems excel in precision but often at the cost of computational intensity and the requirement of high-quality, detailed input data [Song *et al.* 2023]. Recall that an ideal virtual try-on network transforms a person-clothing input pair into a reconstructed image of the individual wearing the target clothing. In practice, virtual try-on needs to understand where to synthesise clothing information and where to leave the original image untouched. To achieve this, researchers [Han *et al.* 2018; Wang *et al.* 2018a] extract features from the person in the input image and concatenate the person representations together to teach generative networks to warp and synthesise clothing in desired regions of interest.

Firstly, the dataset collection procedure paired images consisting of a individual wearing a garment (Figure 2.28a) and a separate image of the garment itself (Figure 2.28c). This process is arduous since the dataset must contain diverse clothing items, skin tones, body shapes, and poses to ensure generative networks are able to generalise to a broad range of scenarios. It also requires that various ethical, privacy and consent issues are considered in such data collection.

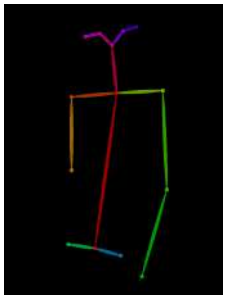
(a) An image of an individual I_p wearing a white shirt.

(b) A clothing item c with its binary mask c_m.

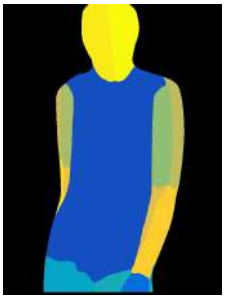
(c) The clothing item c_e extracted from the individual.

Figure 2.28: The first step to constructing a virtual try-on dataset is collecting diverse corresponding image pairs.

Once we have collected the image pairs of individuals and garments, we need to preprocess the images such that the network can understand the spatial context of the clothing item and the body. Neural networks should understand the underlying representations of the human body and their corresponding clothing, including their interaction in transferring garments. To extract the underlying representation, the human body is represented as a set of distinct regions, such as the head, torso, and legs, and the clothing is representated as a set of control points subject to deformation to match the individual's pose. For both the human and the clothing item, we extract binary masks to recognise regions of interest and remove the background. This prevents warping networks from transforming the entire image during the warping phase.



(a) This figure visualises the 2D pose estimation P_p representation that extracted keypoints from the person's image.

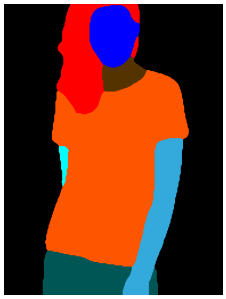


(b) This figure demonstrates the 3D pose estimation P_D mapping human pixels of an image to a 3D surface-based human model.

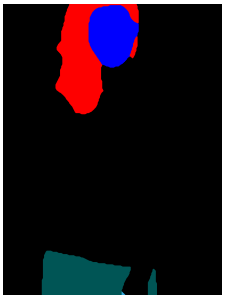
Figure 2.29: This figure visualises the illustrative differences between 2D and 3D pose estimation with respect to underlying representation. Leveraging pose estimation aids the person representation to locate human body parts and localise regions of interest for warping and image composition.

The underlying representations of human body parts guides the warping process. As such, researchers [Han *et al.* 2018; Choi *et al.* 2021] leverage off-the-shelf networks such as OpenPose [Cao *et al.* 2021] to estimate the individual's pose. OpenPose is a sophisticated computer vision algorithm that operates in real-time to identify and track multiple human keypoints, which include joints and features across the body, hands, face, and feet in single images. The system generates a map of keypoints that denote important human anatomical landmarks, allowing for a detailed understanding of a person's pose. They provide a framework for adapting a clothing item's shape to fit and move naturally with the person's body in the image.

2D pose estimation techniques, while effective, have limitations in virtual try-on applications due to their inability to fully capture the complexity of human poses that are 3-dimensional by nature. Such methods might not always accurately predict the pose, which can lead to errors when trying to fit clothing onto a model's image. Given that photos of individuals feature a variety of poses and orientations, warping clothes accurately over these images often requires a sense of depth and form that surpasses what 2D information provides. To surmount these challenges, researchers turn to 3D pose estimation, one of which is known as DensePose [Güler *et al.* 2018]. DensePose maps the pixels of a person's image to a 3D model of the human body. This process yields 3D coordinates for each body pixel, offering a more detailed understanding of the pose. Such depth-aware coordinates are critical for a virtual try-on, as they allow for the clothing item to be manipulated in a way that more authentically mimics real-life draping and contours. Figure 2.29 serves as a visual comparison between the 2D and 3D pose estimation techniques and their respective influence on the clothing warping process. To seamlessly synthesise clothing onto an individual, we remove the clothing from an image of a person, producing a clothing-agnostic representation. By first removing the original clothing item, these representations ensure that the warped clothing blends in the correct area without clashing with the existing clothes. This step is key for the generative network to focus on refining how the target garment fits and integrates with the rest of the image, enhancing the overall realism of the virtual try-on experience. Figure 2.30 illustrates how we leverage instance-level human parsing [Gong *et al.* 2018] to extract segmentation labels corresponding to human body parts. Creating a



(a) This figure represents a segmentation map S that labels different body parts.



(b) This figure represents a clothing-agnostic segmentation map S_{car}.

Figure 2.30: Here, we illustrate how segmentation tasks are used to guide the person representation for training virtual try-on networks. To visualise the human parsing segmentation labels, we attribute each label to a RGB colour.

Figure 2.31: Here, we demonstrate the output of a clothing-agnostic representations that we use to guide the network to localise regions of interest for image composition.

clothing-agnostic representation, depicted in Figure 2.31, for parser-based networks in the context of virtual try-on, involves the tasks explained above that abstracts away specific clothing details while retaining essential body shape and posture information. This process enables the application of various garments onto the base model without the interference of the originally worn clothes Using the clothing-agnostic representations, virtual try-on pipelines composite warped clothing masks onto learned segmentation labels to guide the try-on modules to synthesise target clothing items onto an individual.

2.3 Summary

This chapter provides a comprehensive overview of the developments and methodologies in image-based virtual try-on. In Section 2.1, we introduce virtual try-on and its various approaches, key advancements, and the evolution from parser-based to parser-free architectures that define the current state of virtual try-on. The typical pipeline for virtual try-on involves two main stages: warping the target clothing to match the reference person's body pose and synthesising the resultant warped clothing for garment

transfer. We highlight the complexity of achieving realism in virtual try-on, alongside the significant role of deep learning models, particularly conditional Generative Adversarial Networks (cGANs), in addressing these challenges.

The section discusses various warping and composition (try-on) techniques employed to perform virtual try-on. We delve into detail regarding TPS and elaborate on the mathematical underpinnings of TPS transformations, their integration into Spatial Transformer Networks (STNs), and the role of flow estimation in achieving appearance flows to generate warped clothing items.

Several key parser-based architectures in virtual try-on are reviewed, followed by the transition from parser-based to parser-free architectures. We elaborate on how knowledge distillation techniques enable clothig synthesis using only person-clothing pairs without the need for parser-based supervision.

Finally, in Section 2.2, we lay the groundwork for understanding how the virtual try-on dataset is collected and preprocessed for training. The various steps that we explain help understand how each image plays a significant role in synthesising a final result. Furthermore, this section sets the stage for the next chapter, introducing how we constructed our novel dataset to enhance state-of-the-art virtual try-on by supplementing virtual datasets with varying and diverse body poses.

This background lays the groundwork for understanding the technical concepts and evolving trends in virtual try-on, paving the way for future innovations in the field.

In Chapter 3, we take a closer look at the datasets pivotal to advancing virtual try-on, examining their contributions, shortcomings, and significance. This exploration aims to give a rounded view of the landscape, highlighting the critical role of data in navigating the challenges inherent to virtual try-on solutions.

Then, we unveil our dataset, setting it apart in the context of our findings. We detail its novelty and how it addresses gaps left by prior research, moving the needle forward in virtual try-on technology. This discussion is designed to underscore our unique contribution, stressing the role of diverse perspectives in data collection and processing to enhance the applicability of virtual try-ons.

Chapter 3

Related Work

The development of virtual try-on systems has significantly advanced with the creation of various datasets aimed at improving the accuracy and realism of virtual fitting experiences. Early efforts primarily focused on single-pose datasets [Han *et al.* 2018; Choi *et al.* 2021], which captured models in standard, frontal-facing poses. These datasets, illustrated in Table 3.1, facilitated the initial development of virtual try-on applications by simplifying the alignment of clothing items with human models. However, the lack of diversity in poses limited their applicability in real-world scenarios where clothing fit and appearance change with movement.

Recent advancements have introduced multi-pose virtual try-on datasets [Dong *et al.* 2019; Liu *et al.* 2016b], incorporating images of models in various poses to enhance the realism and versatility of virtual try-on. These datasets demand more sophisticated algorithms capable of accurately simulating clothing behavior across different body positions. While multi-pose datasets represent a significant step forward, they still primarily rely on 2D images, which restricts the ability to fully capture the three-dimensional nature of clothing and body interaction.

Single Pose Virtual Try-On Datasets

Single-pose datasets feature images of models in a fixed, standard pose, simplifying the try-on process by reducing the complexity of aligning clothing items with the model. These datasets are useful for basic virtual fitting rooms in online shopping platforms and early stages of fashion design, offering simplicity in data collection and annotation

Table 3.1: Comparison of different datasets used in virtual try-on and fashion research.

Dataset	Women	Men	Upper body	Lower body	Resolution	Train/Test
Single-pose						
VITON	Y	N	N	N	256 x 192	14,221/2,032
VITON-HD	Y	Y	Y	N	1024 x 768	11,647/2,032
Dress Code	Y	Y	Y	Y	1024 x 768	53000/-
Multi-pose						
DeepFashion	Y	Y	Y	Y	1101 x 750	52,712/-
MPV	Y	Y	Y	N	256 x 192	52,236/10,544
Posecam(Ours)	Y	Y	Y	N	1024 x 768	12,000/4,000

Figure 3.1: Dresscode [Morelli *et al.* 2022] introduces a variety of garments, including those in the lower-body and full-body categories, all presented in high-resolution images.

with lower computational requirements. However, they provide limited realism and versatility as they do not account for how clothes fit in different body positions.

Han *et al.* [2018] introduced the VITON dataset, consisting of 16,253 pairs of frontal-facing human models and corresponding images of clothing that they are wearing, facilitating the study of clothing transfer without the need for 3D body models. The VITON model adeptly addresses the challenge of fitting a 2D clothing image onto an individual, producing realistic results.

Choi *et al.* [2021] compiled a high-resolution virtual try-on dataset, addressing the need for better image quality compared to the dataset previously provided by Han *et al.* [2018]. They collected 13,679 image pairs, each consisting of a frontal-view image of a woman and an upper body clothing item, scraped from an online shopping mall website. These image pairs are divided into a training set and a test set, with 11,647 pairs allocated for training purposes and 2,032 pairs designated for testing. This dataset is utilised to train virtual try-on in two different settings: a paired setting and an unpaired setting. In the paired setting, the objective is to reconstruct the person's image with the original clothing item, while in the unpaired setting, the goal is to change the clothing item on the person's image to a different item. This distinction allows for a comprehensive assessment of the virtual try-on technology's capability to handle both consistent and varied clothing item applications.

In their study, Morelli *et al.* [2022] introduced "Dress Code", illustrated in Figure 3.1, a dataset that significantly improves upon the limitations found in previous collections used for image-based virtual try-on research. Prior datasets predominantly focused on upper-body garments, such as t-shirts, shirts, and tops, and often neglected full-body or lower-body items. This oversight was primarily due to the lack of variety in publicly available datasets, which either featured a limited number of images or did not include paired clothing items for each individual. These constraints have impeded progress in virtual try-on technologies. Dress Code sets itself apart by offering a wide selection of clothing items, including upper-body, lower-body, and full-body garments, in high-resolution images (1024x768). The dataset comprises over 50,000 image pairs, each showing a garment alongside a catalogue image of the item being worn by an individual.

In addition to providing a more diverse and high-quality dataset, the researchers behind Dress Code have developed a novel approach to enhance the virtual try-on experience. They employ a semantic-aware discriminator that focuses on the fine-grained features

Figure 3.2: Figure illustrates the MPV dataset by Dong *et al.* [2019], developed for the MG-VTON project to address virtual try-on challenges across diverse human poses, showcasing extensive variety in poses and clothing.

of the garments, significantly improving the detail and overall quality of the virtual try-on images.

Multi Pose Virtual Try-On Datasets

In contrast, multi-pose virtual try-on datasets include images of models in a variety of poses, requiring more complex algorithms to simulate how clothing fits and moves with the body. These datasets are used in advanced virtual fitting rooms and comprehensive fashion design simulations, offering higher realism and versatility. The increased complexity in data collection and annotation, along with higher computational requirements, are balanced by the robust models that can handle diverse scenarios.

Prior to Dress Code, the challenge of virtual try-on under arbitrary human poses was tackled by Dong *et al.* [2019] in their work on the Multi-pose Guided Virtual Try-On Network (MG-VTON). The authors introduced a new dataset, named MPV, illustrated in Figure 3.2 ,aimed at enhancing the multi-pose guided virtual try-on task. This dataset is notable for its comprehensive coverage of person images across a wider array of poses and clothing diversity. Sourced from the internet, MPV encompasses 35,687 person images alongside 13,524 clothing images. A distinctive feature of each person image within MPV is its capture of different poses, all presented at a resolution of 256x192. From this collection, the authors have compiled 62,780 three-tuples, each consisting of the same individual in the same outfit but depicted in varied poses. These tuples are then allocated into a training set and a testing set, containing 52,236 and 10,544 three-tuples respectively. Notably, the testing set is shuffled to include different clothing and poses, aiming to rigorously evaluate the quality of the dataset.

In contrast, the DeepFashion dataset lacks clothing images, featuring only pairs that showcase the same person in different poses. To assess the generalisation capability of their proposed model, the authors extracted 10,000 pairs from DeepFashion [Liu *et al.* 2016b] and supplemented these with randomly selected clothing images from the MPV's test set for a comprehensive evaluation.

This method was one of the first to facilitate the transfer of clothes on a person's image across a variety of poses. By synthesising a desired human parsing map that accommodates both the target clothes and pose, and employing a deep Warping GAN to align the clothes with the synthesised pose, MG-VTON could generate person images in new clothes and poses with improved texture detail and reduced artifacts.

Despite its advancements, MG-VTON faced limitations in preserving identity and clothing texture details across diverse poses, a gap later addressed by Dress Code's dataset and methodology.

Zhu *et al.* [2020] introduced Deep Fashion3D, depicted in Figure 3.3, the largest dataset of 3D garment models at the time, aimed at advancing the reconstruction of high-fidelity clothing from single images. This dataset comprises 2078 models spanning 10 categories and provides extensive annotations, including 3D feature lines and body poses, accompanied by corresponding multi-view images. Deep Fashion3D facilitates research in garment reconstruction by offering a diverse array of real clothing deformations and a novel baseline method that combines mesh and implicit representations for improved adaptability across clothing types.

Figure 3.3: Deep Fashion3D introduced a collection of 3D clothing models, comprising of 10 distinct categories of clothing, reconstructed from actual garments.

Although Deep Fashion3D significantly enhances the quality and realism of 3D garment reconstruction, its focus on single-view reconstruction highlighted the necessity for methodologies capable of handling multi-pose scenarios, a challenge partially addressed by subsequent works like MG-VTON and Dress Code.

The primary distinctions between single-pose and multi-pose datasets lie in their realism, versatility, data complexity, and computational needs. Single-pose datasets are simpler and easier to manage but less realistic, making them suitable for straightforward applications. Multi-pose datasets, though more complex and resource-intensive, provide a more realistic and versatile simulation of clothing fit and movement, ideal for advanced applications. These differences guide the choice of dataset based on the desired application and level of realism required.

Summary

A significant challenge identified in prior datasets is their inability to accurately perform virtual try-on when the pose deviates from a direct, front-facing camera angle. This limitation severely restricts the applicability of virtual try-on technologies, confining them to a narrow range of poses and, consequently, diminishing their utility in practical, everyday online retail and fashion design.

Our work aims to address these limitations by creating a new dataset that incorporates multiple angles, extending the scope of virtual try-on beyond frontal-facing images. This comprehensive dataset enables the reconstruction of 3D scenes, offering a more immersive and realistic virtual try-on experience. By capturing models from various perspectives, our dataset not only improves the accuracy of clothing fit simulations but also supports real-world applications where a three-dimensional understanding of the scene is crucial.

Chapter 4 unveils the journey of creating a dataset tailored for enhancing virtual try-on. With a focus on multi-view perspectives, this dataset is designed to mimic real-life try-on scenarios closely, expanding the utility of image-based 2D virtual try-on. We explain the significance of introducing a multi-view dataset and highlight the value of diversity in angles and perspectives.

Diving deeper, we outline the intricate process of dataset creation—from selecting varied and representative video sources to the sophisticated extraction of parser-based information. This section delves into the technical challenges of capturing a wide array of poses and garments. By detailing the image processing and data augmentation techniques employed, we aim to provide a comprehensive view of the effort and ingenuity that went into constructing a novel virtual try-on dataset.

Chapter 4

Constructing The PoseCam Dataset

In this chapter, we delve into the steps, exemplified in Figure 4.1, required to build a comprehensive virtual try-on dataset tailored for adapting state-of-the-art techniques to diverse poses. Furthermore, the dataset leverages its unique multi-view perspective to cater for the integration of neural radiance fields for 3D scene reconstruction. Our dataset aids virtual try-on adapt towards real-world scenarios. Moreover, the dataset additionally contains the following information:

1. **Multi-Resolution Images:** The dataset contains parser-based information to cater for both low-resolution and high-resolution virtual try-on tasks.
2. **Multi-view Stereo Imagery:** Given that the video frames are panoramic in nature, the dataset contains a wide array of views to cater for the integration of 3D scene reconstruction. In this regard, the dataset contains a sparse set of images (relatively limited images captured from varying viewpoints around the object) for each individual with corresponding 3D poses.

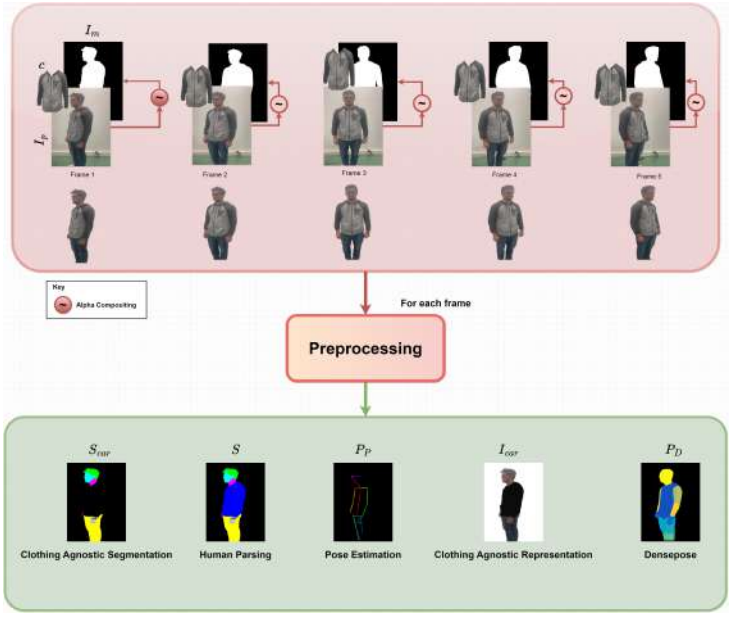


Figure 4.1: Overview of the pipeline essential for generating the **PoseCam** dataset. Firstly, we collect raw image frames and use alpha blending, with the frame's mask, to remove the background. Next, we extract parser-based information from each frame using off-the-shelf networks. Finally, we use the extracted information to build a comprehensive virtual try-on dataset.

Given our reliance on human subjects for our experiments, it is incumbent upon us to declare that we secured ethical clearance from the Ethics Committee at the University of the Witwatersrand. Our duty extends to both the committee and the participants, ensuring all individuals involved were briefed on the experiment's nature. Moreover, each participant was asked to provide consent for recording videos while standing on a rotating platform, and for the use of these recordings in training deep learning algorithms. Specifically, we addressed the possibility of publishing the dataset with their faces displayed, emphasising that the decision to have their faces published was entirely theirs to make. These measures affirm our commitment to ethical research practices, where participants' autonomy and the right to informed consent are paramount.

4.1 Collecting Raw Data

In alignment with our research objectives, we contribute a dataset enabling state-of-the-art virtual try-on networks to adapt to real-world scenarios comprising of diverse and complex poses. To achieve this, we have collected image pairs, comprising clothing items and frames extracted from videos of volunteers standing on a rotating platform to capture 180-degree views. This approach broadens the scope of the virtual try-on dataset to incorporate varying perspectives and diverse poses.

Our dataset encompasses 100 diverse students, varying in height, age, gender, and race, to ensure our model's robustness across different demographics. Specifically, we extracted 200 frames from each video, resulting in a total of 20,000 images. These frames were captured at 29.98 FPS using an iPhone 12 (1080x1920, HEVC) under controlled conditions, featuring a white background, with the camera focusing from the top of the head to the knees. Each student was paired with one clothing item, totaling 100 distinct clothing items in the dataset. To collect paired clothing items, we took individual photos of the clothing each student was wearing by placing the items on a surface. Later in this section, we explain how these images were cleaned and prepared for the task of virtual try-on.

While providing detailed demographic statistics is not standard in this field, as most datasets are obtained through web scraping [Morelli *et al.* 2022; Han *et al.* 2018; Song *et al.* 2023], our work emphasises expanding frontal-facing datasets to incorporate multiple perspectives, enabling virtual try-on to be adaptable regardless of the individual's angle towards the camera. Detailed demographic statistics (e.g., gender, race, height, age) are therefore not the primary focus of our dataset. Nevertheless, our dataset includes diverse participants to better serve as an augmentation to existing datasets, enhancing their utility by providing multiple perspectives and improving model robustness. Future work can investigate the impact of race, height, and gender on virtual try-on applications. By documenting the methods and conditions under which the videos and photos were taken, we aim to facilitate the reproducibility of our results and encourage the adoption of similar approaches in future research.

In Figure 4.2, we employ FFMPEG to extract individual frames from each video, isolating individual subjects. Our objective is to curate a dataset containing a sparse set of varying angles for each individual, necessitating background removal to reduce artifacts during the warping phase of virtual try-on. To this end, we employ Segment Anything [Kirillov *et al.* 2023], an image segmentation model that detects and extract masks from images, to ensure a consistent and noise-free background across each frame. To further

refine our dataset, we use local light field fusion[Mildenhall *et al.* 2019; Müller *et al.* 2022] and Colmap [Schönberger and Frahm 2016] to extract pose information from a set of sparse pictures pertaining to a single individual.

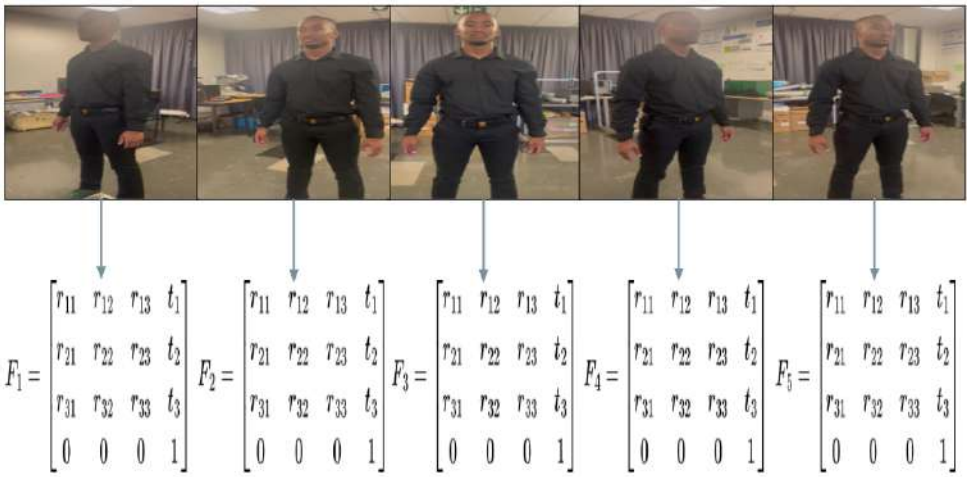


Figure 4.2: This figure represents how we utilised colmap to perform feature matching allowing us to place a camera within the 3D scene. Each frame also contains a pose transformation matrix, extracted using local light field fusion, necessary for training neural radiance fields for each individual. The matrices in this figure are represented as examples of a 4x4 matrix that describe the camera's extrinsic parameters relative to each frame.

The sparse set of image frames, extracted from videos, serves as the foundation for generating parser-based information essential for virtual try-on applications. For every video captured, we acquired an image of the participant's clothing item. Additionally, each frame is accompanied by pose transformation matrices, representing the camera's extrinsic parameters for a particular frame within the set of sparse images. This arrangement allows our dataset to be uniquely tailored for novel view synthesis applications.

For each individual in our study, we pair them with their respective clothing item, creating image pairs that facilitate virtual try-on. This method of pairing is instrumental in constructing our comprehensive dataset, as depicted in Figure 4.3. Clothing masks

Figure 4.3: The "person-clothing" image pairs are essential as practical inputs for the network, facilitating the accurate mapping and realistic simulation of clothing onto various body shapes in the virtual try-on process [Han *et al.* 2018].

play a vital role in preprocessing the input images by removing background details, as depicted in Figure 4.4. This prevents cloth deformation networks from warping background information further impacting the quality of virtual try-on. It ensures that only the clothing item is warped and superimposed on the individual, while the background remains unchanged. This is important for maintaining the realism of the virtual try-on image. Figure 4.5 presents a recurring issue of white spots(artifacts) that

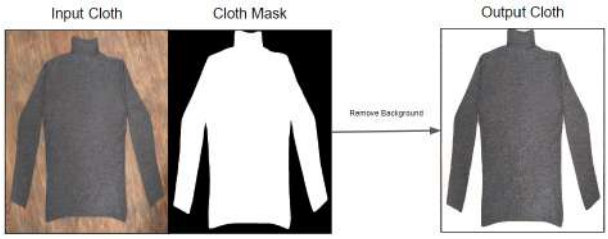


Figure 4.4: Virtual try-on networks warp the clothing item to match the pose of the individual. Here, we show the procedure of creating a mask to prevent the network from warping the background.

were consistent within our dataset during preprocessing. To solve this, we adopted Segment Anything [Kirillov *et al.* 2023] as a more effective alternative for background removal, thereby enhancing the quality and reliability of the clothing masks created for our dataset. This approach prevents the background noise from interfering with the image generation process.

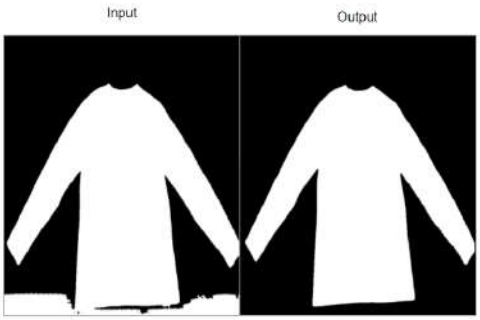


Figure 4.5: Utilising Segment Anything together with morphological processing to remove artifacts from binary masks.

Figure 4.6 summarises the steps to produce the image pairs necessary for virtual try-on.

The acquisition of precise clothing masks is a critical step in our workflow. as it allows for accurate segmentation of clothing items from the background, as shown in Figures 4.6a,4.6b, which is essential for preventing networks from warping extraneous elements.

Lastly, we discussed our employment of Segment Anything for mask creation. This approach has proved more effective in removing background noise and refining mask quality, leading to a substantial reduction in the white spots that were a recurrent issue with the previous method.

By enhancing mask fidelity, we have significantly improved the overall quality of our virtual try-on system, ensuring that the clothing fits seamlessly within the generated images, without the distraction of background artifacts.

(a) Removing the background from each frame to reduce artifacts for virtual try-on.

(b) Removing the background from garments to reduce artifacts for virtual try-on.

Figure 4.6: Here we illustrate the steps necessary to produce the image pairs from raw data. These image pairs are a prerequisite to building a more comprehensive set of features necessary for the virtual try-on.

4.2 Building A Comprehensive Virtual Try-On Dataset

In this section, we delve into the foundational methodologies that underpin the development of a comprehensive dataset designed for parser-based virtual try-on. Our approach begins with instance-level human parsing, where we extract pixel-level labels of clothing items and human body parts. This crucial step enables the detailed categorisation of various human parts, overcoming challenges related to precise body representation and garment fitting. Subsequently, we gather pose estimation data to achieve a detailed representation of the human body's pose. This is essential for enabling virtual try-on networks to accurately warp clothing items onto the individual.

A comprehensive dataset, in this context, signifies one that is suitable for parser-based virtual try-on. It encompasses a person representation incorporating pose estimation, densepose, human parsing, and clothing-agnostic representation. Importantly, our dataset includes both low-resolution (LR) and high-resolution (HR) variants of these elements. Such a diverse dataset ensures versatility and robustness, allowing for the accurate simulation of virtual try-on across different resolutions and enhancing the realism of the virtual fitting experience.

4.2.1 Instance-Level Human Parsing

In Figure 4.7, we associate each body part or clothing with a pixel using a technique called instance-level human parsing. It is a segmentation network that identifies and segregates different human parts and clothing items in an image at a pixel level. It does so by generating a segmentation map, where each pixel of the user's image is classified into different categories corresponding to various body parts and clothing items. With a detailed segmentation map, the system can identify specific areas, such as the torso, by labeling them with distinct identifiers or colours on the map.

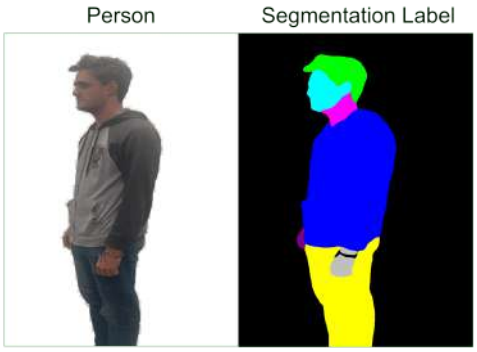


Figure 4.7: This figure illustrates human parsing, an instance-level segmentation task leveraging part-grouping networks that label human body parts [Gong *et al.* 2018]. In this instance, our region of interest is the blue segmentation label corresponding to the shirt.

This precise identification allows the system to accurately overlay clothing items onto the individual's image in alignment with their body pose and clothing labels. Specifically, for the torso, the segmentation map and labels enable the system to adjust the clothing item—be it a shirt, blouse, or jacket—so that it conforms perfectly to the torso's contours, taking into account the individual's unique body shape and size.

4.2.2 Pose Estimation

Human-pose estimation identifies joints of a human body that form a skeleton by detecting and associating body parts with individual keypoints. By identifying and mapping out the joints of a human body to form a skeleton, it provides a foundational understanding of the person's posture and movement. This insight is critical for virtual try-on technologies, as it allows these systems to adapt clothing items to the user's current pose accurately.

A keypoint defines a set of coordinates for each classified joint describing the person's pose outlining the person's pose. The coordinates (x, y) of each keypoint describe its position in the 2D image space, offering spatial information about where a particular body part is located relative to the overall frame. The visibility v indicates whether a keypoint is clearly visible in the image, partially occluded, or entirely hidden from view.

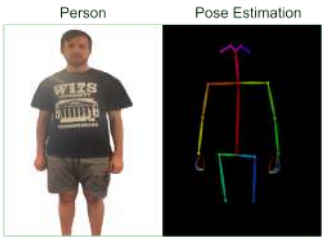
Pose estimation methodologies can be broadly classified into two approaches, each with distinct operational functions:

- **Top-Down Approach**: This approach first detects each person in the image using bounding boxes. Within these detected regions, it then estimates the keypoints of body joints to construct the pose of each individual.
- **Bottom-Up Approach**: In contrast, this method starts by detecting all keypoints of body joints in the image without prior identification of individuals. Subsequently, it groups these keypoints into distinct poses for each detected person based on their spatial proximity and anatomical consistency.

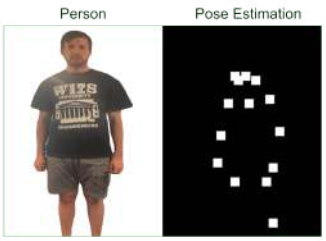
Traditional methods use various hand-crafted feature extraction techniques for individual parts. Recent techniques [Cao *et al.* 2021] utilise multi-stage convolutional neural networks to refine part affinity fields and corresponding confidence maps to perform

multi-person pose estimation. Furthermore, a nonparametric representation defining a set of 2D vector fields, referred to as part affinity fields, encodes the location and orientation associating body parts with individuals.

However, 2D pose-estimation is known to perform poorly given non-typical poses and diverse forms of clothing and occlusion [Dong *et al.* 2019]. The pose is extracted to generate the clothing-agnostic representation that is fed into the virtual try-on generator that warps and fits arbitrary clothing onto the person's region of interest. In Figure 4.8, we generate pose estimation representations for both high-resolution and low-resolution images, enabling us to train virtual try-on networks across multiple resolutions [Cao *et al.* 2021].



(a) High resolution illustration of pose estimation used for virtual try-on networks that accept 1024x768 images as input.



(b) Low resolution illustration of pose estimation crucial for virtual try-on networks that accept 256x192 images as input.

Figure 4.8: We used Cao *et al.* [2021]'s paper to extract pose representation for both high-resolution and low-resolution images. This allows us to train virtual try-on networks across multiple resolutions.

4.2.3 Dense Human Pose Estimation

Figure 4.9 illustrates the process required to annotate dense correspondences between images and a 3D surface model. Firstly, annotators segment the image into semantic regions and pinpoint the corresponding surface point for each sampled point within any of the rendered part images. The red cross serves as an indicator for the currently annotated point. The surface coordinates derived from the rendered views are then used to localise the collected 2D points onto the 3D model.

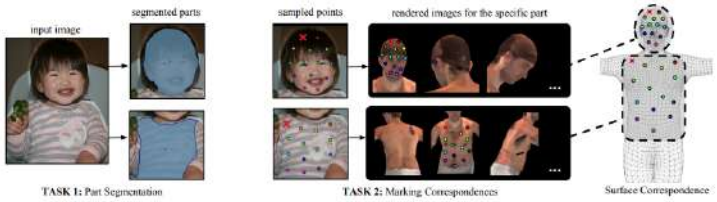


Figure 4.9: Annotators segment the image into semantic regions. Next, they precisely locate the corresponding surface point for each sampled point on any of the rendered part images [Güler *et al.* 2018].

In the case of our dataset, depicted in Figure 4.10, densepose provides a fine-grained, pixel-level representation of human pose and shape suitable for generating clothing-agnostic representations. Densepose provides a dense prediction of body keypoints

and their associated UV coordinates for each pixel in an image. This level of detail allows for precise pose estimation even when individuals are wearing various types of clothing or none at all. It captures the contours and positions of body parts with high accuracy. In specific situations, when we apply a pretrained densepose network to an

Figure 4.10: Here we create a densepose representation from raw image frames to enhance the person representation for virtual try-on networks.

image, it sometimes produces an undesired effect where pieces of clothing appear in the final image as if they are still there, even though they should have been removed. To address this issue, we employ a measurement called "cosine similarity" to assess the similarity between these predictions containing lingering clothing pixels and the densepose representation.

Cosine similarity essentially tells us how alike or different two vectors (in this case, pixel values) are in terms of their direction in a high-dimensional space. When we calculate the cosine similarity between the predictions with lingering clothing pixels and the densepose representation, a higher similarity score suggests that these pixels are more similar to the densepose features and are possibly part of the human body. Conversely, a lower similarity score indicates that these pixels are more different from the densepose representation and are likely closer to the background.

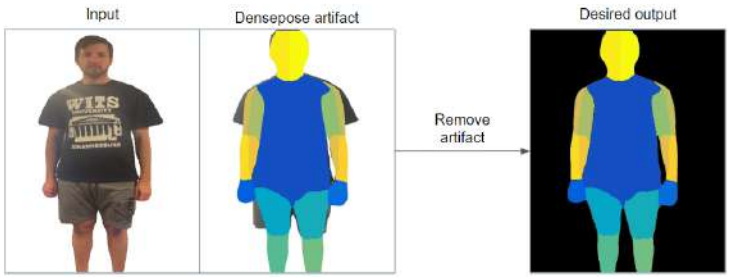


Figure 4.11: Here, we used cosine similarity together with masking to remove pixels corresponding to clothing items from the reference image. In certain instances, the off-the-shelf network that generates the densepose representations with a background. However, it is standard practice for densepose representations to have a black background.

Given the noise presented by densepose in Figure 4.11 , we used cosine similarity to measure the difference between densepose and predicted representations. Pixels that are less similar are considered part of the background and are turned white, and we

subsequently convert the background back to its original appearance by changing these white pixels to black. This process helps eliminate artifacts created by the densepose network in certain scenarios. Thus, the cosine of two non-zero vectors can be derived by using the Euclidean dot product formula:

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\|\|\mathbf{B}\| \cos\theta \tag{4.1}$$

Given two n-dimensional vectors of attributes, $\mathbf{A}$ and $\mathbf{B}$, the cosine similarity, $\cos(\theta)$, is represented using a dot product and magnitude as

$$\text{cosine similarity} = S_C(\mathbf{A}, \mathbf{B}) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, \tag{4.2}$$

where A_i and B_i are the ith components of vectors $\mathbf{A}$ and $\mathbf{B}$, respectively.

4.2.4 Agnostic Representations

Agnostic representations, in the context of generative networks, are images that emphasise the presence of certain segmentation labels or attributes. These representations are designed to enable generative models to manipulate or substitute specific features of the input data while keeping the rest unchanged. In other words, agnostic representations grant generative networks the flexibility to work with various textures, styles, or attributes, regardless of the original content, thereby enhancing their versatility and adaptability.

They allow generative networks to disentangle the clothing's texture or pattern from the underlying garment structure. This means that users can try on clothing items with different textures and patterns without altering the fit or style of the garment itself. This level of texture independence enhances the variety and customisation options available to users during the virtual try-on process. Clothing-agnostic segmentation describes an

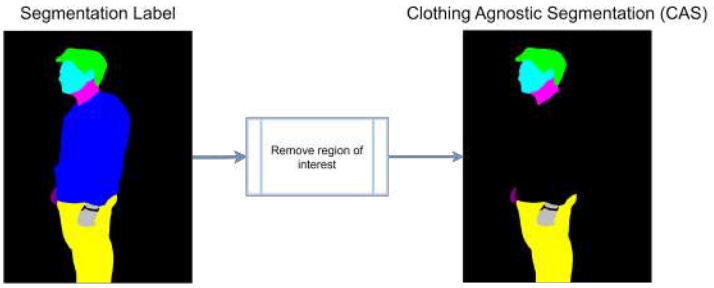


Figure 4.12: We remove clothing items from the human parsing segmentation maps, mentiond in subsection 4.2.1, to generate a clothing-agnostic segmentation labels.

image containing semantic information independent of target clothing items and is a fundamental step in the creation of clothing-agnostic representations. It involves the generation of a segmentation map, where pixel-level annotations distinguish clothing parts from the underlying human body.

As shown in Figure 4.12, these annotations are used to identify and label regions of an image corresponding to clothing, with the specific aim of isolating the human torso from other body parts. In practice, pixels labeled as 'torso' (e.g., label 4) are typically

represented as black (0), while other clothing items may be illustrated in various distinct colours. This segmentation map serves as the foundation for clothing-agnostic representations, allowing us to warp clothing items into empty regions of interest after removing clothing items from the individual.

A clothing-agnostic representation, depicted in Figure 4.13, in the context of virtual try-on, refers to the ability to accurately depict a user's appearance and posture underneath the clothing they are trying on, regardless of its style, texture, or complexity. This representation is pivotal for ensuring that the virtual try-on experience closely mimics the reality of trying on physical clothing items in a store.

We use the segmentation labels to remove pixels pertaining to clothing items of interest. This enables the training of a composition network, facilitating the deformation of a clothing item and the subsequent completion of the void region

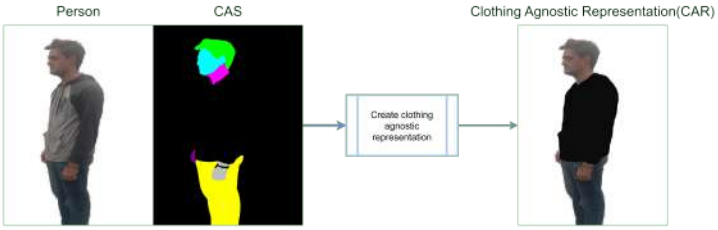


Figure 4.13: We remove clothing items from the human parsing segmentation maps, mentiond in Section subsection 4.2.1, to generate a clothing-agnostic segmentation labels.

Previous techniques [Han *et al.* 2018] provide a coarse body shape mask to synthesise clothing onto an individual. However, these techniques fail to reproduce body parts while preserving skin colour. A clothing-agnostic representation generates a segmentation map that removes clothing regions with their corresponding pose maps that remove arms using pretrained off-the-shelf networks [Gong *et al.* 2018; Cao *et al.* 2021]. Thus, a clothing-agnostic representation eliminates the shape of the clothing item while preserving body parts to prevent the network from synthesising skin onto unrelated regions.

4.2.5 Extracting The Warped Clothing

Figure 4.14 demonstrates how we leverage segmentation maps to identify clothing regions within an image, we can apply alpha compositing to isolate these areas with high precision. The segmentation map's values directly influence the alpha values, allowing for the precise extraction of clothing by making the non-clothing regions transparent. This process not only aids in digital fashion trials but also in various computer vision tasks requiring the isolation of specific image segments.

4.3 Summary

In Section 4.1, we detailed the process of collecting raw data to simulate augmented reality experiences through deep learning techniques. By gathering image pairs from

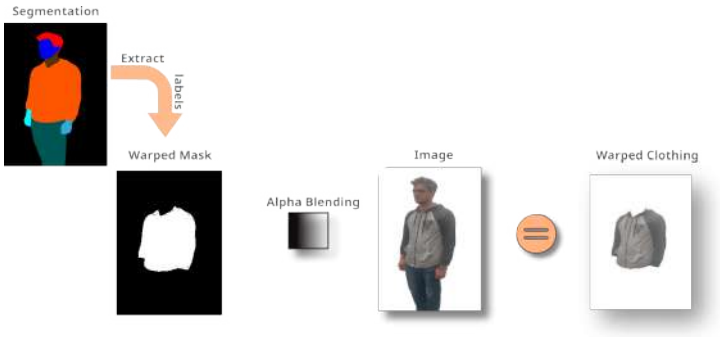


Figure 4.14: To extract the warped clothing, we associate segmentation labels associated with the clothing item and use alpha blending to extract the item from the individual.

diverse volunteers and enhancing the VITON-HD dataset, we have created a comprehensive dataset that captures a wide array of human features and poses from various perspectives. This effort lays the groundwork for developing a robust model capable of accurate virtual garment fitting across different demographics. The adoption of advanced techniques for extracting pose information and generating precise clothing masks marks a pivotal step towards achieving realistic and personalised virtual try-on experiences, significantly improving the system's quality by reducing artifacts and ensuring seamless clothing integration.

Finally, Section 4.2 explored the foundational technologies and methodologies that drive the development of advanced virtual try-on systems, demonstrating our application of sophisticated vision technologies for preprocessing and categorising image pairs. By leveraging insights from existing research to extract pixel-level labels and obtain detailed body pose representations, we have significantly advanced the capability of virtual try-on systems to provide realistic and accurate visualisations of clothing onto digital avatars. The employment of pose estimation techniques for a comprehensive mapping between 2D images and the 3D body surface has been crucial in synthesising lifelike virtual try-ons.

This advancement not only marks a significant leap in digital fashion technology but also sets the stage for a new era where users can explore and experiment with clothing in a virtual environment with unprecedented ease and realism.

In Chapter 5, we discuss the experiment setup that facilitates the evaluation of our dataset's performance after fine-tuning. The intent is to shed light on the process of testing virtual try-on to determine their capacity to replicate a realistic try-on experience in a digital environment. The chapter initiates with a discussion on the division of the dataset into distinct segments: training, validation, and testing.

Next we quantify and qualify the performance of these networks. Quantitative measures—such as accuracy, computational efficiency, and scalability—are discussed in detail. These metrics afford an objective basis for comparison. Simultaneously, the chapter emphasises the importance of qualitative evaluation through visual assessments and user experience metrics. This dual approach recognises the subjective dimension of fashion and apparel, where aesthetic appeal and personal preference play pivotal roles.

Chapter 5

Experimental Setup

In this chapter, we discuss how we utilise our dataset to facilitate the training procedure for virtual try-on. We elaborate on the data splits necessary to ensure that the model generalises on unseen data. Furthermore, we explain the separation between resolutions in the dataset. Then, in Section 5.1, we discuss the training procedure for the virtual try-on networks explaining the different approaches that we use to train the networks and the hyperparameter sweeps that we perform to ensure that the best method is chosen during validation. Finally, in Section 5.2, we discuss the evaluation metrics that we use to compare the performance of each method. We also discuss the importance of hyperparameter sweeps to ensure that the best method is chosen during validation. Finally, we conclude the experimental setup in Section 5.3.

5.1 Training Procedure

Following from Chapter 4 that details how we facilitate our own data collection, we separate the **PoseCam** dataset into high-resolution and low-resolution counterparts where the image sizes are defined as 1024x768 and 256x192 respectively, since virtual try-on architectures are designed to evaluate warping and composition based on their image sizes.

We utilise pretrained state-of-the-art architectures, mentioned in subsection 2.1.3, to train virtual try-on networks on our dataset. Thus, there are 6 virtual try-on architectures, each comprising of warping and composition stages in their respective pipelines.

A single experiment refers to training a single stage, or network, on the dataset. For example, to train a single architecture (HR_VITON), we train separate experiments comprising of warping and composition stages. Hence, for each experiment we categorise the stage as either warping or composition and compare the performance of each architecture based on the evaluation metrics obtained from the testing set.

To follow the best practice of training deep learning algorithms, we split the dataset into training(60%), validation(20%) and testing(20%) sets to ensure that the chosen networks generalise well on unseen data. The dataset splits are kept consistent across the experiments for all architectures to ensure that the performance of each method is evaluated fairly.

Virtual try-on is categorised into two distinct approaches: parser-based and parser-free. Parser-based methods rely on human parsing to segment the clothing and body

regions, while parser-free methods do not require human parsing. Parser-free networks pretrain parser-based networks to distil the knowledge of parsing into the parser-free networks. In other words, parser-free networks train 4 different networks: parser-based warping, parser-based composition, parser-free warping, and parser-free composition. In reporting our experiments, we compare the results across parser-based and parser-free experiments, neglecting the parser-based performances of parser-free networks, to evaluate the performance of each architecture.

To this end, we train 18 different virtual try-on networks, illustrated in Table 5.1, to evaluate the feasibility of our dataset:

Method	**Warping**	**Parser Type**	**No. Networks**
HR_VITON[1]	FlowNet	Parser-Based	2
CP_VTON[2]	STN	Parser-Based	2
CP_VTON+[3]	STN	Parser-Based	2
PF_AFN[4]	FlowNet	Parser-Free	4
FS_VTON[5]	FlowNet	Parser-Free	4
DM_VTON[6]	FlowNet	Parser-Free	4

Table 5.1: In this table, we highlight the number of networks trained for each virtual try-on architecture, distinguishing the warping techniques and parser types used by each approach. Parser-free networks distill knowledge from parser-based networks to perform virtual try-on without requiring human parsing information. As such, they pretrain 2 parser-based networks and their outputs to train 2 parser-free networks. Although we train each network from each architecture, the parser-based results of parser-free networks are not reported because they do not reflect the performance of the architecture.

Evaluating performance during training indicates the model's ability to learn from its training data. However, comparative analysis on performance during validation directly reflects the model's ability to generalise the warping and composition process to unseen data. Hence, our research predominantly performs hyperparameter optimisation for the warping task, on the validation dataset, recognising its critical role in the overall composition process.

For each architecture, we leverage hyperparameters sweeps using weights and biases, mentioned in Appendix A, to ensure that the best performing architecture, during validation, for each stage is chosen.

5.2 Evaluation

5.2.1 Quantitative Evaluation

Objective evaluation methods work directly with pixel values to ascertain their comparisons. A variety of methods are used to compare texture and pixel similarities between images. These allow us to evaluate the quality of computer vision models with respect to popular benchmarks.

Peak Signal To Noise Ratio (PSNR)

The peak signal-to-noise ratio measures the image quality obtained from noise compression and restoration tasks. It measures the ratio between the maximum possible signal value (maximum pixel intensity) and the power of distorted noise (MSE between original and reconstructed image) that affects the quality of its representation. Supposedly, a higher PSNR indicates a better image quality. However, it struggles to perceive textural details and assess the quality of the image that do not correlate well with human judgement.

$$PSNR = 20\log_{10}(\frac{MAX_f}{\sqrt{MSE}})$$

where MSE is given by:

$$MSE = \frac{1}{mn}\sum_{0}^{m-1}\sum_{0}^{n-1}||I(i,j) - \hat{I}(i,j)||^2$$

where I and $\hat{I}$ represent the groundtruth and reconstructed image respectively.

Structural Similarity Index (SSIM)

The Structural Similarity Index (SSIM) [Wang *et al.* 2004], calculated on arbitrary windows of an image, measures similarity between two images by quantifying the quality degradation from reference image to processed image. Since generative models introduce nuisance factors that degrade the perceptual quality of the synthesised image, SSIM quantifies the perceptual similarity between two images based on their luminance, contrast and structure. The range of decimal values is between-1 and 1, where 1 indicates that the images are identical. The SSIM index is described as follows:

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

The equation above is described by the following variables:

- μ_x : the average of x
- μ_y : the average of y
- σ_x^2 : the variance of x
- σ_y^2 : the variance of y
- σ_{xy} : the covariance of x and y
- $\boldsymbol{L}$: the dynamic range of the pixel values
- $c_1 = (k_1 L)^2, c_2 = (k_2 L)^2$: stabilise the division with weak denominator

Learned Perceptual Image Patch Similarity (LPIPS)

The Learned Perceptual Image Patch Similarity, typically used for generative models and image synthesis, measures the perceptual similarity between two images [Zhang *et al.* 2018]. Traditional methods, such as MSE, perform pixel-wise differences to define an image reconstruction loss. However, LPIPS describes a reconstruction loss on higher-level perceptual details based on deep learning methods, such as convolutional neural networks. Pre-trained convolutional neural networks extract abstract high-dimensional

features to simulate human perception. A lower LPIPS value implies a higher perceptual similarity between compared images.

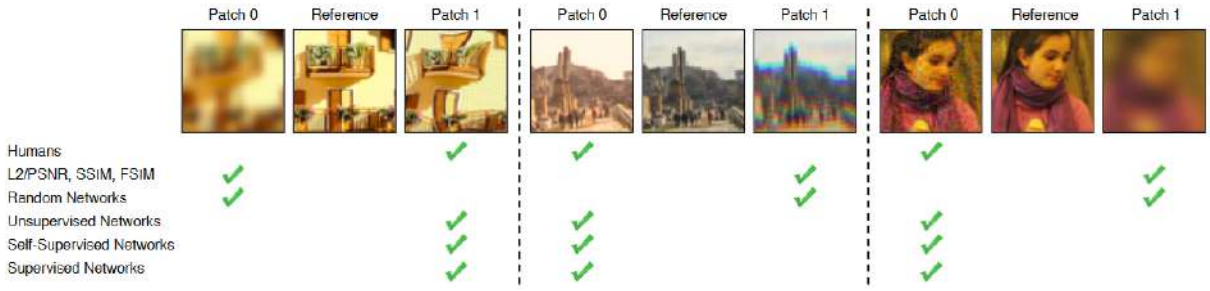


Figure 5.1: Zhang *et al.* [2018] asks the question: *Which patch (left or right) is 'closer' to the middle patch*

Inception Score (IS)

The Inception Score measures the quality of generated images by capturing a score that measures the realism of the output. GANs acquire a distribution that mirrors the training dataset, making it evident that assessing how effectively the distribution produces distinct images measures the model's diversity. The inception score, based on a KL-divergence, utilises a pre-trained inception network trained on the ImageNet dataset to summarise predictions made from classifying outputs of the generative network. The generated results are scored highly if the perceptual quality is good. Thus, a better perceptive quality reflects a higher inception score. Hence, the equation is defined as follows:

$$IS(G) = \exp\left(\frac{1}{N}\sum_{i=1}^{N} D_{KL}\left(p\left(y \mid x^i\right) \| \hat{p}(y)\right)\right)$$

where x indicates the generated image samples and y indicates the labels. Also, the probability density functions describe [Barratt and Sharma 2018]:

- $\exp$: Makes it easier to compare values
- $p(y|x^i)$: Conditional class distribution
- $\hat{p}(y)$: Empirical distribution of classification labels

In conclusion, the inception is restrcted to the inception network trained on the ImageNet dataset. This means that it suffers from out-of-distribution performance evaluation. Considering the inherently natural image-based nature of our datasets, the Inception Score should remain a reliable metric for our purposes.

5.2.2 Qualitative Evaluation

Qualitative evaluation in the context of computer vision refers to the assessment and analysis of visual results or outputs using human judgment and subjective criteria, rather than quantitative measurements or metrics. It involves examining the quality, characteristics, and perceptual aspects of computer vision algorithms, models, or systems, often through visual inspection and human interpretation. Qualitative evaluation plays a crucial role in understanding the strengths and limitations of computer vision solutions, as well as in identifying areas for improvement. Here are some key aspects of qualitative evaluation in computer vision:

Human Study

Human studies are a fundamental approach to conducting qualitative evaluation in various fields, including computer vision. In these studies, researchers engage human participants to assess and analyse the perceptual and experiential aspects of computer vision systems or applications. Typically, human studies involve defining research objectives, selecting a diverse group of participants, and designing scenarios or tasks that replicate real-world interactions with the technology. During the study, qualitative data is collected through observations, user feedback, and open-ended questions.

The collected qualitative data is then subjected to rigorous analysis techniques, such as coding or thematic analysis. Researchers identify recurring patterns, themes, and insights from participants' feedback and behaviors. These qualitative findings offer valuable perspectives on how users perceive and interact with computer vision systems, revealing strengths, weaknesses, usability challenges, and opportunities for improvement. Qualitative evaluation through human studies provides a holistic understanding of user experiences and ensures that technology aligns with user needs and expectations.

The results of human studies inform the development and refinement of computer vision systems, enabling researchers and developers to make evidence-based decisions. Through iterative cycles of qualitative evaluation, systems can evolve to better serve users, enhance usability, and address usability issues. Overall, human studies play a pivotal role in assessing the real-world applicability of computer vision technologies and ensuring they are user-centered, effective, and user-friendly. As such, we define our questionnaire to address the following:

- Which image is the most photo-realistic?
- Which image closely resembles the desired clothing item?
- Which image preserves the characteristics of the person (i.e skin, hair)?

Limitation and Artifact Analysis

In the context of virtual try-on systems, limitation and artifact analysis involves assessing the constraints and potential drawbacks associated with the technology. This analysis is crucial for understanding the system's shortcomings and identifying areas for improvement. Here are some criteria commonly considered in limitation and artifact analysis for virtual try-on:

- **Clothing Realism:** Evaluate how realistically the virtual clothing items are rendered on the user's body. Look for issues like inaccurate draping, unrealistic texture mapping, or clothing clipping through the user's body, as these can detract from the authenticity of the try-on experience.
- **Fit and Sizing:** Assess the accuracy of sizing and fit prediction. Identify instances where the virtual clothing items do not conform well to the user's body shape or size. Inaccurate fit can lead to an unrealistic representation of how the clothing would actually look and feel.

5.3 Summary

In this chapter, we have discussed the experimental setup for fine-tuning state-of-the-art architectures on our **PoseCam** dataset. We have explained the structure of the dataset

and the data splits that are necessary to ensure that the model generalises on unseen data. We have also discussed the training procedure for the virtual try-on networks, explaining the different approaches that we use to train the networks and the hyper-parameter sweeps that we perform to ensure that the best method is chosen during validation.

We established a clear distinction between quantitative and qualitative analysis methods, underscoring the importance of this bifurcation. Quantitative metrics offer an objective assessment of performance, allowing for direct comparisons across different methods based on numerical data. These include measures such as accuracy, speed, and scalability, which provide a tangible gauge of each network's efficacy.

Conversely, qualitative analysis brings to light the subjective aspects of virtual try-on results, such as visual realism and user satisfaction. This type of analysis often involves user studies and expert reviews, offering insights that quantitative metrics alone cannot capture. The distinction between these analysis types is crucial, as it ensures a comprehensive evaluation that accounts for both the technical performance and the user-centered experience of virtual try-on solutions.

We conclude this chapter, and set the stage for an in-depth discussion on the outcomes of our experimental endeavors. In Chapter 6, we will unveil the comparative performance of each method, illuminated through the lens of our rigorous evaluation framework. This will include a critical analysis of the strengths and limitations inherent in our approach.

Chapter 6

Results

This chapter discusses the results of our experiments. It separates the sections according to the virtual try-on stage of the pipeline. To facilitate the experiments, we load the checkpoints of each architecture pretrained on their original datasets. Then, we fine-tune these architectures on our dataset to leverage the knowledge stored withinin each network's weights. This setup is crucial for understanding how we compare different approaches. For each stage, we analyse the qualitative and quantitative performance on the validation dataset giving insight for which network to choose for the training phase.

In our research, we explored various methodologies to evaluate and compare the training performance of different architectures, each designed with unique goals and employing distinct loss functions for virtual try-on tasks. Comparing training objectives of different architectures is challenging because each model optimises for a specific characteristic or quality of the virtual try-on experience, such as fit accuracy, fabric draping, or visual realism. These goals are directly influenced by their respective loss functions. For instance, a virtual try-on model utilising a structured loss focuses on the precise placement and fit of garments [Ge *et al.* 2021]. In contrast, a model employing a perceptual loss prioritises the realistic rendering of fabric textures and colours, aiming to enhance the overall visual appeal and believability of the try-on experience [Choi *et al.* 2021; Lee *et al.* 2022]. Thus, the diverse goals and corresponding loss functions of these models highlight the inherent challenge in directly comparing their performance, often illustrated by their training curves.

Initially, we attempted a promising methodological approach that involved normalising each model's loss function to a range between 0 and 1—effectively comparing losses as percentages of their initial values. This normalisation was intended to mitigate the inherent disparities in scale and magnitude of the raw loss values, thus providing a standardised basis for comparison. However, upon a detailed review of the literature [Lecun *et al.* 1998; Glorot and Bengio 2010; Cheng *et al.* 2021; Johnson *et al.* 2016; Han *et al.* 2018] and extensive preliminary analyses, we confronted significant challenges inherent in this approach, primarily due to the distinct nature of the loss functions employed by different models. These challenges include:

- **Scale, Magnitude, and Behavior Differences:** Each model's loss function operates on a different scale and with different sensitivities. For instance, a model utilising mean squared error (MSE) might begin with very high loss values that rapidly decrease, whereas a model using a perceptual loss such as VGG might start lower and decrease at a different rate [Lecun *et al.* 1998]. Consequently, even when normalised, a value of 0.5 in MSE does not necessarily correspond

qualitatively or quantitatively to a 0.5 in VGG loss, rendering direct comparisons not only inappropriate but potentially misleading. These disparities make it difficult to interpret normalised losses directly since the same numerical value across different loss functions can imply different things about model performance.

- **Convergence Patterns:** Different loss functions also exhibit unique convergence behaviors. Some might show a steep decline initially and plateau quickly, while others might decrease slowly but continuously [Glorot and Bengio 2010]. This difference in how loss values evolve during training makes it challenging to use normalised values to compare models effectively, as a visually similar convergence pattern can mask fundamentally different underlying model behaviors.
- **Non-Commensurability:** The metrics optimised by different loss functions are often not directly comparable. A loss function designed to optimise pixel-wise accuracy (like MSE) is fundamentally measuring a different aspect of the model's output than one optimised for perceptual quality (like VGG) [Johnson *et al.* 2016]. This non-commensurability means that even normalised, the loss values are not measuring the same phenomena, rendering comparisons across models conceptually flawed.

Given these findings and the lack of substantive evidence in the field supporting the effectiveness of normalised loss comparisons for drawing meaningful conclusions, we opted against employing this method in our study. Instead, we adopted a multi-faceted evaluation approach, incorporating both quantitative metrics suitable to each model's design objectives and qualitative assessments to capture the perceptual and practical aspects of model performance. This decision aligns with contemporary best practices [Cheng *et al.* 2021; Lecun *et al.* 1998] in model evaluation, emphasising the necessity of contextually appropriate and task-specific assessment methods over generalised or overly simplified metrics.

Appendix A explains how we select the best combination of hyperparameters for each architecture using hyperparameter sweeps. This analysis allows us to determine the generalisability of each architecture on unseen data. Lastly, we demonstrate the limitations and failure cases on these state-of-the-art methods on the **PoseCam** dataset.

6.1 Clothing Warping

In this section, we discuss the outcomes and implications of warping clothing items on the PoseCam dataset. By examining the results, we aim to evaluate the impact and effectiveness of introducing diverse perspective and camera angles. Additionally, we explore the challenges of realistic representations of clothing on varying poses.

In Figure 1.1, we illustrate the significance of our dataset by demonstrating how state-of-the-archictures, without finetuning on the **PoseCam** dataset, fail to accurately warp clothing. Following this, Figure 6.1 illustrates the result, across all state-of-the-art virtual try-on architectures, of warping on the **PoseCam** dataset to compare their qualitative performance. Here, we observe each architecture's ability to accurately warp the clothing item and mimic the ground truth's contours.

As depicted in Figure 6.1, these methods demonstrate proficiency in style transfer with correct orientation of textual elements on clothing. In the case of each model, the results demonstrate that the architecture warps the clothing accurately with the excep-

tion of replicating lighting conditions, folds and creases present in the clothing item. This indicates a lack of robust feature-mapping that preserves high-level details across transformations. Furthermore, these models exhibit limitations in dealing with complex articulations, specifically shoulder rotations.

Despite the anticipated challenges of warping downscaled interpolations of clothing items, HR_VITON demonstrates competitive performance. This indicates that even when trained on downscaled images, HR_VITON is capable of capturing essential features necessary for the warping task. In fact, the steep curve leading to the lowest validation warping loss for HR_VITON suggests that the model is capable of learning complex warping patterns and generalising on unseen data.

The current results show that while the edges of clothing are adjusted to match the pose, the alignment does not consistently reproduce the 3D rotation of the shoulder joints. For CP_VTON and CP_VTON+, the right arm of the ground truth (3^{rd} row) image, which is occluded due to the rotational angle, is not faithfully replicated in the warped prediction. From the results, we can deduce that precise 3D-aware replication of body rotations appears to be a challenge for 2D warping architectures.

In comparison to the parser-free networks, PF_AFN, DM_VTON, and FS_VTON, demonstrate a higher degree of accuracy in replicating the ground truth clothing items. The models exhibit a high level of detail and precision in replicating the texture, patterns, and folds of the clothing, as well as the overall fit and alignment with the body. However, PF_AFN (3^{rd} row) struggles with rotating the garment and synthesising the arms portion of the clothing. This is evident from the misalignment of the sleeves and the inaccurate replication of the arm's posture.

Figure 6.1: Results of applying parser-based warping networks to transform the input clothing item. The leftmost image is the original clothing inpu.

The warping models are designed to map a 2D representation of clothing onto a target body image. While these models can handle certain affine transformations, such as translation and scaling with relative success, they inherently struggle with the complex non-affine transformations associated with rotation, particularly in a 3D context.

The crux of the issue lies in the model's limited understanding of the depth and the resulting occlusions when garments are rotated in real-world scenarios. In the case of

shoulder rotation, for instance, the warping model must not only adjust the garment to match the new position but also account for the changes in the garment's appearance due to the rotation. Real-world fabric does not merely shift; it bends, folds, and wraps around the body. These dynamic changes are not fully captured by current 2D warping models.

Moreover, when a part of the body or garment becomes occluded due to rotation, the model must infer what is unseen, a task that requires a sophisticated understanding of the garment's structure beyond what is visible in the input image. The ability to predict occlusion involves reconstructing the occluded parts of the body and garment, which, in turn, requires a form of spatial reasoning that current models lack. The challenge is compounded when the rotation leads to self-occlusion, where different parts of the same garment overlap. This type of occlusion is particularly problematic because it involves predicting the interaction between different parts of the fabric, which may not be directly observable in the training data.

In Table 6.1, we begin with PSNR (Peak Signal-to-Noise Ratio) as it is commonly used to measure the reconstruction quality of lossy transformations. A higher PSNR indicates a higher fidelity of reconstruction.

Following FS_VTON's performance, the architecture led the ranking in all the evaluation metrics, implying its superior capability in minimizing reconstruction errors and producing warping predictions that are perceptually closer to the target images. To support this claim, Dam *et al.* [2024] emphasize the importance of inference time alongside traditional quality metrics such as SSIM and LPIPS. Their approach focuses on refining the warping and synthesis stages to maintain identity consistency and minimize artifacts, supporting the high performance of FS_VTON in terms of both perceptual similarity and overall image quality. Cui *et al.* [2023] provide further evidence for the effectiveness of sophisticated warping techniques used in methods like FS_VTON.

Considering SSIM (Structural Similarity Index), which evaluates the visual impact of three characteristics of an image: luminance, contrast, and structure, a higher SSIM score represents better preservation of these qualities.

Table 6.1: Inspecting the visual quality of warping architectures (* refers to optimal performance).

Method	PSNR ↑	SSIM ↑	LPIPS ↓	IS(mean) ↑
CP_VTON	36.94	0.8867	0.096	4.39
CP_VTON+	36.59	0.8617	0.1224	3.56
HR_VITON	35.27	0.8855	0.1200	3.43
PF_AFN	37.09	0.8968	0.089	4.25
FS_VTON	37.10*	0.8981*	0.0860*	4.72*

FS_VTON shows the highest values for both PSNR and SSIM (37.10 and 0.8981, respectively), indicating that it produces high-quality and structurally accurate images. PF_AFN follows closely with slightly lower values (37.09 for PSNR and 0.8968 for SSIM), still indicating high performance. PF_AFN is designed to handle a wide range of poses without relying heavily on pose estimation. This robustness to pose variation can lead to better preservation of structural details and fewer warping artifacts, resulting in higher PSNR and SSIM values [Ge *et al.* 2021]. In contrast, HR_VITON might strug-

gle with more complex poses, leading to lower quality reconstructions. Furthermore, PF_AFN's use of flow-based warping techniques, where it calculates per-pixel flow fields to align the garment with the person's body, can be more effective in preserving fine details and minimizing artifacts compared to traditional warping techniques used by HR_VITON. This can result in higher fidelity reconstructions (higher PSNR) and better structural similarity (higher SSIM) [Cui *et al.* 2023; Dam *et al.* 2024]. CP_VTON and CP_VTON+ show competitive values but are slightly lower than FS_VTON and PF_AFN. HR_VTON has the lowest PSNR and SSIM values, suggesting it might produce lower quality and less structurally accurate images compared to the others.

FS_VTON has the highest IS (Inception Score) of 4.72, suggesting it produces the highest quality and most diverse images. PF_AFN and CP_VTON also show high IS values (4.25 and 4.39, respectively), indicating good performance. CP_VTON+ and HR_VTON have lower IS values (3.56 and 3.43, respectively), indicating that these methods might produce less diverse or lower quality images compared to the others.

Based on the provided metrics, FS_VTON emerges as the most effective virtual try-on method for warping, with the highest scores in PSNR, SSIM, and IS, and the lowest LPIPS value. PF_AFN also performs very well, closely following FS_VTON in most metrics. CP_VTON and CP_VTON+ show moderate performance, while HR_VTON lags behind the other methods. This comprehensive analysis underscores the superiority of FS_VTON in delivering high-fidelity, perceptually accurate, and diverse virtual try-on results.

6.2 Image Composition

In this section, we discuss the outcomes of training the try-on module of each architecture's ability. We evaluate each architecture's to synthesise clothing items onto images of individuals, assessing the visual realism and synthesis quality of the predictions. Through an of examination of the qualitative performance, we highlight the effectiveness in generating convincing virtual try-on images, the challenges encountered in consistently synthesising clothing items across different perspectives.

Similarly, in Figure 1.2, we illustrate how state-of-the-art virtual try-on methods trained without the **PoseCam** dataset fail to accurately synthesise warped clothing onto individuals. Following this, Figure 6.2 compares the composition performance across each virtual try-on architecture in a paired setting, where the clothing image represents the same item that the individual in the person image is wearing. The results reveal that each composition network exhibits a high degree of visual consistency and realism, with the garments seamlessly integrated into the person images. The clothing items conform naturally to the body contours, and the overall visual quality is impressive, reflecting the models' proficiency in generating realistic virtual try-on compositions.

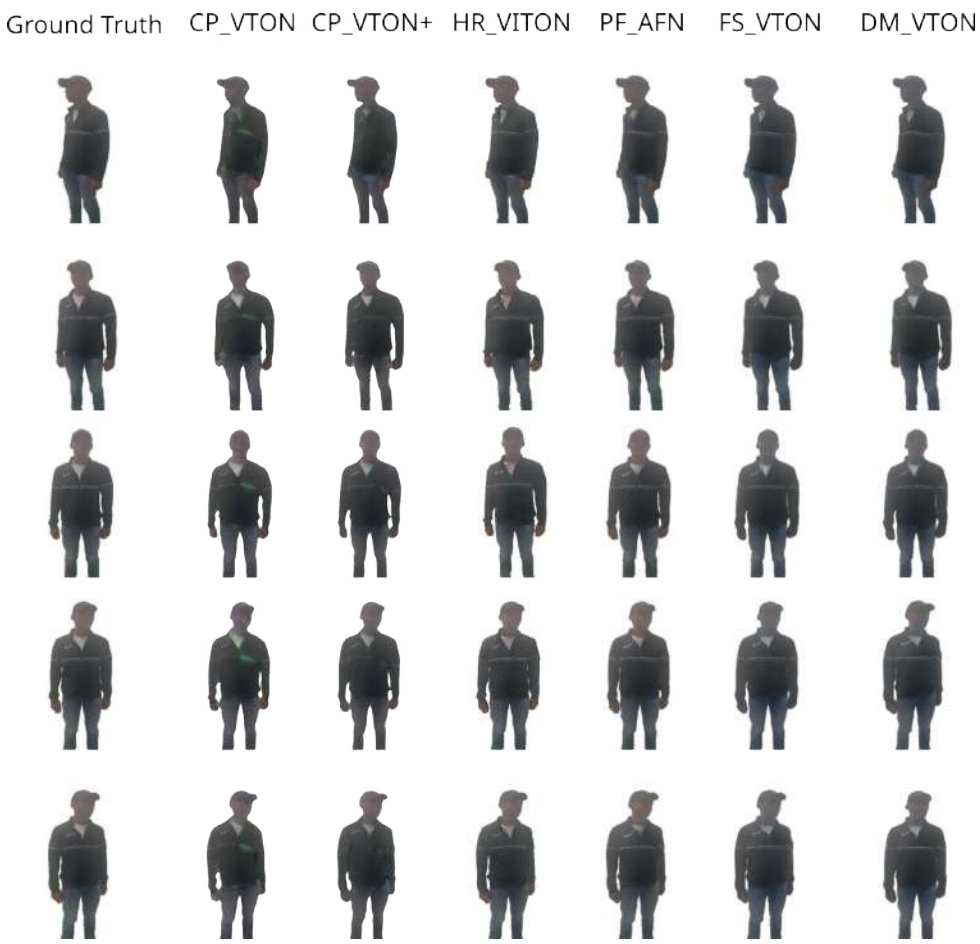


Figure 6.2: Comparing virtual try-on methods within a paired setting.

Looking closely, CP_VTON struggls to maintain the white line across the jersey while CP_VTON+ is able to maintain the white line across the jersey. This is a clear indication that CP_VTON+ hallucinates the greens pixels surrounding the torso area. Considering visual inspection, the garments in CP_VTON are more accurately aligned with the body contours, and the overall visual fidelity is higher, indicating a more realistic and immersive virtual try-on experience.

Figure 6.3 compares the performance of each composition network in an unpaired setting, where the input clothing item is not matched with the person image. CP_VTON accurately warps and synthesises the clothing item onto the individual. However, CP_VTON+ appears to present overlapping pixels between the clothing item and the individual's skin. In the 4^{th} and 5^{th} row, the network struggled to warp the jacket appropriately affecting the quality of composition.

On the other hand, HR_VITON blends the original and target clothing items causing the network to attempt synthesising both items onto the individual. We hypothesise that the network overfit and memorised the training data from the original dataset and adequately adapt to the new dataset. This may be indicative of inadequate style transfer whereby the network fails to adapt to the new dataset's characteristics. We summise that early stopping on the original dataset may improve the network's performance on the new dataset. In this instance, the network would learn the overall structure and contours of the clothing items, and allow finetuning to adapt to newer style and textures.

Figure 6.3: Comparing virtual try-on methods within an unpaired setting.

The parser-free networks outperform the parser-based networks in the unpaired setting. Each of these networks adequately warp and synthesise clothing items on the individual across different angles. Furthermore, these networks preserve miniature details, such as the white line across the jersey and the logo, while leaving the original characteristics of the individual intact. This indicates that the parser-free networks are more adept at virtual try-on in both paired and unpaired settings given diverse perspectives. We hypothesise that parser-free networks outperform parser-based networks, on the **PoseCam** dataset, as a result of more training time. Since parser-free networks distill knowledge from pretrained parser-based networks, we essentially train parser-free networks twice as many times as parser-based networks. This allows the parser-free networks to learn the intricate details of the dataset and adapt to the new dataset's characteristics.

Table 6.2: Inspecting the visual quality of composition architectures (* refers to optimal performance).

Method	PSNR ↑	SSIM ↑	LPIPS ↓	IS(mean) ↑
CP_VTON[2]	28.162	0.8763	0.05	2.63
CP_VTON+[3]	28.030	0.8258	0.11	2.86
HR_VITON[1]	36.62*	0.9150	0.01*	2.26
PF_AFN[4]	35.950	0.9214*	0.02	2.67
DM_VTON[6]	34.94	0.890	0.05	2.96*
FS_VTON[5]	35.160	0.9070	0.03	2.68

In Table 6.2, a detailed evaluation of composition architectures in virtual try-on applications is presented, highlighting the nuances of visual quality assessment through various metrics. Among the contenders, PF_AFN is the optimal architecture with regards to SSIM and IS values, coupled with low minimal LPIPS scores. This underscores PF_AFN's proficiency in producing compositions that are not only visually appealing

and realistic but also closely aligned with the original visual fidelity, exhibiting minimal perceptual discrepancies.

With the highest PSNR, HR_VITON underlines its ability to produce images that are closely matched to the original in terms of pixel accuracy. In the context of virtual try-on, this means that the garments superimposed onto the model's image retain their structure, colour, and detail to a degree that closely resembles a natural photograph. High PSNR values indicate that the visual distortion introduced during the virtual try-on process is minimal. However, PSNR does not account for perceptual differences, which are crucial in evaluating the style and texture of the garments.

On the other hand, achieving the lowest LPIPS score showcases PF_AFN's ability in capturing the perceptual likeness of the target garment on the virtual model. LPIPS is designed to measure the similarity between images in a way that mirrors human visual perception, focusing on aspects like texture, colour, and structural integrity. A low LPIPS score suggests that the synthesised images not only have high pixel-wise accuracy but also maintain the integrity of the garment's appearance in a way that feels natural and realistic to human observers

Contrastingly, CP_VTON and CP_VTON+ lag in terms of PSNR and SSIM, suggesting their generated images may not consistently match the visual quality of PF_AFN. Their higher LPIPS scores further imply a noticeable gap in achieving perceptual similarity with the ground truth. Moreover, their IS metrics fall short compared to other methods, indicating a potential limitation in encapsulating the dataset's varied characteristics.

FS_VTON also shows high PSNR and SSIM scores, indicative of high visual consistency. Additionally, its IS score does not reach the benchmark set by PF_AFN, suggesting room for improvement in capturing the dataset's breadth of diversity and intricacy.

6.3 Limitations and Failure Cases

Virtual try-on systems face various limitations that need to be systematically analysed to improve their accuracy and realism. Some of the primary challenges include [Shim *et al.* 2023; Chen *et al.* 2024b; Zhang *et al.* 2024]:

- **Clothing Fit and Alignment:** Ensuring that virtual garments align accurately with the body shape and pose is critical. Misalignment issues, such as improper warping and deformation, can result in unrealistic appearances. Quantitative measures, like intersection-over-union (IoU) for overlap accuracy, can help assess this dimension.
- **Texture Preservation:** Maintaining the original textures of garments during the warping and rendering processes is essential for visual realism. Issues like texture blurring or stretching can degrade the quality of the output. Perceptual metrics like the Structural Similarity Index (SSIM) can be used to evaluate texture preservation.
- **Garment-Body Interactions:** Ensuring natural draping and interaction between garments and the body, without unrealistic penetrations or gaps, is important for realism. This involves accurately modelling how garments should behave and interact with different body types.
- **Semantic Consistency:** Accurately segmenting and aligning garments where they meet different body parts is necessary to avoid visual artefacts. This involves

ensuring that the virtual try-on system can correctly interpret and process various body regions in relation to the garments.

- **Detail Fidelity:** Preserving fine details like buttons, seams, and patterns is crucial for high-quality virtual try-on results. Fine details are often lost during the processing, which can reduce the authenticity of the virtual garments.

Our research focused primarily on three key aspects: clothing fit and alignment, texture preservation, and garment-body interactions. This decision was driven by the need to expand the utility of virtual try-on systems to dynamic views, where garments must maintain their fit and appearance across different poses and angles.

We discuss the limitations and failure cases observed in the warping and composition processes of virtual try-on, notably issues such as bleeding artifacts, inadequate performance in handling occlusion, and inaccuracies in representing 3D-aware rotational poses. To address these challenges, we developed methods to circumvent common issues such as misalignment and improper garment deformation. This involved leveraging alpha blending to ensure a more accurate fit of virtual garments on various body shapes and poses

Figure 6.4 demonstrates an artifact encountered in warping architectures where adjusting a clothing item to fit the pose and shape of a virtual model fails to enclose the pixel transformation within the contours of the garment. This artifact manifests when the warped clothing item extends beyond the intended boundaries of the garment, causing parts of the jacket to appear as if they are 'bleeding' over into the background or onto other regions of the virtual model.

Figure 6.4: Here, we demonstrate the artifacts generated when fine-tuning state-of-the-art warping networks on the **PoseCam** dataset. The left-hand side represents the ground truth. The right-hand side is the prediction of a warping network.

This bleeding effect is often occurs when the warped clothing mask is not bounded by the contour of the individual's torso. Hence, the warping process struggles to adapt to convex outlines that bound the pose of the individual. Edges, creases and folds of the clothing item are exagerrated beyond the body's shape and pose. In the case for HR_VTION, it results in inaccurate segmentation maps, which define the regions where the clothing should be placed. If these maps are incorrect, the garment may not conform correctly to the model's body, leading to the observed bleeding.

In response, Figure 6.5 illustrates how we address this issue by applying alpha blending between the warped clothing and its binary mask such that the bleeding artifacts are removed.

Current models lack the depth perception required to fully understand the 3D nature of human poses. Referring to Figure 6.4, the warping network fails to occlude the right

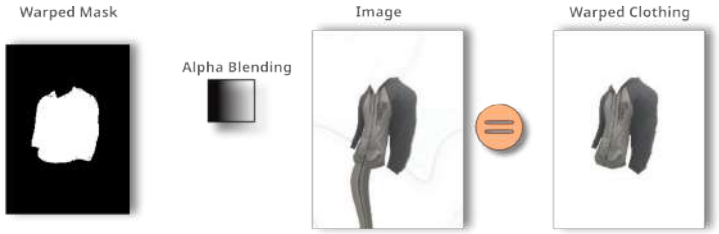


Figure 6.5: Applying alpha blending between the warped clothing and its binary mask to remove bleeding artifacts created by misaligned regions.

jacket arm due to the alternation of the angle. Although the network can adjust the angle of the shoulder region, it fails to rotate the warped prediction to match the pose of the ground truth. Enhancements in pose estimation algorithms and the inclusion of diverse rotational poses in the training set may provide the necessary data for a more accurate representation of shoulder mechanics.

Another contributing factor could be the limitations in the current warping architecture's ability to handle fine details and extreme angles, which are common challenges in virtual try-on technology. When a model adopts a pose that is significantly different from those seen in the training data, the warping model may struggle to predict the correct garment shape and placement.

In the case of preserving identity, Figure 6.6 illustrates how virtual try-on currently fails to faithfully maintain the skin pigmentation of individuals when superimposing garments that failed to remove the skin during warping. This is an important factor in maintaining the realism of the virtual try-on outputs, ensuring that the garments interact naturally with diverse skin tones without causing unrealistic visual artifacts. As such, character preservation remains a challenge for real-world virtual try-on.

Figure 6.6: Synthesising garments from darker skin tones onto individuals with lighter skin tones creates skin pigmentation artifacts

Our primary contribution was aimed at expanding the utility of virtual try-on systems to support dynamic views. This included ensuring that the virtual garments maintain

their fit and alignment even when the user changes poses or angles, thereby providing a more comprehensive and realistic virtual try-on experience.

6.4 Summary

A significant limitation in existing literature is the reliance on frontal-facing datasets, which restricts the ability to perform virtual try-on when individuals are not facing the camera [Chen *et al.* 2024a]. Our main contribution and findings address this limitation by extending the versatility of 2D virtual try-on. Our dataset allows clothing items to be superimposed from various angles, beyond just frontal-facing images, thus providing a more comprehensive and realistic virtual try-on experience. This advancement demonstrates the potential for more flexible and accurate virtual try-on applications in real-world scenarios

In this chapter, we discuss the results of our experiments to demostrate the feasibility of incorporating various perspectives for virtual try-on. Our dataset allows state-of-the-art architectures to adapt to diverse scenarios that aid in the task of real-world virtual try-on. In Section 6.1, we anaylse the experiments conducted during the warping phase, and compare the result of fine-tuning virtual try-on architectures on the **PoseCam** dataset for accurate cloth modeling accounting for diverse poses.

Following cloth warping, we evaluate the performance of image composition, in Section 6.2, and discuss the experiments that facilitate synthesising warped clothing items onto individuals. Our results show that parser-free networks outperform parser-based networks during composition. The validation of our warping and composition methods is quantified through extensive experiments. We present validation curves for both parser-based and parser-free approaches, providing insights into the performance and reliability of each method across various scenarios. These curves highlight the robustness of our models in handling complex warping and composition tasks.

Although 3D reconstruction is not the primary aim of our dataset, it is designed with features that facilitate 3D scene reconstruction as a by-product. This includes multi-angle image captures and depth information that can be utilised for creating 3D models. Consequently, we focus our findings on 2D virtual try-on within this chapter, showcasing the significant improvements our dataset offers in this domain. However, recognising the future potential of our research, we introduce a pilot study in Chapter 7. This pilot study explores the preliminary results of using our dataset for 3D virtual try-on, paving the way for future advancements and demonstrating the broader applicability and versatility of our dataset for comprehensive virtual try-on experiences.

In Section 6.3, we address the limitations and failure cases encountered in our research. We openly discuss challenges such as handling diverse poses, garment interactions with occlusion, and the inability for 2D virtual try-on to represent depth.

In Chapter 8, we conclude this dissertation with a forward-looking perspective, outlining potential avenues for future work to overcome these hurdles.

Chapter 7

A Pilot Study For Reconstructing A 3D Virtual Try-On Scene Using The PoseCam Dataset

Our goal is to enable users to try-on virtual clothing items in a realistic and immersive manner. To achieve this, we begin the dataset construction phase by capturing videos of individuals standing on a platform. These videos are then processed to extract image frames suitable for 3D scene reconstruction. This process allows us to create a detailed representation of the scene, including the pose and shape of the individual.

Next, we leverage virtual try-on techniques to render clothing items on the reconstructed 3D scene. We utilise warping techniques to transform the clothing items to match the pose of the individual. This ensures a realistic fit and appearance of the virtual clothing. To further enhance realism, we employ generative modelling techniques to synthesise the warped clothing items on the individual. This allows us to generate realistic textures and details, making the virtual try-on experience more immersive and visually appealing.

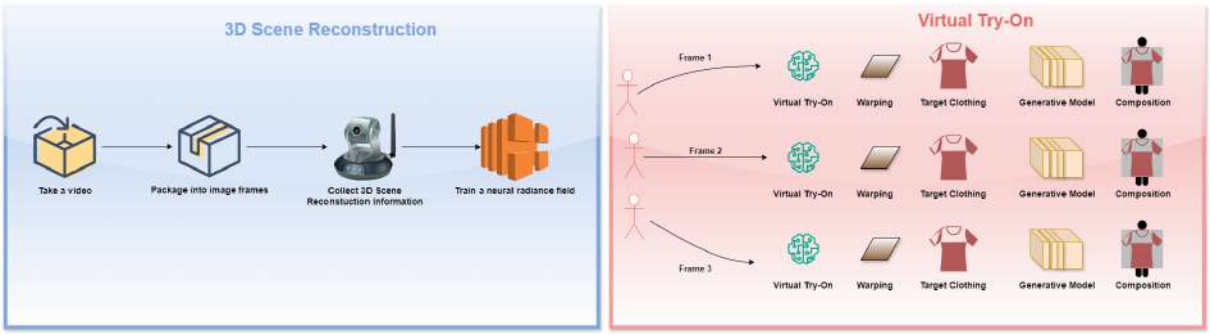


Figure 7.1: In this figure, we illustrate the pipeline that facilitates the interaction of various computer vision techniques required to integrate 3D scene reconstruction with virtual try-on. Firstly, we capture videos of individuals standing on a platform. Then, we extract image frames suitable for 3D scene reconstruction and represent the scene in pose transformation matrices allowing us to train neural radiance fields to render a novel view. For each frame generated from a neural radiance field, we leverage virtual try-on to warp clothing items to match the pose of the individual. We then use generative modelling techniques to synthesise the warped clothing items on the individual.

Figure 7.1 gives a broad overview of how our pipeline utilises the **PoseCam** dataset to integrate 3D scene reconstruction with virtual try-on to simulate a real-world try-on experience. In the following sections, we will delve into the background of each step required in our approach and elaborate on vision techniques required to implement the pipeline.

7.1 Background

Seminal work in computer graphics aimed to reconstruct 3D scenes by interpolating between rays in space, setting the groundwork for subsequent advancements. Building on this, novel view synthesis emerged, utilising a relatively small number of photographs—referred to as a sparse set of pictures—from varying angles to reconstruct a 3D scene with greater efficiency. Unlike a dense set of pictures, where the scene is captured from nearly every conceivable angle, a sparse set means there are fewer images and, consequently, larger intervals between the vantage points from which each photo is taken. This approach led to the development of neural volume rendering, a process that traces a ray into a scene and integrates along the ray's path to represent the scene with increased accuracy.

7.1.1 3D Scene Reconstruction

At the heart of this process are continuous differential models, which encode functions that map a coordinate space to scalar quantities such as density and colour, providing a detailed representation of the scene. Through these models, neural rendering methods are able to yield images that offer novel views of a 3D scene, showcasing the technique's ability to generate perspectives from arbitrary directions. The end goal of encoding a scene within the weights of a neural network, as facilitated by novel view synthesis, is precisely to achieve this flexibility in generating novel views, thereby inferring 3D scene properties from images.

This technique presents a valuable alternative to the expensive process of storing 3D scenes as collections of voxels or polygon meshes, highlighting novel view synthesis's efficiency in reconstructing 3D representations of objects. Furthermore, the differentiable and continuous nature of these representations allows neural networks to incorporate learning, enabling them to more effectively encode a 3D scene.

7.1.2 From Images to 3D Models: Structure from Motion and Multi-View Stereo

Structure from Motion (SfM) is a computer vision technique that constructs detailed 3D models of a scene from images captured at different viewpoints. It identifies and matches unique features across these images using sophisticated algorithms, then minimises reprojection errors to accurately estimate the camera's position and the 3D coordinates of scene features. SfM iteratively refines the camera poses and feature coordinates with each added image, ensuring high accuracy and model consistency. The process further enriches the 3D model by applying multi-view stereo (MVS) to the generated point cloud, producing denser and more detailed representations. This efficient and iterative approach allows SfM to create high-quality 3D models from a minimal

set of images, making it invaluable in various applications like heritage preservation, virtual reality, and robotics.

Multi-View Stereo (MVS) is a sophisticated computer vision technique designed to reconstruct 3D models from a collection of images captured from various viewpoints. This method intricately processes a set of overlapping images of a particular scene or object, leveraging the unique intrinsic (e.g., field of view) and extrinsic (e.g., camera position and orientation) camera parameters associated with each image. At the heart of the Multi-View Stereo (MVS) process is a step known as feature extraction. In this stage, specialised algorithms search for and pinpoint standout points within the collection of images. These points are distinctive and consistent across different photos, much like unique landmarks that can be recognised from several viewpoints. Techniques such as SIFT (Scale-Invariant Feature Transform) or SURF (Speeded Up Robust Features) are employed for this task. Once these landmarks are identified, the method uses them to align pairs of images and create depth maps, which are akin to intricate topographical charts that denote the distance of each part of the scene. These maps are then pieced together to assemble a comprehensive 3D model. The initial form is refined by applying a mesh (think of it as a digital framework for the sculpture) and textures (which add surface details and colour), significantly enhancing the model's realism and intricacy.

Essentially, SfM sketches the initial outline of a scene. MVS, assuming camera positions are known (often determined by SfM), enriches this outline by detailing every aspect of the scene, producing a dense and detailed 3D model. While SfM lays the groundwork by mapping out key points based on camera movements, MVS completes the picture by filling in the intricate details, transforming the preliminary sketch into a full, lifelike 3D representation.

Both MVS and SfM benefit from the contributions of applications like COLMAP, which streamline the 3D reconstruction process. COLMAP [Schönberger *et al.* 2016; Schönberger and Frahm 2016] is an open-source photogrammetry application that automates camera calibration and point cloud reconstruction, effectively bridging the gap between MVS and SfM by simplifying the intricate tasks of feature matching and 3D model creation. This integrated approach is invaluable in various domains, including heritage preservation, virtual reality, and robotics, by enabling the efficient and accurate creation of high quality 3D models from a minimal set of images.

7.1.3 Neural Radiance Fields

Mildenhall *et al.* [2020] introduce the baseline NeRF that performs implicit scene representation (density and colour transformed in the same stage) on a single scene using a multi-layer perceptron. In practice, the NeRF model requires input images with corresponding camera parameters. The camera parameters consist of the following information:

- **Intrinsic parameters**: focal length modeling the distance to the focal plane.
- **Extrinsic parameters**: position and orientation of the camera (i.e translation and rotation).

Neural radiance have a special property that motivate the reason for its notoriety, namely that it outputs a view-dependent colour and view-independent density. This means that the model can capture specular reflections and various appearance effects.

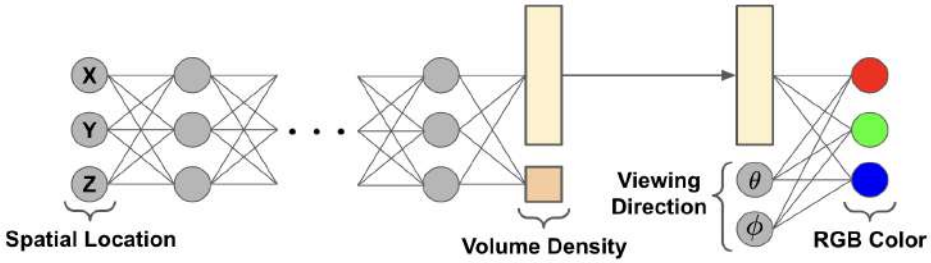


Figure 7.2: The radiance field is modeled by a neural network which takes a 5D input (3D coordinate vector with 2D viewing direction). The 3D coordinate vector representing the spatial location is passed through a feed-forward network to output a feature vector and the volume density. The volume density is concatenated with the viewing direction to predict the RGB colour where an image taken from a sparse set of frames is represented as the groundtruth. In this instance, colour is view-dependent and volume density is view-independent [Wolfe 2019].

Given the view dependency properties, depicted in Figure 7.2, of the output colour and volume density (amount of light passing through point), the model performs the following functions in this order:

- **Volume density network**: A NeRF trains a preliminary network to input spatial location and output the volume density and a feature vector to model the opacity of pixel.
- **RGB network**: The RGB network concatenates the output feature vector of the volume density network with the viewing direction to output a colour (RGB value) representing the viewpoint.

Thus, the network encapsulates the input dependency properties of colour and volume density from the inputs to output a single RGB value to synthesise a novel view. However, given that neural radiance model low-frequency functions, they transform the input spatial location and density into a higher dimensional space using high frequency positional encoding. This allows the network to represent a high frequency model that generates textural details and geometry. Thus the 5D input to the network is transformed to represent the following:

$$x_{\text{input}} = \{x, y, z, \theta, \phi\} \longrightarrow x_{\text{input}} = \{x_1, x_2, \ldots, x_d\} \tag{7.1}$$

Furthermore, differentiable volume rendering aggregrates colour accumulated across multiple rays(query points are accumulated to aggregrate colour for a single ray) from different camera viewpoints to produce an image from samples of discrete data (i.e colour and volume density) representing a 3D scene. In simpler terms, volume rendering renders the camera rays and allocates an RGB value to each pixel forming an image.

A NeRF synthesises a novel view through neural rendering by minimising the reconstruction loss between the predicted image and groundtruth. The reconstruction measures the L_2 objective function between the RGB colours of the synthesised image and the groundtruth for each of the rays that propagate through a scene. The equation can be defined as [Mildenhall *et al.* 2020]:

$$L_{\text{reconstruction}} = \frac{1}{N} L_2(C_{\textbf{pred}}, C_{\text{gr}}) \tag{7.2}$$

where N is the number of pixels.

Radiance Fields

As described previously, the radiance field interacts with surfaces as the ray travels through the scene to describe its colour and volume density. A NeRF encapsulates radiance fields to optimise an underlying volumetric scene function with sparse images from multiple viewpoints to encode a 3D scene within the weights of a neural neural. Thus, the input of the network is a 5D vector-valued function that describes the position(x, y, z) and orientation(θ, ϕ) of each point of the object in the scene. Hence, the model is trained to predict the view-dependent emitted radiance, described as the colour, and volume density (content of the scene) for every viewing direction in the scene. The equation is described as follows:

$$F(\vec{x}, \vec{\theta}) \longrightarrow \{\boldsymbol{c}, \sigma\} \tag{7.3}$$

where the spatial location $\vec{x} = (x, y, z)$ and 2D viewing direction $\vec{\theta} = (\theta, \phi)$ that outputs a colour $\boldsymbol{c} = (r, g, b)$ and the volume density σ. Since volume rendering is naturally differentiable, the network only requires a sparse set of images with their corresponding transformations (i.e camera poses) to synthesise a novel view.

Positional Encoding

As stated in the previous section, neural radiance fields represent a 5D vector-valued function

$$F_\theta(\quad \underbrace{x, y, z}_{\text{position}}, \quad \underbrace{\theta, \phi}_{\text{viewing direction}} \quad) \longrightarrow \{\boldsymbol{c}, \sigma\}$$

that transforms a position and viewing direction into an output consisting of colour (RGB value) and volume density (opacity). Given the 5D vector representation, neural networks transform inputs (position and viewing direction) to synthesise a novel view. Although neural networks are universal function approximators [Hornik *et al.* 1989], they are not well suited to represent high-frequency functions. As such, Mildenhall *et al.* [2020] argue that training a multi-layer perceptron to reconstruct a 3D scene from input coordinates renders novel views with low-frequency variation in colour and geometry. Low-frequency variation in colour means that changes in colour across the image occur more gradually, leading to smoother transitions rather than sharp, detailed colour changes. Similarly, low-frequency variation in geometry refers to broader, less detailed changes in shapes and surfaces within the scene. Hence, neural radiance fields transform input coordinates using high-frequency positional encoding to enable the model to represent high-frequency details. Therefore, low-frequency variation, as a result of not using positional encoding, results in poor texture and geometry representation with oversmoothed appearance. Thus, for any scalar p, the function $\gamma(p)$ converts the input coordinates into a higher dimensional space $\mathbb{R}^{2L}$ such that:

$$\gamma(p) = \left(\sin 2^0 \pi p, \cos 2^0 \pi p, \cdots, \sin 2^{L-1} \pi p, \cos 2^{L-1} \pi p\right)$$

where L is the number of frequencies in the positional encoding. The higher the value of L, the more high-frequency details can be captured.

Shooting A Ray

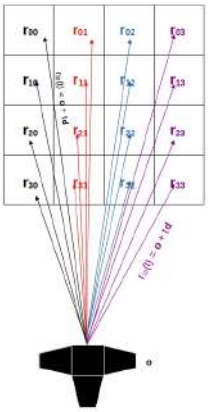


Figure 7.3: Ray marching through camera pixels of an image plane.

A novel view is represented by the model prediction of a neural radiance field. The prediction that forms on the image plane is comprised of pixels such that a ray, depicted in Figure 7.3, is marched from the camera through each pixel location into the scene. Each ray that is defined at every individual pixel location on the image plane is described as a parametric equation of a line:

$$r(t) = \underbrace{\mathbf{o}}_{\text{Camera origin}} + t \underbrace{\mathbf{d}}_{\text{Ray direction}} \tag{7.4}$$

where t is a specific query point defined along a ray. The prediction renders the scene as a projection on the image plane through a process called backwards tracing. The output image is represented by combing the colours along each ray based on the density.

Volume Rendering

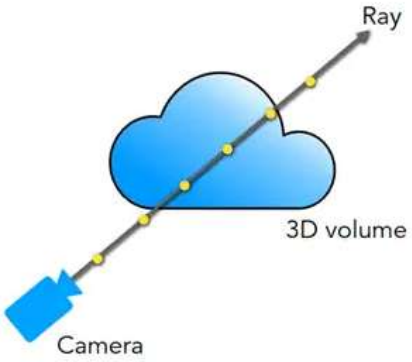


Figure 7.4: Illustration of differentiable volume rendering. This diagram depicts a simplified view where a camera captures a ray passing through a cloud-shaped 3D volume. The ray is discretized with several points, each contributing to the final image formation by simulating the interaction of light within the volume. This approach allows gradients of the image with respect to volume density and color to be computed, enabling gradient-based optimization of the volume [Matt Tancik 2022].

Volume rendering, as demonstrated in Figure 7.4, creates an image of a 3D scene by calculating the colour that a ray would acquire as it traverses the scene. This is achieved by using predictions of the scene's volume density and colour, which are obtained from the outputs of a multi-layer perceptron. We observe the colour as a result of light reflecting partially in the location of the viewer. Thus, the volume rendering equation is defined as [Yagel *et al.* 1992]:

$$\boldsymbol{C}(r) = \int_{t_1}^{t_2} T(t) \cdot \sigma(r(t)) \cdot c(r(t), d) \cdot dt \quad \text{where } T(t) = exp(-\int_{t_n}^{t} \sigma(\boldsymbol{r}(\boldsymbol{s}))ds) \tag{7.5}$$

where the symbols are labelled as

- $\boldsymbol{T(t)}$: Transmittance representing the probability that the ray travels from t_1 to t_2 without interception.
- $\boldsymbol{\sigma(r(t))}$: Volume density obtained the camera ray from each pixel.
- $\boldsymbol{c(r(t), d)}$: Colour along the camera ray with viewing direction $\boldsymbol{d}$.
- $\boldsymbol{dt}$: Differential distance the ray travels at each integration step.

This equation accumulates the colours defined along the ray to represent a 2D image from a 3D scene. The camera rays $C(r)$ are traced through each pixel to synthesise novel views. Fundamentally, the goal of 3D scene reconstruction using volume rendering is to view a scene directly in 3D space. However, computers are limited to 2D representations thus we use complex projections and mathematical transformations to translate a 3D object onto a 2D screen for viewing. To visualise a 3D scene, we carry out rendering from a specific viewpoint that is defined by the camera's intrinsic and extrinsic parameters. The intrinsic parameters relate to the camera's internal characteristics, such as focal length and lens distortion, which affect how the scene is projected onto the image plane. The extrinsic parameters, on the other hand, describe the camera's position and orientation in the world, determining the scene's perspective captured by the camera.

Given the world-to-camera transformations that occur to project a 3D scene onto a 2D image plane, rays are emitted at each pixel into the scene and sampled at different timesteps to determine the colour and the density occuring at the intersection of the ray and the surface. The discrete rendering equation defines the estimate of the volume rendering integral as a weighted average.

$$C(\boldsymbol{r}) = \sum_{i=1}^{N} T_i(1 - exp(-\sigma_i\delta_i))\boldsymbol{c}_i \quad \text{where } T_i = exp(-\sum_{j=1}^{i-1} \sigma_j\delta_j) \tag{7.6}$$

where $\delta_i = t_{i+1} - t_i$ is the distance between adjacent samples. t_i is drawn uniformly at random where the query points along the ray are partitioned into N evenly spaced bins using stratified sampling

$$t_i \sim \mathcal{U}[t_n + \frac{i-1}{N}(t_f - t_n), t_n + \frac{i}{N}(t_f - t_n)] \tag{7.7}$$

This approach allows the network to represent a continuous scene representation since the MLP is evaluated at continuous positions during optimisation.

Hierarchical Volume Sampling

Given a large number of points along a ray, i.e a large bin value, rendering will accumulate free space and occluded regions that do not contribute to the rendered image.

As a result, Mildenhall *et al.* [2020] increase rendering efficiency by allocating samples proportionally to their expected effect on the final rendering. Thus, hierarchical sampling allocates more samples to regions that are expected to contain visible content. First, they optimise a **coarse** network and sample a set of N_c locations using stratified sampling. This allows the network to find where the valuable regions are

$$\hat{C}_c(\boldsymbol{r}) = \sum_{i=1}^{N_c} w_i c_i, \quad w_i = T_i(1 - exp(-\sigma_i \delta_i)) \tag{7.8}$$

where the weights are normalised, $\hat{w} = \frac{w_i}{\sum_{j=1}^{N_c} w_j}$, to produce a piecewise-constant probability density function along the ray. Given the output of the **coarse** network, they use the valuable regions to sample more informative points N_f along each ray where samples are biased towards relevant parts of the volume. They sample a second set of N_f locations from the piecewise constant probability density function to evaluate the **fine** network at the unio of the first and second set of samples. Finally, they computed the rendered colour of the ray $\hat{C}_f(\boldsymbol{r})$ using all $N_c + N_f$ samples.

7.2 Inference Pipeline

The inference pipeline serves to allow the interaction between pretrained generative network paradigms to synthesise garments on novel views. The pretrained neural radiance field encodes a 3D scene representation using 2D-based images to synthesise novel views. Given the vector-valued function representing position (x, y, z) and direction (θ, ϕ), the neural radiance field samples the coordinates along camera rays and performs volume rendering to predict the emitted colour and density. A virtual try-on network is employed during inference to synthesise garments on each rendered novel view. In other words:

$$\begin{aligned} F_\theta([R|T]) \quad &\underset{\textbf{MLP}}{\Longrightarrow} \quad \{c, \sigma\} \\ VITON(\{c, \sigma\}, i_c) \quad &\Longrightarrow \quad \{I_c\} \end{aligned}$$

where

- $\mathbf{F}_\theta$: represents the neural radiance field parameterised by its weights θ.
- $\{c, \sigma\}$: The colour and density that is predicted by the neural radiance field for each novel view.
- $[R|T]$: represents the extrinsic parameters $\begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ of a camera at viewing directions $\{\theta, \phi\}$.
- i_c represents the in-shop clothing item.
- I_c represents the person wearing in-shop clothing item.

Firstly, we represent an individual's orientation within the scene as a continuous 5D function represented by a 3D scene. The continuous scene is constructed by extracting image frames from frontal-facing videos of a person. These videos were taken with a camera phone while participants stood on a rotating device. As explained in Chapter 4,

we use FFMPEG to extract individual frames from each video. Then we collect transformation poses using local light field fusions from each of the image frames (extracted from videos).

Thus, given 5D coordinates (location and viewing direction) and a clothing image $c \in \mathbb{R}^{3\times H\times W}$, our goal is to synthesise target clothing c onto a novel view of a person. For sub-task of generating a representing a continuous 3D scene using 2D images, we use Müller *et al.* [2022]'s instant neural graphic primitives to render novel views. For each novel view, we infer a parser free network to synthesise virtual try-on for each sequence of novel views reconstructing a 3D virtual try-on scene using 2D images.

Figure 7.5 illustrates the pipeline that leverages the **PoseCam** dataset and infers a parser-free composition network to generate a novel try-on view. Focusing on the qualitative performance of the novel try-on view highlights key areas of improvement in integrating virtual try-on networks with neural radiance fields for fitting clothing onto an individual. The fit and draping of the garment appear unnatural, with floating artifacts around the torso, while the seams and edges of the clothing show noticeable inconsistencies, undermining the overall realism. The texture quality is compromised, especially where the fabric should show natural bends and folds.

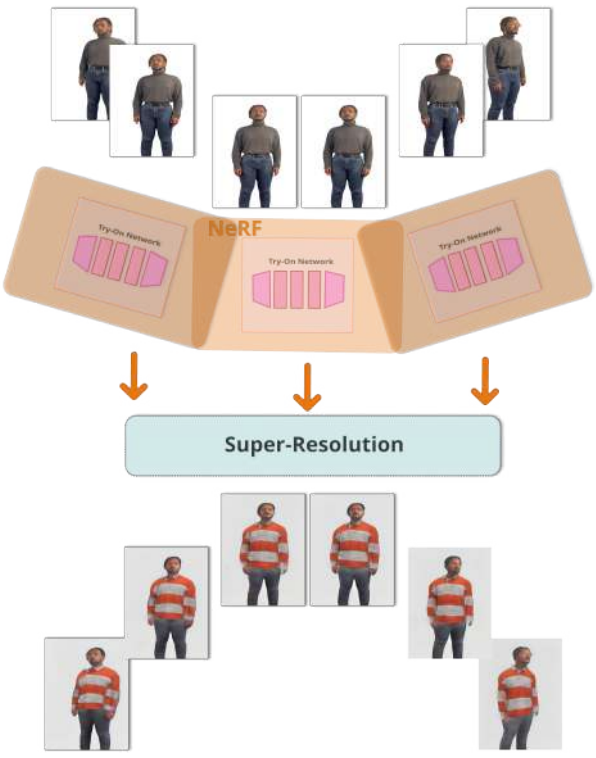


Figure 7.5: Here we demonstrate the goal of our research project, which is to infer that a pretrained neural radiance field can be integrated with virtual try-on to render novel try-on views. Super-resolution, using stable diffusion [Yue *et al.* 2023], is an optional integration but can be leveraged to upsample try-on views(256x192) to higher resolutions (1024x768). The pipeline leverages the **PoseCam** dataset and infers a parser-free composition network to generate a novel try-on view.

7.3 Results

The evaluation of the integrated system of virtual try-on and neural radiance fields (Nerf Studio) was structured around both paired and unpaired scenarios. This dual approach is crucial to understanding the adaptability and effectiveness of the virtual try-on technology when subjected to different levels of available data consistency.

We selected our evaluation criteria by testing a dataset composed of 10 different students. The inference pipeline serves to facilitate the interaction between pretrained generative network paradigms to synthesise garments on novel views. The pretrained neural radiance field encodes a 3D scene representation using 2D-based images to synthesise novel views. During inference, a virtual try-on network synthesises garments on each rendered novel view.

For the quantitative analysis, we employed a combination of paired and unpaired evaluation methods. In the paired evaluation, we focused on direct comparisons between the original clothing item images and their corresponding images from the video frames. Metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) were used to quantify the fidelity and visual similarity of the virtual try-on results compared to the original clothing items.

In contrast, the unpaired evaluation assessed the performance of the virtual try-on in scenarios where ground-truth images of the individual wearing the desired clothing item were not available. Instead, the generated images consist of the individual wearing a different clothing item from the ground-truth image. This approach tested the model's capability to generalise across different clothing items and poses not paired in the training data.

Evaluating virtual try-on systems in an unpaired setting presents significant challenges due to the inherent differences in clothing items. Variances in colour, logos, and embroidery can lead to large discrepancies in specific regions, making traditional metrics less effective. Given these challenges, traditional metrics like PSNR and SSIM are not always suitable [Cheng *et al.* 2021]. Therefore, we employed the Learned Perceptual Image Patch Similarity (LPIPS) [Zhang *et al.* 2018] metric to evaluate the perceptual similarity between the generated images and the ground truth.

In both paired and unpaired settings, the results were averaged across all students. Specifically, for each of the 10 students, approximately 100 images were taken from video frames for the task of 3D scene reconstruction. Our inference pipeline was applied individually to each student's images, and the evaluation metrics were recorded.

To illustrate, for a model like DM_VTON, we performed the evaluation separately for each of the 10 students, resulting in 10 distinct PSNR values. These individual PSNR values were then averaged to provide an overall performance measure for the model. This averaging process gives a more comprehensive understanding of how the method performs across different subjects, ensuring that the evaluation reflects varied scenarios and conditions. This approach ensured a robust assessment of the performance across different individuals and scenarios.

In Table 7.1, we analyse the paired setting and unpaired settings of virtual try-on methods CP_VTON+, HR_VITON, PF_AFN, DM_VTON, and FS_VTON based on their paired setting results for SSIM and PSNR metrics. These metrics provide insight into the image quality and structural similarity achieved by each method.

Table 7.1: Comparison of SSIM, PSNR, and LPIPS metrics for different virtual try-on methods under paired and unpaired settings (* refers to optimal performance).

	Paired			Unpaired		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
CP_VTON+[3]	18.61	0.68	0.36	-	-	0.42*
HR_VITON[1]	15.00	0.73*	0.40	-	-	0.50
PF_AFN[4]	19.09*	0.68	0.37	-	-	0.46
DM_VTON[6]	15.10	0.58	0.43	-	-	0.52
FS_VTON[5]	16.78	0.68	0.35*	-	-	0.48

CP_VTON+ achieves a PSNR of 18.61, an SSIM of 0.68 and an LPIPS of 0.36. This method utilises a two-stage process with a geometric matching module and a refinement module. The geometric matching ensures accurate alignment of the clothing item, while the refinement module integrates the clothing more seamlessly into the person's image. This results in good overall image quality and structural similarity, although the moderate SSIM value suggests that there is room for improvement in preserving fine details. The LPIPS score indicates that the perceptual quality is fairly good but not the best among the compared methods.

HR_VITON attains a PSNR of 15.00, an SSIM of 0.73 and an LPIPS of 0.40. This method focuses on generating high-resolution virtual try-on images, significantly enhancing visual quality and detail. The method employs a two-stage process involving a try-on condition generator and a try-on image generator. The try-on condition generator deforms the clothing and produces a segmentation map that helps align the clothing with the person's body. This process minimizes misalignment and pixel-squeezing artifacts, contributing to a higher SSIM value, indicating better preservation of structural details. Despite the high structural similarity, the lower PSNR value suggests that while HR-VITON excels at maintaining structural details, it may introduce more noise or artifacts, affecting overall image quality. The LPIPS score, reflecting perceptual quality, is reasonable but not the highest among methods, indicating that there is still room for improvement in terms of visual perception The LPIPS score reflects reasonable perceptual quality, though not the highest.

PF_AFN shows a PSNR of 19.09, an SSIM of 0.68 and an LPIPS of 0.37. This method employs a novel "teacher-tutor-student" knowledge distillation scheme to produce high-quality virtual try-on images without the need for human parsing. The method uses an appearance flow warping module (AFWM) to establish dense correspondences between the person and clothing images, which accurately aligns the clothing with the person's image. This approach results in high PSNR values, indicating superior image quality. However, the SSIM value, while still high, suggests that there is room for improvement in preserving fine structural details. The LPIPS score indicates good perceptual quality, slightly better than HR_VITON but not the best.

DM_VTON records a PSNR of 15.10, an SSIM of 0.58 and an LPIPS of 0.43. This method employs a knowledge distillation approach involving Teacher and Student networks. The Teacher network generates synthetic images to train the Student network, which focuses on faster inference and reduced memory usage. Despite this advanced training pipeline, the method's lower PSNR and SSIM values indicate challenges in preserving overall image quality and structural details. These issues might stem from difficulties in

achieving precise garment deformation and seamless integration without introducing artifacts. The higher LPIPS score further suggests that the perceptual quality is not as high compared to other methods The higher LPIPS score suggests lower perceptual quality compared to the other methods.

FS_VTON achieves a PSNR of 16.78, an SSIM of 0.68 and LPIPS of 0.35. This method uses a few-shot learning paradigm, enhancing generalisation with limited examples. The method performs reasonably well in terms of both PSNR and SSIM, indicating a good balance between image quality and structural preservation. The few-shot learning approach likely helps in maintaining robustness across different clothing styles and body types, resulting in competitive performance. The lowest LPIPS score among the methods indicates the best perceptual quality in the paired setting.

In conclusion, PF_AFN stands out with the highest PSNR, indicating excellent image quality, likely due to its pose-free adaptive approach. HR_VITON excels in SSIM, suggesting superior structural detail preservation, attributed to its high-resolution refinement. CP_VTON+ and FS_VTON offer balanced performance with good image quality and structural similarity. DM_VTON lags in both metrics, indicating potential issues with disentangling and re-integrating clothing features without artifacts. These performance differences highlight the strengths and trade-offs of each method's architectural choices and their impact on virtual try-on quality.

While these quantitative results provide valuable insights, it is essential for future work to incorporate user studies to capture subjective perceptions of image quality and realism, which quantitative metrics alone cannot fully gauge. User feedback will provide a more holistic understanding of virtual try-on performance. Additionally, the unpaired evaluation of virtual try-on remains an active area of research, emphasising the need for developing methods to assess performance in unpaired settings to address practical use cases.

Another promising direction for future research is to evaluate 3D scene reconstruction and 2D virtual try-on in unison, rather than assessing them separately. By pretraining 3D scene reconstruction methods, such as neural radiance fields, and integrating virtual try-on, this combined approach leverages the strengths of both techniques. However, combining two networks created for different purposes introduces new evaluation challenges, as traditional metrics may not fully capture the integrated system's performance. Future work should develop new evaluation methods that balance the dual objectives of synthesising realistic clothing images and accurately reconstructing 3D scenes from 2D images.

This research expands the applicability of virtual try-on by leveraging neural radiance fields to reconstruct 3D virtual scene from 2D images. By integrating these networks, we are setting the groundwork for future explorations into more immersive and realistic virtual try-on applications. The dual evaluation approach, considering both paired and unpaired scenarios, lays a comprehensive foundation for assessing the robustness and flexibility of these integrated systems in virtual try-on research.

7.4 Summary

This pilot study investigates the reconstruction of a 3D virtual try-on scene using the **PoseCam** dataset, adopting a comprehensive approach that incorporates Multi-View

Stereo (MVS) and Neural Radiance Fields (NeRFs). In subsection 7.1.1, we discuss Multi-View Stereo (MVS), a sophisticated computer vision technique designed to reconstruct 3D models from a series of images captured from various viewpoints. This method processes overlapping images and employs feature extraction methods such as SIFT and SURF to identify and align distinctive landmarks across images, creating depth maps that form a comprehensive 3D model. The model is further refined with a mesh and textures, enhancing its realism and detail.

Structure-from-Motion (SfM), discussed alongside MVS in subsection 7.1.2, outlines the scene's initial structure based on camera movements, which MVS then enriches with detailed 3D information. Applications like COLMAP streamline this process by automating camera calibration and point cloud reconstruction, facilitating the creation of high-quality 3D models from minimal image sets.

In subsection 7.1.3, we introduce Neural Radiance Fields (NeRFs), a method for implicit scene representation. NeRFs transform input images and corresponding camera parameters into detailed 3D scenes, capturing view-dependent colours and view-independent densities to render realistic textures and appearance effects. The model synthesises novel views by predicting RGB colours and volume densities using high-frequency positional encoding.

Our inference pipeline, detailed in Section 7.2, integrates pretrained generative network paradigms to synthesise garments on novel views. By encoding a 3D scene with NeRFs and employing virtual try-on networks, the system predicts and renders clothing onto individuals in the reconstructed 3D scene. This process includes extracting image frames from videos, collecting transformation poses, and synthesising target clothing onto novel views.

Lastly, in Section 7.3 we provide a quantitative evaluation of the integrated system, involving both paired and unpaired scenarios. Paired evaluation compares the original clothing item images with the generated virtual try-on results using metrics such as PSNR and SSIM. Unpaired evaluation assesses the model's performance when ground-truth images of the individual wearing the desired clothing item are unavailable, employing the Learned Perceptual Image Patch Similarity (LPIPS) metric to evaluate perceptual similarity.

The study demonstrates the adaptability and effectiveness of virtual try-on, highlighting its potential in various real-world applications. By integrating MVS, SfM, and NeRFs, this research lays the groundwork for future explorations into more immersive and realistic virtual try-on experiences.

As we transition to the concluding chapter of this study, Chapter 8 summarises the key insights gained from this study, emphasising how the PoseCam dataset has proven instrumental in advancing 3D scene reconstruction. We will discuss the broader implications of our work for the virtual try-on research, particularly how our approach can contribute to more immersive and realistic experiences. This concluding chapter will encapsulate the journey of our research, reaffirming the value of our dataset and methods, and setting the stage for future explorations in the realm of virtual try-on and fashion in computer vision.

Chapter 8

Conclusion And Future Work

8.1 Conclusion

Our research details the development of our novel **PoseCam** dataset that overcomes existing dataset limitations to broaden the applicability of virtual try-on. We utilised state-of-the-art techniques, such as segmentation algorithms and part-grouping networks, to extract detailed semantic information, and DensePose to provide essential 3D pose data, collectively enriching our dataset in a single integrated process.

The incidental facilitation of 3D scene reconstruction through our dataset highlights its potential to expand virtual try-on capabilities beyond traditional front-facing camera angles, thereby offering a more comprehensive application within digital fashion scenarios.

Looking ahead, we envisage that our contributions will significantly enhance the development of more sophisticated virtual try-on systems, capable of synthesising garments across varied viewing angles and seamlessly integrating with emergent technologies like mixed reality. These advancements aim to utilise the outputs of our research to offer more dynamic and immersive user experiences.

In conclusion, the introduction of the **PoseCam** dataset marks a significant enhancement in the realm of digital fashion, addressing critical challenges of realism, adaptability, and perspective diversity in clothing synthesis. Through an exhaustive exploration of parser-based and parser-free networks, alongside the strategic use of neural radiance fields (NeRFs), our research has substantially advanced the state of virtual try-on. By enabling more sophisticated training on a diverse dataset, we have improved the accuracy of garment warping and facilitated the synthesis of clothing across diverse viewing angles, thus broadening the scope of virtual try-on to better reflect real-world conditions.

8.2 Future Work

The analysis of synthesising warped cloth predictions has highlighted several avenues for future research. Addressing these challenges will require a multifaceted approach, focusing on the integration of three-dimensional data, occlusion handling, attention mechanisms, and the use of synthetic training data.

Formal Qualitative Human Evaluation

While our study has provided a comprehensive evaluation of virtual try-on models using a multi-faceted approach, incorporating both quantitative metrics and qualitative assessments, we acknowledge the absence of a formal human evaluation in our results. The qualitative evaluations presented in this study were conducted through internal assessments, focusing on the perceptual and practical aspects of model performance as interpreted by expert reviewers. We recognise that a formal human evaluation, involving participants who rate the outputs or identify error types, would provide valuable insights into the subjective quality and user experience of the virtual try-on models. Such an evaluation would enhance the robustness of our findings and offer a more comprehensive understanding of the models' performance in real-world scenarios. To address this, we propose the following directions for future research:

- **Formal Human Evaluation** [Song *et al.* 2023]: Implementing a structured human evaluation study where a diverse group of participants rate the virtual try-on outputs based on various criteria, such as fit accuracy, visual realism, and overall satisfaction. This study would involve detailed protocols to ensure unbiased and representative feedback.
- **Error Type Identification**: Conducting user studies to identify specific types of errors in the virtual try-on results, providing deeper insights into the models' weaknesses and areas for improvement.
- **User Experience Studies**: Exploring the broader user experience by assessing how users interact with the virtual try-on system, their preferences, and their satisfaction levels over prolonged use.

In conclusion, while our current study provides a solid foundation, integrating formal human evaluations in future research will significantly enhance the depth and applicability of our findings, aligning with contemporary best practices in model evaluation and user-centred design.

3D Data Integration and Pose Estimation

To achieve a more realistic rendering of garments, future warping models should integrate a nuanced understanding of the 3D structure of clothing and the human body. Utilising 3D convolutional networks to process three-dimensional data could significantly enhance the model's ability to comprehend garment dynamics and rotations. Additionally, improved 3D pose estimation techniques could provide deeper insights into the relative positioning of the human skeleton to the camera, which is vital for accurately handling cases involving pronounced rotation.

Occlusion-Aware Mechanisms

An occlusion-aware mechanism is essential for predicting and visualising the portions of clothing that are not immediately visible. The incorporation of synthetic training data, generated with 3D body models, could provide comprehensive garment views and train the model to manage occlusions more effectively. High-resolution images depicting diverse clothing types in various poses and rotations are also critical for the continued enhancement of dataset quality.

Attention Mechanisms in Warping Architectures

Implementing attention mechanisms could enable warping architectures to focus on intricate garment details and generate unseen parts, thus improving warping accuracy. Leveraging 3D body scanning technologies to create detailed datasets containing depth information would be beneficial. These datasets could include varying degrees of rotation and occlusion, aiding in the development of more sophisticated virtual try-on systems.

Synthetic Data and Deep Learning Techniques

The use of synthetic data to simulate fabrics and garments in a virtual environment could address the variety of occlusion scenarios needed for comprehensive training. Such data, rich in detail and variety, would allow models to learn from complex situations that are challenging to capture in real-life datasets. Further, adapting deep learning techniques like Generative Adversarial Networks (GANs) could help in generating the unseen parts of garments. When combined with occlusion-aware loss functions, these networks could be trained to focus on the continuity of fabric and accurately predict occlusions caused by rotation.

Comprehensive Development of Virtual Try-On Systems

The ultimate objective is to develop a warping model that can accurately predict garment appearance on a person from any viewpoint and in any pose. This ambition necessitates a collaborative effort in dataset development, model architecture design, and advanced training methodologies. With the evolution of these technologies, we can anticipate the emergence of highly realistic and accurate virtual try-on systems that enhance the user experience.

References

[Barratt and Sharma 2018] Shane T. Barratt and Rishi Sharma. A note on the inception score. *ArXiv*, abs/1801.01973, 2018.

[Bookstein 1989] 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.

[Cao *et al.* 2021] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(01):172–186, 2021.

[Chen *et al.* 2024a] Haodong Chen, Yongle Huang, Haojian Huang, Xiangsheng Ge, and Dian Shao. Gaussianvton: 3d human virtual try-on via multi-stage gaussian splatting editing with image prompting. 2024.

[Chen *et al.* 2024b] Mengting Chen, Xi Chen, Zhonghua Zhai, Chen Ju, Xuewen Hong, Jinsong Lan, and Shuai Xiao. Wear-any-way: Manipulable virtual try-on via sparse correspondence alignment. 2024.

[Cheng *et al.* 2021] Wen-Huang Cheng, Sijie Song, Chieh-Yun Chen, Shintami Chusnul Hidayati, and Jiaying Liu. Fashion meets computer vision: A survey. *ACM Comput. Surv.*, 54(4), July 2021.

[Choi *et al.* 2021] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14131–14140, June 2021.

[Cui *et al.* 2023] Aiyu Cui, Sen He, Tao Xiang, and Antoine Toisoul. Learning garment densepose for robust warping in virtual try-on. 2023.

[Dam *et al.* 2024] Phuong Dam, Jihoon Jeong, Anh Tran, and Daeyoung Kim. Time-efficient and identity-consistent virtual try-on using a variant of altered diffusion models. 2024.

[Dong *et al.* 2019] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[Ge *et al.* 2021] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8485–8493, June 2021.

[Glorot and Bengio 2010] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[Gong *et al.* 2018] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[Goodfellow *et al.* 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[Guan *et al.* 2012] Peng Guan, Loretta Reiss, David A. Hirshberg, Alexander Weiss, and Michael J. Black. Drape: Dressing any person. *ACM Trans. Graph.*, 31(4), July 2012.

[Gupta *et al.* 2024] Aayush Gupta, Aditya Gulati, Himanshu, and Lakshya LNU. Human shape and clothing estimation. 2024.

[Güler *et al.* 2018] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[Han *et al.* 2018] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[He *et al.* 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[He *et al.* 2022] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3470–3479, June 2022.

[Hornik *et al.* 1989] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5):359–366, July 1989.

[Hui *et al.* 2018] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[Ilg *et al.* 2017] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estima-

tion with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[Isola *et al.* 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[Jabbar *et al.* 2021] Abdul Jabbar, Xi Li, and Bourahla Omar. A survey on generative adversarial networks: Variants, applications, and training. *ACM Comput. Surv.*, 54(8), oct 2021.

[Jaderberg *et al.* 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[Johnson *et al.* 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. volume 9906, pages 694–711, 10 2016.

[Kalantidis *et al.* 2013] Yannis Kalantidis, Lyndon Kennedy, and Li-Jia 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*, ICMR '13, page 105–112, New York, NY, USA, 2013. Association for Computing Machinery.

[Karras *et al.* 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[Kirillov *et al.* 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, October 2023.

[LeCun and Bengio 1998] Yann LeCun and Yoshua Bengio. *Convolutional Networks for Images, Speech, and Time Series*, page 255–258. MIT Press, Cambridge, MA, USA, 1998.

[Lecun *et al.* 1998] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Lee *et al.* 2022] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pages 204–219. Springer Nature Switzerland, 2022.

[Lin and Lucey 2017] Chen-Hsuan Lin and Simon Lucey. Inverse compositional spatial transformer networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[Liu *et al.* 2016a] Kuan-Hsien Liu, Ting-Yen Chen, and Chu-Song Chen. Mvc: A dataset for view-invariant clothing retrieval and attribute prediction. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, page 313–316, New York, NY, USA, 2016. Association for Computing Machinery.

[Liu *et al.* 2016b] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, 2016.

[Lombaert 2002] Herve Lombaert. Manual registration with thin plates. 2002.

[Long *et al.* 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[Mao *et al.* 2017] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[Massip *et al.* 2018] Eric Massip, Shintami Chusnul Hidayati, Wen-Huang Cheng, and Kai-Lung Hua. Exploiting category-specific information for image popularity prediction in social media. In *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 45–46, 2018.

[Matt Tancik 2022] Ben Mildenhall Matt Tancik. Nerf tutorial eccv 2022. 2022.

[Mildenhall *et al.* 2019] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: practical view synthesis with prescriptive sampling guidelines. *ACM Trans. Graph.*, 38(4), July 2019.

[Mildenhall *et al.* 2020] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421, Cham, 2020. Springer International Publishing.

[Minar *et al.* 2020] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[Morelli *et al.* 2022] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2231–2235, June 2022.

[Müller *et al.* 2022] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022.

[Nguyen-Ngoc *et al.* 2023] K. Nguyen-Ngoc, T. Phan-Nguyen, K. Le, T. V. Nguyen, M. Tran, and T. Le. Dm-vton: Distilled mobile real-time virtual try-on. In *2023*

IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), pages 695–700, Los Alamitos, CA, USA, oct 2023. IEEE Computer Society.

[Park *et al.* 2019] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[Pons-Moll *et al.* 2017] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J. Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Trans. Graph.*, 36(4), July 2017.

[Raj *et al.* 2018] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. Swapnet: Image based garment transfer. In *European Conference on Computer Vision*, 2018.

[Ranjan and Black 2017] Anurag Ranjan and Michael J. Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[Ronneberger *et al.* 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015.

[Ross 2022] Kassandra Ross. *Artificial Intelligence in Fashion Manufacturing: From Factory Operation to Advisory Role*, pages 95–116. Springer International Publishing, Cham, 2022.

[Sandler *et al.* 2018] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[Santesteban *et al.* 2022] Igor Santesteban, Miguel Otaduy, Nils Thuerey, and Dan Casas. Ulnef: Untangled layered neural fields for mix-and-match virtual try-on. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 12110–12125. Curran Associates, Inc., 2022.

[Schönberger and Frahm 2016] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[Schönberger *et al.* 2016] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[Shim *et al.* 2023] Sang-Heon Shim, Jiwoo Chung, and Jae-Pil Heo. Towards squeezing-averse virtual try-on via sequential deformation. 2023.

[Song *et al.* 2023] Dan Song, Xuanpu Zhang, Juan Zhou, Weizhi Nie, Ruofeng Tong, Mohan Kankanhalli, and An-An Liu. Image-based virtual try-on: A survey. 2023.

[Wang *et al.* 2004] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[Wang *et al.* 2018a] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. 2018.

[Wang *et al.* 2018b] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[Wolfe 2019] Cameron Wolfe. Beyond nerfs: Part one. 2019.

[Yagel *et al.* 1992] Roni Yagel, Daniel Cohen, and Arie Kaufman. Discrete ray tracing. *IEEE Comput. Graph. Appl.*, 12(5):19–28, sep 1992.

[Yamaguchi *et al.* 2012] Kota Yamaguchi, M. Hadi Kiapour, Luis E. Ortiz, and Tamara L. Berg. Parsing clothing in fashion photographs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3570–3577, 2012.

[Yue *et al.* 2023] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. 2023.

[Zhang *et al.* 2018] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume abs/1801.03924, 2018.

[Zhang *et al.* 2024] Jeffrey Zhang, Kedan Li, Shao-Yu Chang, and David Forsyth. Acdg-vton: Accurate and contained diffusion generation for virtual try-on. 2024.

[Zhu *et al.* 2020] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. 2020.

Appendix A

Appendix

This appendix delineates the methodological framework and constraints underpinning our investigation into the integration of neural radiance fields within virtual try-on systems. The training of our networks was subject to several constraints. Notably, the computation was performed on the school's shared cluster, a resource that, while powerful, is not dedicated solely to this project. Additionally, the training process was periodically interrupted by load shedding, further complicating our computational efforts. Given the substantial computational demands of vision models, which often necessitate days of continuous training, these interruptions were non-trivial.

Our research specifically concentrated on the feasibility of integrating neural radiance fields with virtual try-on technologies. Consequently, our experimental focus was narrowed to hyperparameter sweeps solely within the domain of warping architectures. The selection of hyperparameters for these architectures adhered to two principal guidelines. Initially, values were directly adopted from those suggested within the literature referenced in our study, ensuring a foundation in previously validated research. Subsequently, Bayesian optimization techniques were employed to refine these selections, aiming to identify the optimal hyperparameter values within the context of our specific investigative focus.

The hyperparameter sweeps were conducted using the weights and biases library, which provides a modular and extensible platform for optimizing experimental parameters. This library was selected for its compatibility with PyTorch, the deep learning framework used in our research. The hyperparameter sweeps were conducted on the **PoseCam** dataset, which was used to train the virtual try-on networks. The dataset was divided into training and validation sets, with the training set comprising 60% of the data and the validation set the remaining 20%. The validation set was used to evaluate the generalisability of the trained models to unseen data. The hyperparameter sweeps were conducted to investigate the effect of varying coefficients of loss terms for each of the following architectures: CP_VTON, CP_VTON+, HR_VITON, PF_AFN, and FS_VTON. The results of these sweeps are presented in the following figures.

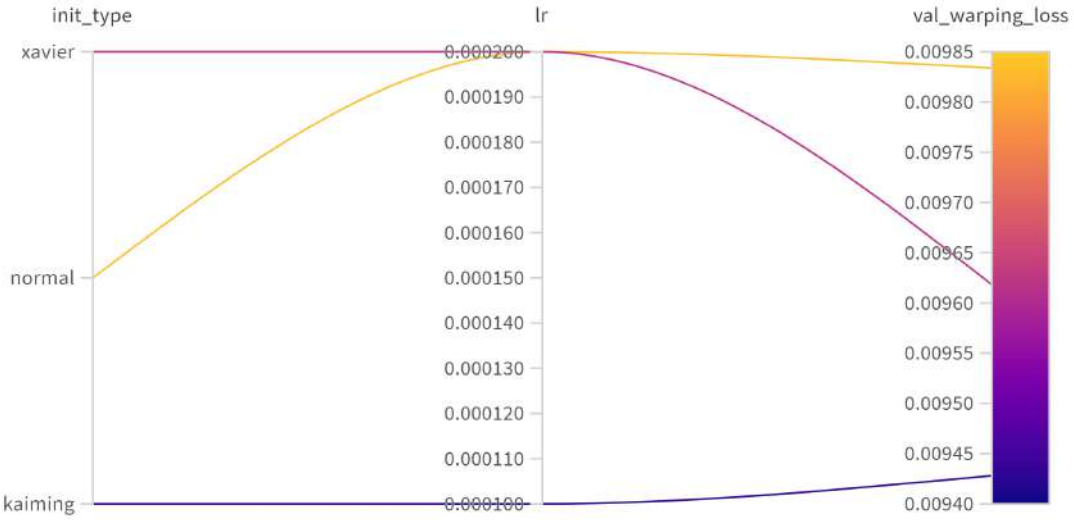


Figure A.1: In this sweep, CP_VTON[2] varied the learning rate of the warping architecture and the initialisation technique between xavier, kaiming and normal. The figure illustrates that selecting the learning rate of 0.0001 and the kaiming initialisation technique resulted in the lowest validation warping loss. The validation warping loss reflects the model's ability to generalise the warping of clothing articles to new data not seen during training.

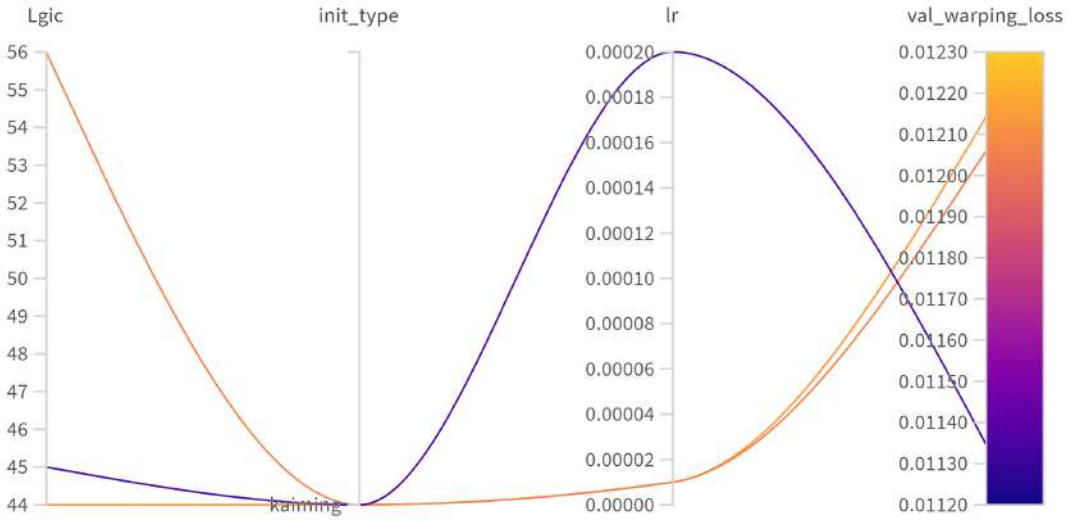


Figure A.2: In this sweep, CP_VTON+[3] varied same parameters for CP_VTON with an added λ_{reg} hyperparameter that alters the effect of the loss function [Minar *et al.* 2020].

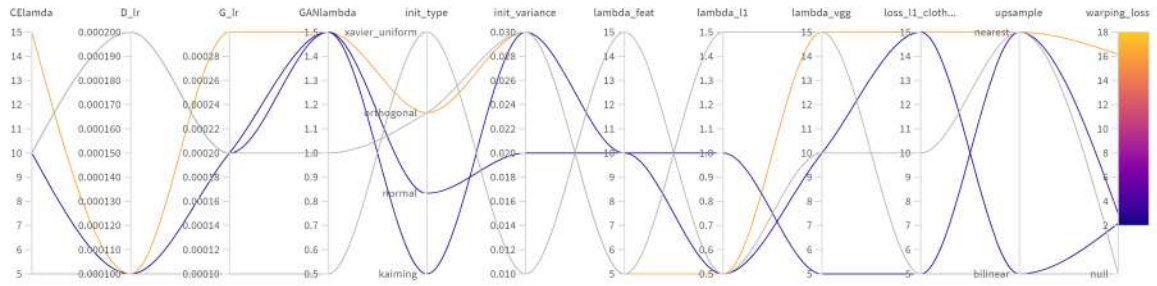


Figure A.3: In this sweep, HR_VITON[1] varied the loss coefficients for the warping architecture. The figure illustrates the combination of loss coefficients that resulted in the lowest validation warping loss.

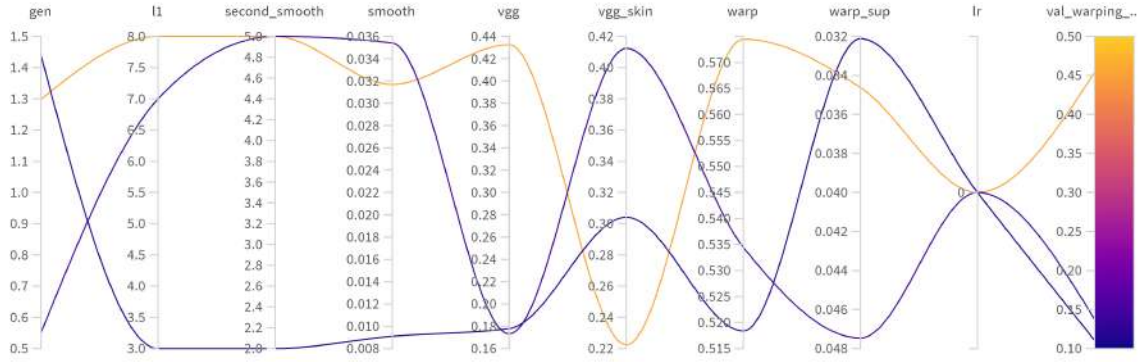


Figure A.4: In this sweep, PF_AFN[5] varied the loss coefficients for the warping architecture. The figure illustrates the combination of loss coefficients that resulted in the lowest validation warping loss.

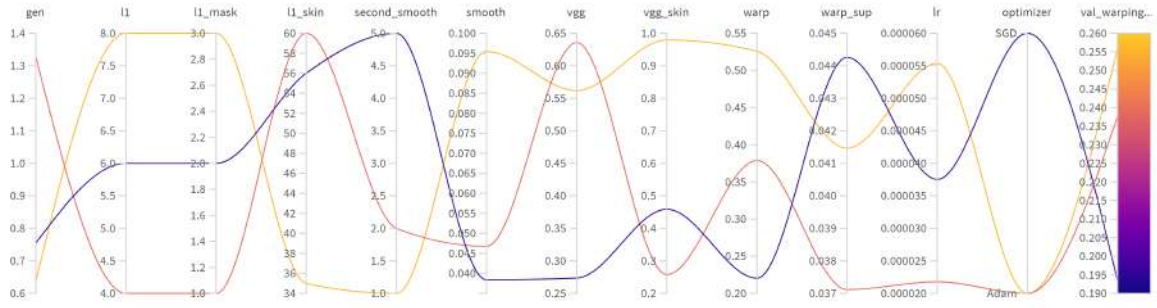


Figure A.5: In this sweep, FS_VTON[6] varied the loss coefficients for the warping architecture. The figure illustrates the combination of loss coefficients that resulted in the lowest validation warping loss.

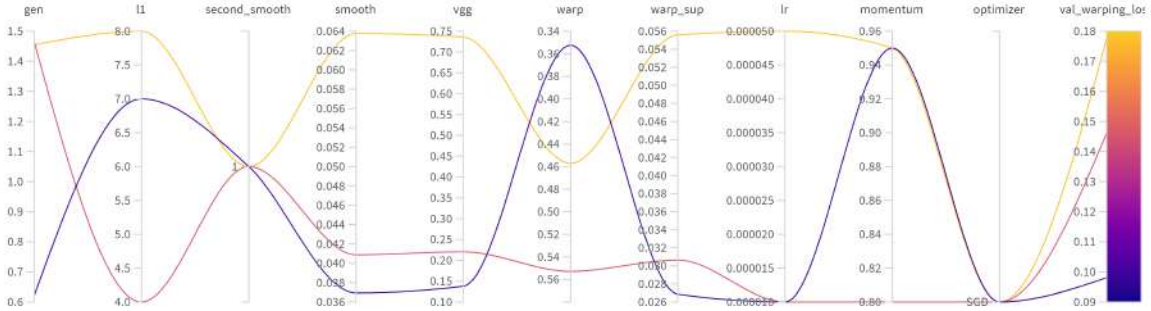


Figure A.6: In this sweep, DM_VTON[7] varied the loss coefficients for the warping architecture. The figure illustrates the combination of loss coefficients that resulted in the lowest validation warping loss.