

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

INTELLIGENT BODY-AWARE 2D VIRTUAL
TRY-ON AND TILEABLE TEXTURE GENERATIONS
FOR 3D VISUAL TRY-ON APPLICATIONS

HE HONGHONG

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

School of Fashion and Textiles

Intelligent Body-aware 2D Virtual Try-on and
Tileable Texture Generations for 3D Visual Try-on
Applications

HE Honghong

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

August 2024

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

(Signed)

HE Honghong

(Name of student)

Abstract

The worldwide reach of e-commerce enables individuals and businesses to buy and sell products online with ease at anytime and anywhere. Nevertheless, there are still specific difficulties or consumer concerns that hinge online purchases of fashion, largely because clothing products, different from other typical products, are soft and depend on wearers' body forms. More specifically, these concerns include clothing fitting, shape and size, the compatibility with other fashion items, how well the clothing matches the wearer's skin tone, and so forth. These concerns are reflected at the sky-high rates of return for clothing products shopped online. High returns generate a lot of carbon emissions that pollute the environment. To improve customer experience while promote sustainable development of the fashion industry, virtual try-on (VTON) technology offers a viable solution to these problems.

This thesis explores VTON methods for fashion applications, allowing users to preview garments before making a purchase decision. VTON methods can be generally classified into 2D-VTON and 3D-VTON methods. On one hand, 2D VTON methods are *image-based* virtual try-on schemes that utilize deep generative models to visualize garments on human images. Existing 2D-VTON methods learn pixel-wise cross-domain transformations to synthesize fashion images, yet they fail to consider the relationship between garment size and user body shape. A good VTON should not only provide excellent visual representation, but also help individuals evaluate the fit, shape,

and size of garments on wearers' bodies. With this motivation, a shape-faithful virtual try-on method is developed in this study based on body size dependent fashion landmark transformation (called BSLT-VTON). The proposed BSLT-VTON transfers wearable fit and wearing position between potential wearers and related garments with two key developments: a fashion landmark localization model (*i*) and body-aware garment deformation model (*ii*). Given the unavailability of massive fashion data with special body shapes, it is difficult to learn the relationship between different human body shapes and clothing sizes from images. Therefore, a fashion landmark localization model is proposed in this study (*Model i*) based on pose-aware segmentation to obtain fashion masks, and this will combine with human joints from a pose estimation method, both are inputs to locate the key functional points on the clothing. With this *Model i*, the relationship between different human body shapes and functional areas of clothing is obtained. Next, the body-aware clothing deformation model aiming to deform and scale clothing according to different body shapes (*Model ii*) is developed. This is a 2-step small incremental clothing deformation strategy to prevent excessive deformation of clothing during fitting. The effectiveness of the proposed method has been demonstrated in terms of the impact of clothing height and length (i.e., overall fit) on people of different body shapes and heights/proportions. Even for complex texture patterns, the proposed BSLT-VTON can provide pleasing visual effect, better than other state-of-the-art methods.

On the other hand, 3D VTON methods offer try-on experience to users based on 3D

garment models. In this study, a high-quality texture generation method (*Model iii*) and a new cloth digitization pipeline (*Model iv*) are proposed for 3D VTON. Most existing work for deep learning-based 3D VTON focus on reconstructing 3D garments from images, while the texture or texture quality of 3D garments are largely ignored, resulting in 3D garments with low quality of texture not suitable for real try-on fashion applications. The proposed texture reconstruction method combines the latest advances in deep texture synthesis, adversarial neural networks, and latent space manipulation, enabling the synthesis of high-quality, tileable textures for 3D garments, ready for Augment Reality 3D VTON fashion applications.

In summary, this study proposed 2 new VTON methods including both 2D and 3D approaches. Extensive experiments have shown that, comparing to the state-of-the-art 2D VTON methods, the new 2D VTON method (BSLT-VTON), integrating *Models i and ii*, exhibits realistic visualization try-on results, in terms of clothing patterns, textures, and colors, while allowing users to judge the realistic try-on effect of the clothing on users' bodies, in terms of clothing fit and shape. Comparing to the existing 3D VTON solutions, the new 3D VTON method, namely cloth digitization pipeline (*Model iv*) can easily and flexibly map high-quality, tileable textures generated (by *Model iii*) to 3D garments. Not only textures are generated, tileable albedo and normal maps are generated simultaneously for more realistic 3D visual try-on by 3D garment rendering.

Keywords: Fashion landmark localization; 2D virtual try-on; generative adversarial

network; tileable texture generation; 3D visual try-on; Augment Reality.

Publications and Awards

List of Publications:

1. **He, H.**, Zhou, Y., Sun, Z., Fan, J., & Mok, P. Y. (2024, January). Human Skeleton-aware Virtual Try-on via Fashion Landmarks and Garment Deformation. *In International Textile and Apparel Association Annual Conference Proceedings (Vol. 80, No. 1)*. Iowa State University Digital Press.
2. Ding, Y., **He, H.**, & Mok, P. Y. (2024). Personalizing human avatars based on realistic 3D facial reconstruction. *Multimedia Tools and Applications*, 1-28.
3. **He, H.**, Sun, Z., Fan J & Mok, P. Y. (2024). TiPGAN: Synthesising High Quality Tileable Textures with Intrinsic Priors for Cloth Digitization. *Computer-Aided Design*. (Under Review).
4. **He, H.**, Sun, Z., Fan J & Mok, P. Y. (2024). A Shape-faithful Virtual Try-On Method Based on Body Size Dependent Fashion Landmark Transformation. *Computers & Graphics*. (Under review)
5. **He, H.**, Sun, Z., Fan, J., & Mok, P. Y. (2023). SP3F-GAN: GENERATING SEAMLESS TEXTURE MAPS FOR FASHION. *IADIS International Journal on Computer Science & Information Systems*, 18(2).
6. Sun, Z., Zhou, Y., **He, H.**, & Mok, P. Y. (2023, October). Sgdiff: A style guided diffusion model for fashion synthesis. *In Proceedings of the 31st ACM International Conference on Multimedia (pp. 8433-8442)*.

7. **He, H.**, Zhou, Y., Fan, J., & Mok, P. Y., Pose-aware segmentation-based landmarks localization for fashion, accepted for 2022 International Conference on Clothing and Textiles (oral presentation), May 26-27, Seoul, Korea.
8. **He, H.**, Zhou, Y., Fan, J., & Mok, P. Y., Detection of Fashion Landmarks Based on Human Paring and Pose Estimation, 16th International Conference on Computer Graphics, Visualization, Computer Vision and Image Processing, 2022.

List of Awards

- **He, H.**, Zhou, Y., Fan, J., & Mok, P. Y., Detection of Fashion Landmarks Based on Human Paring and Pose Estimation, 16th International Conference on Computer Graphics, Visualization, Computer Vision and Image Processing, 2022. **Best paper award.**
- **He, H.**, Sun, Z., Fan, J., & Mok, P. Y. (2023). SP3F-GAN: GENERATING SEAMLESS TEXTURE MAPS FOR FASHION. *IADIS International Journal on Computer Science & Information Systems*, 18(2). **Best paper award.**

Acknowledgements

Firstly, I wish to convey my profound appreciation to my supervisor, Prof. Tracy P. Y. Mok, and my co-supervisor Prof. Fan Jintu, for acknowledging my capabilities and extending to me the opportunity to pursue an PhD degree at The Hong Kong Polytechnic University (PolyU). Her unwavering assistance throughout the last 3 years, enabling me to delve into diverse areas and providing guidance in academic endeavors, has been immeasurable. Any advancements I achieve in my academic journey will undoubtedly reflect her substantial impact.

Secondly, I would like to sincerely thank Dr. Yanghong Zhou and Ms. Yangping Zhou for many rich academic and technical discussions with me. I would also like to thank Dr. Xin Zhang, Dr. Hong Qu, Mr. Zhengwentai Sun and colleagues in rooms ST604, ST605 and W203, who have given me great help in my work and life.

Then, I sincerely thank my family for their strong support and respect for me. Without this premise, I would not be able to engage in the scientific research that I am interested in without worries. I also thank Dr. Jia for his encouragement and help. He accompanied me all the way, from Shanghai to Shenzhen, to the United Kingdom, and then to Hong Kong. When I faced challenges and difficulties, he was always the first to give me patient comfort and selfless help.

Finally, I would like to thank PolyU and the Laboratory for Artificial Intelligence in Design (AiDLab) for the financial support of my PhD research study and the city of HK for allowing me to broaden my horizons and see the diversity of the world.

Table of Contents

Abstract.....	i
Publications and Awards	v
Acknowledgements.....	vii
Table of Contents.....	ix
List of Figures	xiii
List of Tables.....	xviii
Chapter 1. Introduction	1
1.1 <i>Research Backgrounds.....</i>	1
1.1.1 State-of-the-art of 2D Virtual Try-on Technology (2D VTON) ...	3
1.1.2 State-of-the-art of 3D Virtual Try-on Technology (3D VTON) ...	5
1.2 <i>Statements of the Problem.....</i>	7
1.2.1 Limitations of 2D virtual try-on.	7
1.2.2 Limitations of texture generation for 3D VTON.....	8
1.3 <i>Research Aim and Objectives.....</i>	9
1.4 <i>Methodology Overview</i>	10
1.5 <i>Organization of the Thesis</i>	13
Chapter 2. Literature Review.....	15
2.1 <i>Overview of Virtual Try-on Technology.....</i>	15
2.1.1 2D Virtual Try-on Technology.....	15
2.1.1.1 2D VTON based on reference image and target clothing without 	

2.2.2	Pose Estimation	30
2.2.3	Fashion Landmark Detection	32
2.2.3.1	<i>Regression Based Models</i>	34
2.2.3.2	<i>Pose Estimation Models</i>	35
2.2.3.3	<i>Attentive Models</i>	36
2.2.3.4	<i>Constraints Based Methods</i>	37
2.2.4	Image deformation.....	38
2.3	<i>Generative Adversarial Networks</i>	42
2.3.1	Overview of GAN	43
2.3.2	Optimality of GAN.....	45
2.3.3	GAN for Generated Image/Texture Quality	49
2.4	<i>Augmented Reality</i>	51
2.4.1	Development of Augmented Reality	51
2.4.2	Augmented Reality application for fashion industry.....	53
2.5	<i>Chapter Summary</i>	56
Chapter 3.	Fashion Landmark Localization	59
3.1	<i>Introduction of fashion landmark</i>	59
3.2	<i>Related Work</i>	62
3.3	<i>Method</i>	65
3.3.1	Initial Mask Extraction	65
3.3.2	Clothing Mask Completion	68
3.3.3	Coarse Clothing Key-points Sampling	72
3.3.4	Clothing Key-points Refinement.....	74
3.4	<i>Experiments</i>	78
3.4.1	Datasets and Implementation Details	79
3.4.2	Evaluation metrics	79
3.4.3	Results and Discussions	80
3.5	<i>Limitation</i>	83
3.5.1	Limitations of dataset selection	83

3.5.2	Failure Cases Discussions	84
3.6	Chapter Summary.....	85
Chapter 4.	Body-aware 2D Virtual Try-on	86
4.1	Introduction of 2D virtual try-on	86
4.2	Related Work.....	90
4.2.1	Virtual Try-on	90
4.2.2	Fashion landmark detection.....	93
4.2.3	Clothing deformation.....	95
4.3	Method	96
4.3.1	Mathematical notation	96
4.3.2	Human body-aware garment deformation model	100
4.3.2.1	<i>Definition the size of garment and human body</i>	<i>100</i>
4.3.2.2	<i>Body-aware try-on garment landmark relocation</i>	<i>104</i>
4.3.2.3	<i>Garment deformation via fashion landmark.....</i>	<i>106</i>
4.3.3	The virtual try-on synthesis	110
4.4	Experiments	111
4.4.1	Experimental Setup	111
4.4.2	Universality	112
4.4.2.1	<i>The Try-on results for different body shapes.....</i>	<i>112</i>
4.4.2.2	<i>Try-on results for different garment types.....</i>	<i>114</i>
4.4.3	Ablation study	115
4.4.3.1	<i>Overview of try-on influences</i>	<i>115</i>
4.4.3.2	<i>The influence of fashion landmarks on garment deformation ..</i>	<i>116</i>
4.4.3.3	<i>The deformation scales of different deformation algorithms....</i>	<i>118</i>
4.4.4	Experimental Results.....	119
4.4.4.1	<i>Qualitative results</i>	<i>119</i>
4.4.4.2	<i>Quantitative results</i>	<i>121</i>
4.4.5	Limitation	124
4.5	Chapter Summary.....	125

Chapter 5.	Tileable Texture Generation for 3D Visual Try-on	126
5.1	<i>Introduction of tileable texture generation</i>	<i>126</i>
5.2	<i>Related Work.....</i>	<i>132</i>
5.2.1	Texture Synthesis.....	132
5.2.2	Seamless Texture Synthesis.....	135
5.3	<i>Method.....</i>	<i>138</i>
5.3.1	TiPGAN Architecture	138
5.3.2	Seamless Intrinsic Priors Guided Synthesis	140
5.3.3	Seamless Quality Metric – Relative Total Variation.....	144
5.3.4	Cloth Digitization Pipeline	146
5.4	<i>Experiments.....</i>	<i>146</i>
5.4.1	Datasets and Implementation Details	147
5.4.2	Quantitative Evaluation	149
5.4.3	Qualitative Evaluation	151
5.4.4	Ablation Study.....	154
5.4.4.1	<i>Ablation on the PSM & PTM Modules</i>	<i>154</i>
5.4.4.2	<i>Ablation on the Choice of Patch Size p</i>	<i>156</i>
5.5	<i>Discussion.....</i>	<i>159</i>
5.5.1	Effectiveness of the proposed RTV metric.....	159
5.5.2	Capability in Cloth Digitization	160
5.5.3	Limitation of TiPGAN.....	163
5.6	<i>Augmented Reality Mobile Application.....</i>	<i>164</i>
5.7	<i>Chapter Summary.....</i>	<i>166</i>
Chapter 6.	Conclusions and Recommendations for Future Work	169
6.1	<i>Conclusions.....</i>	<i>169</i>
6.2	<i>Recommendations for Future Work</i>	<i>171</i>
References		174

List of Figures

Figure 1-1	Example 2D VTON method of image-based virtual try-on (Choi et al., 2021).	5
Figure 1-2	Examples of the sewing pattern-based 3D VTON software.	7
Figure 1-3	A comprehensive summary of the study's objectives.	10
Figure 1-4	The overview method.....	12
Figure 1-5	The roadmap of the thesis.	14
Figure 2-1	Illustration of 2D VTON with different input types.....	17
Figure 2-2	An overview of VITON (Han et al., 2018).	19
Figure 2-3	An overview of CP-VTON (Wang et al., 2018b).....	19
Figure 2-4	The overall pipeline of SwarpNet (Raj et al., 2018).	21
Figure 2-5	The overall pipeline of M2E-TON (Wu et al., 2019).....	22
Figure 2-6	The Re-identification Network (Wang et al., 2019a).	26
Figure 2-7	A fully textured 3D Avatar Generation of (Lazova et al., 2019).....	26
Figure 2-8	The Pix2Surf model and results of 3D VTON (Mir et al., 2020).....	27
Figure 2-9	Example of fashion image parsing.	30
Figure 2-10	Illustration of pose estimation frameworks. (a) regression-based pose estimation. (b) heatmap-based pose estimation.	32
Figure 2-11	Example of fashion landmarks detection (Liu et al., 2016c).	33
Figure 2-12	The FashionNet model (Liu et al., 2016a).	34
Figure 2-13	Illustration of the FPM (Hu et al., 2018b).....	35
Figure 2-14	The attention maps learned by the Attention Branch (Li et al., 2019c). 37	
Figure 2-15	The overview method of the global-local embedding (Lee et al., 2019b).	37
Figure 2-16	The method overall of the BCRNN model (Wang et al., 2018d).....	38
Figure 2-17	The example of a TPS-based fish warps (Ma et al., 2014).	41
Figure 2-18	The Prototype of GAN.	44

Figure 2-19	The StyleGAN framework (Karras et al., 2019).	50
Figure 2-20	The texture generation framework is based on StyleGAN (Lin et al., 2023).	51
Figure 2-21	An example of an haute couture virtual try-on based on AR technology.	56
Figure 3-1	Illustration of the proposed landmark's definition.	60
Figure 3-2	The overall framework of the fashion landmarks localization module.	62
Figure 3-3	The process of the initial mask extraction.	68
Figure 3-4	Examples of connected components with different poses: $n=1$ in (a) and (b), $n=2$ in (c), $n=3$ in (d).	69
Figure 3-5	Joints definition of pose estimation (Cao et al., 2017).	70
Figure 3-6	Illustration of estimated joints and self-defined clothing feature line.	71
Figure 3-7	Illustration of the new mask created in the occluded area.	72
Figure 3-8	Matching results when using garment outline points.	74
Figure 3-9	Collar landmarks localization.	77
Figure 3-10	Shoulder landmarks localization.	77
Figure 3-11	Examples of Fashion Landmarks Localization Module qualitative results.	82
Figure 3-12	Examples of failure cases.	85
Figure 4-1	BSLT-VTON introduces a body-aware geometric deformation optimization algorithm that enables realistic virtual try-on. (a) An example of BSLT-VTON's garment looseness during wear. (a) Top: T-shirt try-on for a mesomorph shows regular fit. Bottom: Shirt try-on for an endomorph shows tight fit. (b) Illustration of BSLT-VTON's garment position during wear. Top: T-shirt try-on for a teenager shows oversize. Bottom: skirt try-on for a girl with a short and long skirt respectively.	88
Figure 4-2	Illustration of the human body size dependent fashion landmark	

	transformation VTON method.	98
Figure 4-3	Body-aware target garment fashion landmark relocation, where illustrates source garment landmark <i>lis</i> (green dots in (a)), target garment landmark <i>lit</i> (green dots in (b)), and body-aware target garment landmark <i>lit</i> ∇ (rosy dots in (b))......	105
Figure 4-4	The illustration of the virtual try-on image synthesis.	110
Figure 4-5	The virtual try-on results for different body shapes.....	114
Figure 4-6	Application of BSLT-VTON capability to try-on garment for various categories using fashion landmarks to guide the garment deformation.	115
Figure 4-7	Virtual try-on effects of different fashion landmark detection methods: columns (a), and (b) show the source and target garments; columns (c), (f), and (i) show the fashion landmarks using the DFA framework, Match-RNN and the proposed Model i column (d), (g), (j) show segmented garment using these methods; column (e), (h), (k) show virtual try-on effects based on landmarks and segmented garment.	118
Figure 4-8	Virtual try-on effects of different deformation algorithms: columns (a), and (b) show the source garment and target garments; columns (c), (f), (i) show garment deformation using the methods of TPS, MRLS and TPS+MRLS; column (d), (g), (j) show the clothing deformation scales of; columns (e), (h), (k) show virtual try-on effects.....	119
Figure 4-9	Qualitative comparison of SOAT VTON methods along the different column as follows of (c) PFAFN's, (d) DiOr's, (e) Flow-style VTON's, (f) MCG VITON, (g) LaDI-VTON, (h) OOTDiffusion and (i) the proposed BSLT-VTON, respectively. Columns (a) and (b) are the input source garment and target human, respectively.	122
Figure 5-1	A cloth digitization pipeline, where high-quality seamless textures, generated by TiPGAN, can be tiled onto 3D garments. Here, 3D meshes	

	with rich texture designs, such as seamless patterns of ④ , ⑨ , ⑭ , ⑮ , ⑯ , and materials, e.g., seamless normal map of knitted rib ⑳ or knitted cotton ㉑ are rendered, allowing textures to be rendered on different materials while maintaining consistent appearance.	127
Figure 5-2	Texture continuity: (a) non-seamless vs. (b) seamless after tiling or swapping.	129
Figure 5-3	Realistic 3D seamless texture generation for cloth digitization application.	138
Figure 5-4	Architecture of TiPGAN, including the Patch Swapping Module and Patch Tiling Module.	139
Figure 5-5	Qualitative comparison with Image Quilting (Efros & Freeman, 2023), Histogram Blending (Deliot & Heitz, 2019), PSGAN (Bergmann et al., 2017), Self-Organizing (Niklasson et al., 2021), and SeamlessGAN (Rodriguez-Pardo & Garces, 2022). To more intuitively visualize texture tileability, all the results are processed by 2×2 tiling. The proposed TiPGAN excels at preserving the continuity of the overall pattern of textures while providing seamless and semantically consistent boundaries.	154
Figure 5-6	Qualitative ablation results on patch swapping module and patch tiling module. All outputs are tiled 2×2 to visualize texture tileability.	156
Figure 5-7	Effectiveness of the proposed RTV metric in evaluating seamless quality.	160
Figure 5-8	Comparison of 3D cloth digitization performance. The proposed Model iv' results outperform others with high-resolution textures and flexible pattern tiling sizes. Zoom in for a better view.	162
Figure 5-9	Other applications of the proposed tileable texture: the tileable textures are generated by TiPGAN, utilizing the non-flatten and non-seamless reference textures ① to ⑧	163

Figure 5-10	Failure examples synthesized by TiPGAN.	164
Figure 5-11	Screenshot of demo video for of Easy Wear APP.	166

List of Tables

Table 3-1	Naming of landmarks.....	60
Table 3-2	Comparison of human parsing performance with 12 SOTA on the LIP validation set.	66
Table 3-3	Quantitative Results of Fashion Landmark Localization Model.	82
Table 4-1	The mathematical notation used for different variables.....	99
Table 4-2	The size (length & width) of different parts of the garment and the corresponding human body size.....	102
Table 4-3	Quantitative comparison the performance of different virtual Try On method.....	123
Table 5-1	Quantitative comparison.	151
Table 5-2	Quantitative ablation result.	155
Table 5-3	Quantitative ablation result.	158

Chapter 1. Introduction

1.1 Research Backgrounds

Ubiquitous Internet access has accelerated the global expansion and penetration of e-commerce, making it easy for individuals and businesses to buy or sell products without time or geographic limitations. According to Statista, the global retail e-commerce revenue is projected to grow to \$7.0 trillion in 2025 (Emarketer, 2024). Since more and more retailers and brands are moving their battlefield from offline to online, online marketplaces have become the main channel where customers shop nowadays. The COVID-19 pandemic further accelerated the conversion of consumers' shopping patterns and behaviors to an online mode around the world. Among the many products that purchased online, clothing is purchased much more frequently than other categories (Statistics, 2024). Although online clothing shopping has its own advantages in terms of categories, styles, brand diversity and price transparency, the lack of physical touch and trying-on/fitting services have made the evaluation of clothing products difficult and may adversely lower consumers' shopping enjoyment (Batoool & Mou, 2023; Hua et al., 2024; Lavoye et al., 2023). In most traditional online shopping websites, clothing products are displayed, like any other products, in forms of 2D fashion images and/or videos, which has the obvious limitations in addressing consumer's needs, regarding clothing fitting, shape and size, whether or not they match with other fashion pieces, or match with the wearer's skin tone (Bialkova & Barr, 2022; Chidambaram et al., 2024; Hwangbo et al., 2020). In response to various uncertainties in purchase decision making,

e-commerce platforms then offer return policies as a compensation to encourage sales, this however results in a sky-high rate of return for clothing products purchased online. For example, according to a survey by the National Retail Federation, a total of \$761 billion of merchandise was returned to retailers in 2021 in the US (National-Retail-Federation, 2021). This amount surpassed the amount spent on national defense in 2021, which was \$741 billion in total. Moreover, returns also generate tons of carbon emissions and landfill waste, because merchandise, in most cases, cannot be resold as it was originally produced.

To offer consumers new online shopping experience and support sustainable development of the fashion industry, there is a growing interest and research efforts in digital technologies like augmented reality (AR) and virtual reality (VR) that allow consumers to virtually try on clothing (VTON) before making a purchase. The capability of VTON has become a pivotal tool in enhancing user experience, reducing return rates, and bridging the gap between physical and digital shopping. Virtual try-on tasks, whether in 2D or 3D, leverage advanced computer vision, artificial intelligence, and graphic rendering technologies to simulate how garments fit and move on a user's body. This capability not only enables consumers to make more informed purchasing decisions but also offers significant benefits to retailers, designers, and researchers.

The significance of VTON research extends far beyond convenience. It addresses critical challenges in the fashion industry, such as sustainability, inclusivity, and

technological innovation. By reducing the need for physical try-ons and minimizing returns, VTON contributes to a more sustainable retail ecosystem. Moreover, it promotes body positivity and inclusivity, allowing users of various body types and sizes to visualize how clothes would look on their unique figures. From a technical perspective, virtual try-on tasks drive advancements in fields like image generation, 3D modeling, human pose estimation, and machine learning, propelling the evolution of digital fashion.

Overall, research into VTON holds multifaceted significance, examining its impact on consumer behavior, retail efficiency, technological progress, and broader societal trends.

1.1.1 State-of-the-art of 2D Virtual Try-on Technology (2D VTON)

Virtual try-on techniques can be broadly classified into two categories. The first category is 2D image-based virtual try-on (2D VTON).

Early 2D VTONs attempted to achieve simple try-on effects by replacing color and texture of the clothes (Pan et al., 2005; Wang et al., 2003) or substitute human characters (Shilkrot et al., 2012). More recently, 2D VTON studies are considered as 2D image deformation or generation tasks, and Figure 1-1 illustrates example of a 2D VTON method. Most of these 2D VTON studies, including the pioneer work of (Han et al., 2018), and its variants CP-VTON (Wang et al., 2018b), and LA-VTON (Lee et al., 2019a), use Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) to

implicitly learn image generation. Other studies such as LaDI-VTON (Morelli et al., 2023), OOTDiffusion (Xu et al., 2024), and TryOnDiffusion (Zhu et al., 2023) considered virtual try-on via latent diffusion models (Rombach et al., 2022), which significantly improve output image quality by iteratively denoising potential features (Ho et al., 2020; Rombach et al., 2022). However, most of the existing image-based VTON methods do not consider the critical relationship between the size of the garment and the user’s body size, and that a one-size-fits-all approach is often used when visualizing clothing items. Moreover, exiting image-based VTON solutions are deep learning-based that they train on datasets characterized by slender human models, resulting in a bias towards specific body types (Choi et al., 2021; Dong et al., 2019; Morelli et al., 2022). There are a few recent attempts trying to assist users to visualize how a clothing is worn on a body, e.g. some visualized the effect of ‘try-on’ with user measurements (Chen et al., 2023a), and some showed how different sizes of clothing being worn by models of different sizes (Rincon, 2023). However, these efforts appear somehow sorrowful, and users are still not provided with practical help with the real try-on effects. One of the key limitations of existing VTON studies is that they have not taken into account of the complex relationship between a user’s actual body shape and the size of the garment, nor considering a user’s personal preference on fit, e.g. preferring tight fit or loose fit. It thus points out the need for further development of 2D VTON technology in order to better address issues of user sizes and fit preferences. To this end, a novel **body size dependent fashion landmarks transformation VTON** method, called **BSLT-VTON**, is developed. The proposed BSLT-VTON consists of two models,

including a **fashion landmark localization model** and **body-aware garment deformation model**, which will be discussed in detail in later Chapter 3 and Chapter 4, respectively.

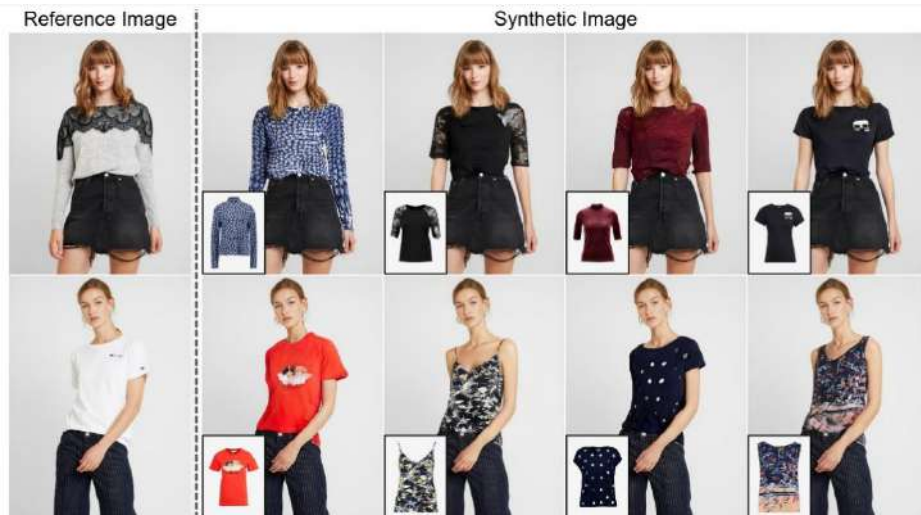


Figure 1-1 Example 2D VTON method of image-based virtual try-on (Choi et al., 2021).

1.1.2 State-of-the-art of 3D Virtual Try-on Technology (3D VTON)

The second category is 3D virtual try-on technology (3D VTON) involving 3D avatars and 3D garment models. In 3D VTONs, users usually first design 2D clothing patterns for a given design, and then virtually sew these 2D patterns virtually to obtain the desired 3D garments through physical simulation or iterative optimization. For example, some research studies (Bartle et al., 2016; Meng et al., 2010; Umetani et al., 2011) have successively implemented the real-time interactive virtual fitting system for 2D clothing patterns and 3D garment styles. This approach is also adopted by most commercial fashion CAD software, such as Optitex 3D (U.S.) (Optitex, 2024), Browzwear (Singapore) (Browzwear, 2024) and CLO3D (South Korea) (CLO3D,

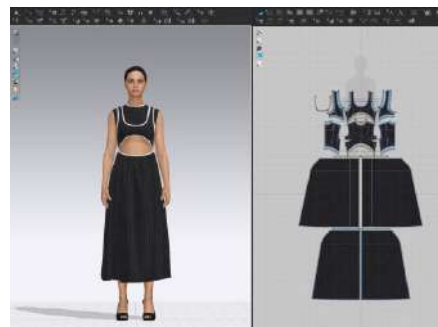
2024), etc., and these software systems offer complete solution for clothing style design, 2D patternmaking, grading, and 3D virtual try-on. The 3D garments offer close true to life effect on how the clothing will drape on a body, see Figure 1-2 for some example CAD solutions. These methods can provide faithful virtual try-on feedback on whether the garment is suitable for a particular person. However, these commercial software are expert-only operations, preventing the widespread use of 3D VTON technology.

Until recently, some deep learning-based 3D VTON methods were proposed. For example, Wang et al. (2018c) presented a novel shared shape space spanning multimodal inputs that allows designers to seamlessly work across multiple domains for interactively design and edit garments based on a range of body shapes without requiring access to expensive physical simulation software. Liu et al. (2023a) introduced a two-level transformer model named Sewformer to generate sewing patterns and 3D garment through one RGB image input, which is training on their synthesis versatile dataset SewFactory and combines more pose variation information into the generation process, so this method allows 3D garment various shape editing. 3D editable garments are usually effective with virtual try-on data. In this study, **a cloth digitization pipeline** is proposed for 3D VTON in Chapter 5, exactly following the same approach for creating 3D garments. Nevertheless, most existing 3D VTON methods are mainly focused on the creation of 3D avatars or 3D garments in realistic shapes and geometric forms, while little attention was given to the garment surface texture creation so far, which are indeed fundamental for realistic 3D try-on effects. In this regard, a new method called **TiPGAN** is proposed in Chapter 5 for **tileable textures**

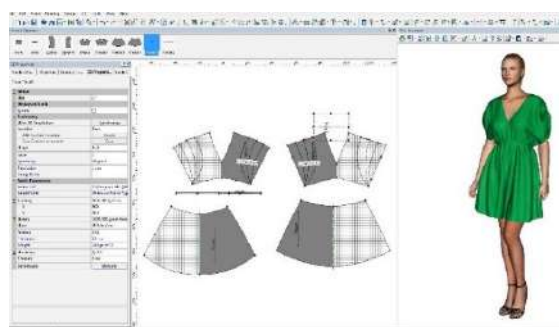
generation into 3D garments for 3D VTON.



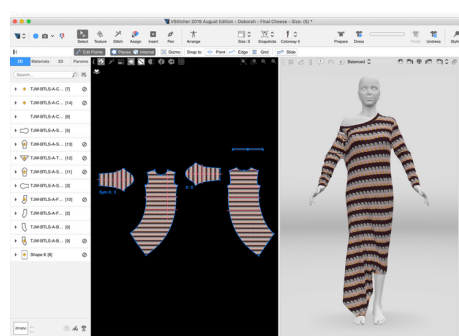
(a) Blender Interface



(b) CLO Interface



(c) Optitex 3D Interface



(d) Browzwear Interface

Figure 1-2 Examples of the sewing pattern-based 3D VTON software.

1.2 Statements of the Problem

The relationship between human body size and garment structure is crucial for virtual try-on. However, there are some known limitations of existing research on image-based virtual try-on.

1.2.1 Limitations of 2D virtual try-on.

Many VTON methods fail to consider the critical relationship between garment size and user body size. They often take a one-size-fits-all approach when visualizing garments, assuming that garments fit all body types in the same way. This results in a

visualization that omits key information and fails to show the user how the garment fits on the user’s body. Recent 2D VTON methods based on generative models, including GANs (Goodfellow et al., 2014) and diffusion model (Ho et al., 2020), need to rely on massive amounts of data to learn image generation models. These methods are mainly trained on datasets featuring slim models (or standard models), which may result in datasets that are biased towards specific body types (Choi et al., 2021; Dong et al., 2019; Morelli et al., 2022). Some other methods gather fashion data relating to various body types and simulate the wearing effects of clothing on different sizes of fit models. However, such data of unique body types are precious and not available on e-commerce and social media websites. Consequently, it is challenging to obtain reliable models when training data is limited.

1.2.2 Limitations of texture generation for 3D VTON.

Existing deep learning-based 3D VTON methods are mainly focused on reconstructing 3D garments from images while ignoring the texture of the model surface or its quality. For example, Wang et al. (2019a) attempted to generate garment textures directly from 3D human body surface, rather than from the garment surface. But the output garment textures are very blurry. Chaudhuri et al. (2021) generated 3D avatars with clear textures, yet their bodies and clothes are combined as a single 3D model, making it impossible to try the same clothing on different avatars. Mir et al. (2020) used an input image and a 3D clothing reference model to output textured 3D garments. However, this work has several limitations. Firstly, their model can only input perfect (in-shop)

fashion images with clean backgrounds. Secondly, the variety of 3D clothing models is limited, only three categories are available, including T-shirts, pants, and shorts. Thirdly, the model has limited learning capability, resulting in misalignment of fashion image and 3D garments. Lastly, even when the input is a high-quality fashion image, the output textured 3D garments still are of low resolution. It is therefore essential to develop a new model generating high-quality textures for 3D garments for 3D VTON.

1.3 Research Aim and Objectives

To overcome the current limitations of 2D and 3D virtual try-on, we have set two primary focuses for this study. The first research focus is to develop an intelligent body-aware 2D virtual try-on system that can capture the relationship between the user’s actual size and the size of the garment. The second one is to develop seamless textures for 3D visual try-on, aiming to produce a texture that can be applied to any 3D model for accurate try-on. Ultimately, we aim to combine the 2D and 3D virtual try-on systems into an augmented reality platform, providing consumers with an immersive virtual try-on experience. The specific research objectives are summarized as follows:

- i. In-depth study and comparison of virtual try-on solutions.
- ii. Comprehensively review the traditional and artificial intelligence technologies related to fashion image understanding and deepen the relevant theoretical knowledge.
- iii. Develop a shape-faithful VTON method based on body size dependent fashion landmark transformation that can effectively transfer the wearable looseness and

wearing position between the source garment and the target per-son, benefiting from a fashion landmark localization model and human body-aware garment deformation model.

- iv. Create high-quality tileable textures for 3D garment and propose a novel cloth digitization process to support 3D visual try-on.
- v. To discuss the integration of the above proposal into Augmented Reality to enable consumers to achieve an immersive virtual try-on experience.

Figure 1-3 provide a comprehensive summary of the study’s objectives.

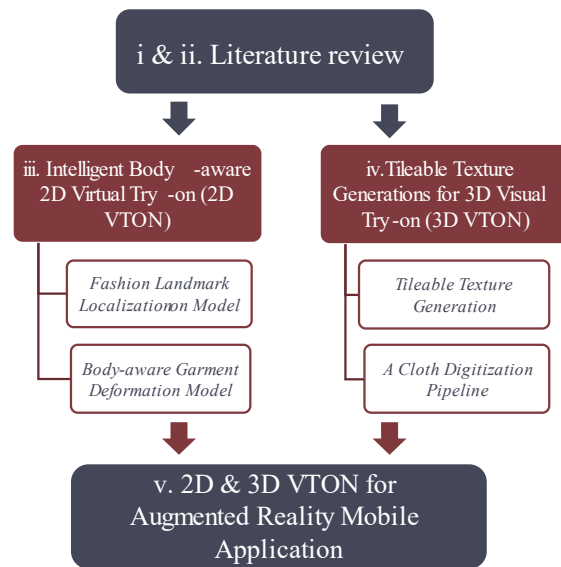


Figure 1-3 A comprehensive summary of the study’s objectives.

1.4 Methodology Overview

According to defined research objectives, Figure 1-4 illustrates the overall methodology of this study.

In terms of 2D VTON, a new method is developed to fully capture the relationship

between the user's actual size and the size of the garment, covering two core developments: (i) fashion landmark positioning model and (ii) body-aware garment deformation model. With these 2 models, the proposed BSLT-VTON method can effectively transfer the wearable looseness and wearing position between the source garment and the target population, allowing users to select clothing following their own preferences of fit.

In terms of 3D VTON, a new texture synthesis model is developed. This model (named TiPGAN) combines recent advances in deep texture synthesis, adversarial neural networks, and latent space operations to synthesize high-quality tileable textures while maintaining perceptual consistency with reference textures. Taking advantage of the properties of tileable texture, infinite tile mapping on the surface of 3D objects, a cloth digitization pipeline is then proposed for the fashion industry. Comparing with existing 3D VTON solutions, the proposed cloth digitization pipeline allows for easy and flexible mapping of high-quality textures to 3D garments. In addition, in 3D garment rendering, different seamless normal maps can be used to simulate the shadow effects of various fabrics or materials to achieve a more realistic 3D visual try-on.

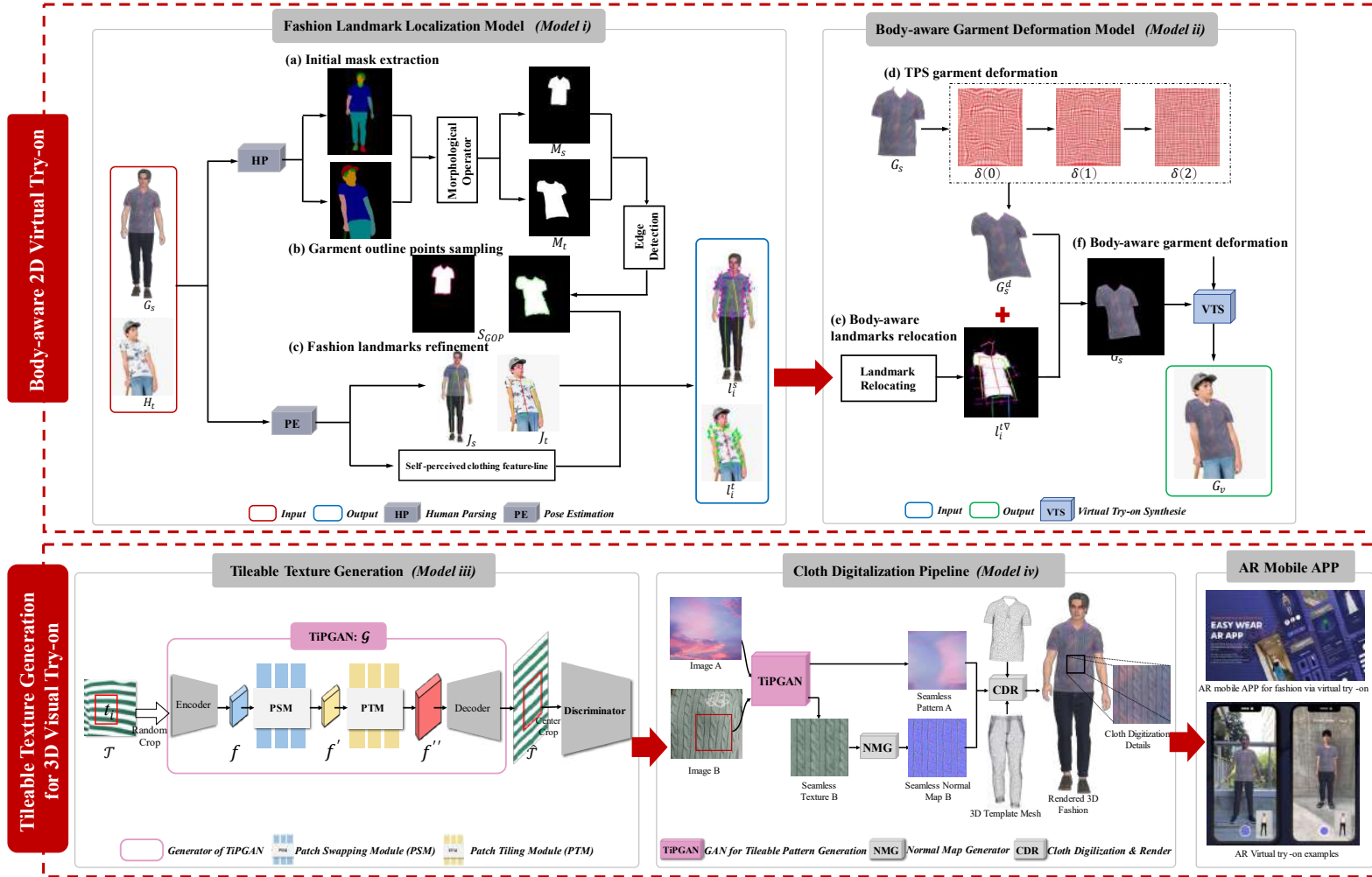


Figure 1-4 The overview method.

1.5 Organization of the Thesis

A concise roadmap of this thesis is presented in Figure 1-5. Chapter 1 of the thesis provides an overview of the fashion industry and its strong connection to research related to virtual try-on, emphasizing the importance of human body shape differences and texture quality for virtual try-on.

Chapter 2 comprehensively reviews the development of related 2D VOTN and 3D VTON solutions. Considering that virtual try-on is a comprehensive computer vision technology, the related technologies are fully reviewed including fashion image understanding, fashion image detection, and image deformation to support the research development in subsequent chapters. In addition, this study proposes a high-quality tileable texture generation model for 3D VTON. Therefore, generative adversarial networks and their improvements on the quality of image generation are comprehensively reviewed. Finally, this chapter also reviews AR technology to prepare for augmented reality 3D VTON fashion applications.

Chapter 3, Chapter 4 and Chapter 5 introduce the specific methods for implementing 2D virtual try-on and 3D visual try-on models, as well as many verification experiments.

Finally, Chapter 6 discusses and summarizes the main findings, contributions and suggests potential future work.

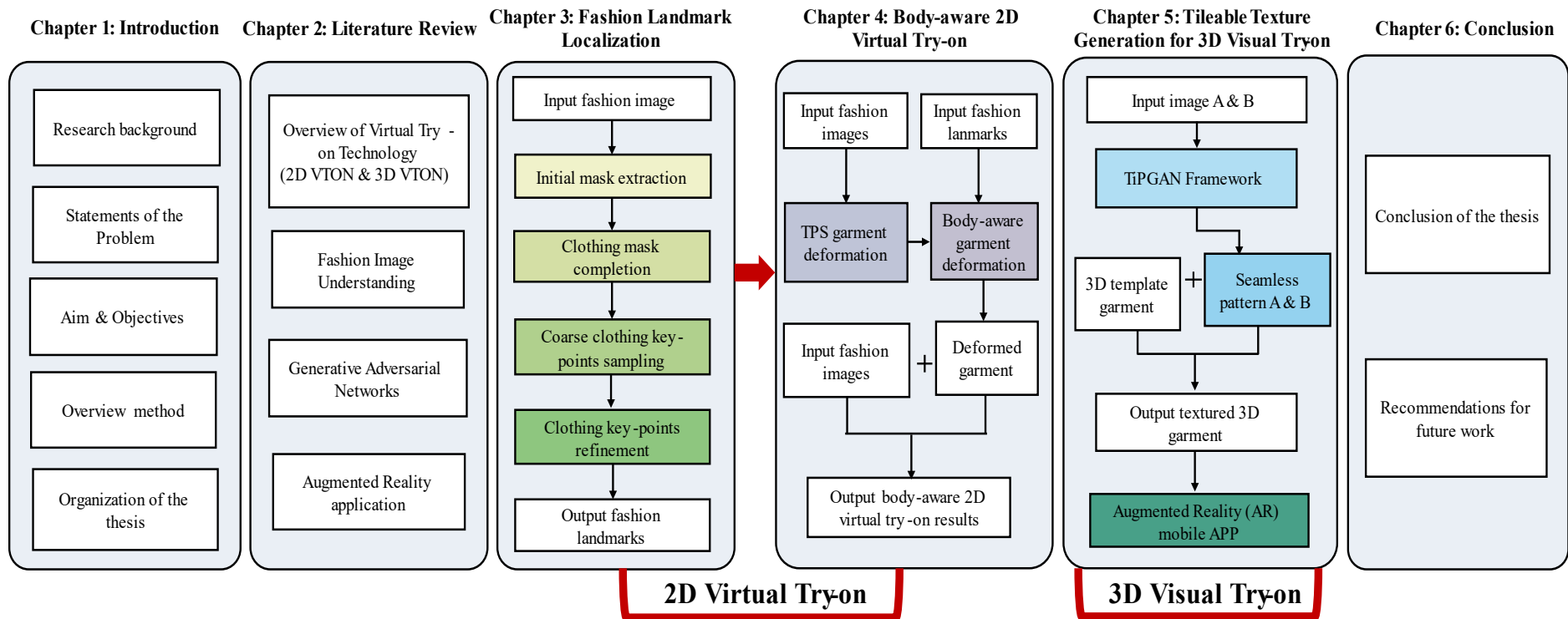


Figure 1-5 The roadmap of the thesis.

Chapter 2. Literature Review

This study aims to develop practical 2D virtual try-on (2D VTON) and 3D visual try-on (3D VTON) solutions. Before presenting the new solutions, this chapter first reviews the relevant work in related areas as follows. First, this chapter comprehensively reviews the current development of related 2D VOTN and 3D VTON solutions. Second, considering that virtual try-on was largely regarded as a computer vision task in the deep learning community, the tasks of fashion image understanding are defined and the relevant methods in fashion image detection and image deformation are reviewed. Third, to provide more thorough background of texture generation model in proposed 3D VTON pipeline, the generative adversarial networks and their advancements in output image quality are comprehensively reviewed. Lastly, this chapter also reviews AR technology to prepare for Augmented Reality (AR) 3D VTON fashion applications.

2.1 Overview of Virtual Try-on Technology

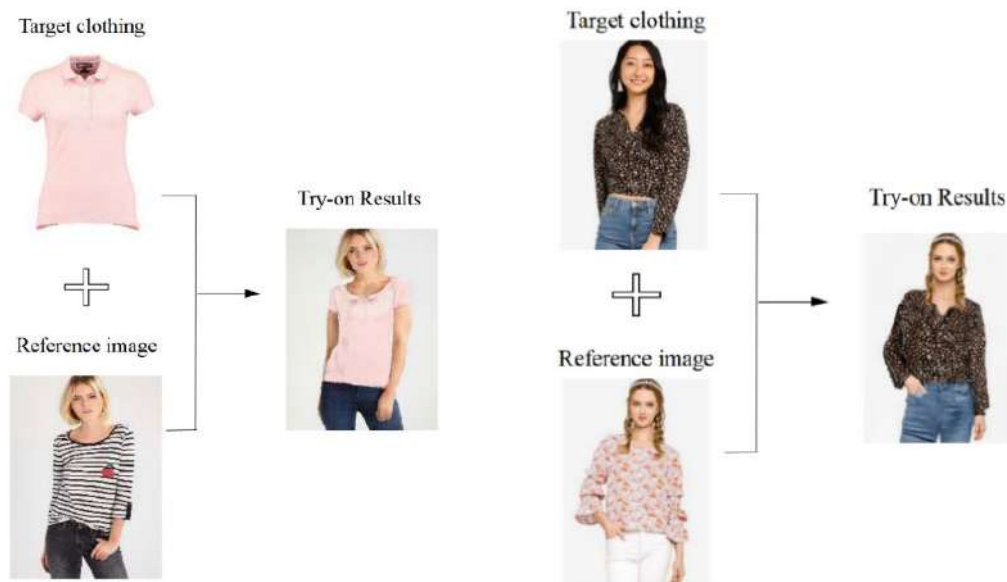
To date, a variety of virtual try-on methods have been proposed. In the past few years, the VTON method has received great research attention due to its various applicability in real-world problems. There is a large amount of literature in this field, mainly including two mainstream solutions: 2D virtual try-on (2D VTON) and 3D virtual try-on (3D VTON). Section 2.1.1 and Section 2.1.2 respectively introduce the current common 2D virtual try-on and 3D virtual try-on technologies.

2.1.1 2D Virtual Try-on Technology

Given that traditional e-commerce platforms usually display 2D fashion images to show consumers. Many researchers have started to study virtual try-on technology based on 2D images, allowing the texture of online retail store clothing images to be digitally worn on user images, thereby enabling virtual try-ons. This kind of try-on solution has become a hot task in recent years due to its ability to adapt to mass consumption scenarios such as bulk of ready-to-wear online shopping.

Virtual try-on based on a 2D image is to wear new clothing into the user/reference image to be tried on to achieve the fitting effect. The traditional 2D VTON approach is usually based on the 2D modeling method, which takes the user/reference image and the target image as inputs and artificially fusing the target clothing image into the user/reference image through image processing. However, the method is inefficient and produces poor results, for example, there are often problems such as overlay errors between the user/reference image and the target clothing. Recently, deep learning based 2D VTON has made remarkable progress, and the goal of these solutions is to be consistent, that is, to input a user/reference image and a target clothing image and output an image of the user wearing the target clothing. In general, the method of 2D VTON is also different according to the content of the input image: 1): input the user/reference image and the target clothing image without model, and then integrate the image of the user wearing the target clothing; 2) input the user/reference image and the target clothing image with model, and then replace the model in the target clothing image with user in the user/reference image, and finally synthesize the image of the user wearing

the target clothing. Figure 2-1 (a) and (b) show the desired virtual try-on results for the input and output of these two types of methods respectively.



(a) Target clothing without model

(b) Target clothing with model

Figure 2-1 Illustration of 2D VTON with different input types.

Both approaches require that the final synthetic result be real and natural. At the same time, the user's body shape and posture remain unchanged, and the attributes and details of the target clothing are still retained in the final composite image. Therefore, in order to meet these requirements, researchers have also made in-depth exploration and research on the 2D VTON task from different perspectives.

2.1.1.1 2D VTON based on reference image and target clothing without model

The reference image and target clothing-based 2D VTON methods refer to the user image and clothing product as input, which are the common input contents of 2D VTON.

For example, Han et al. (2018) proposed VITON as one of the well-known tasks, which consists of three parts. First, an encoder-decoder network is used to generate coarse results. The encoder-decoder loss function is written as the sum of the perceptual loss and the L_I loss, expressed as follows:

$$L_{G_C} = \sum_{i=0}^5 \lambda_i \|\phi_i(I') - \phi_i(I)\|_1 + \|M - M_0\|_1 \quad (2-1)$$

In the initial component, $\phi_i(y)$ represents the feature map of image y within the i -th layer of the visual perception network ϕ . Then in the clothing deformation stage, the target clothing (without model) is deformed using Thin Plate Spline (Belongie et al., 2002). Finally, the rough image obtained in the first step is fused with the deformed clothing to get the final result. Figure 2-2 shows the flowchart of VITON. Although this method preserves the details of the final clothing to a certain extent, there are still some problems in practical applications, such as the difference between the clothing attributes (such as neckline type, sleeve length, etc.) in the synthesis results and the target clothing.

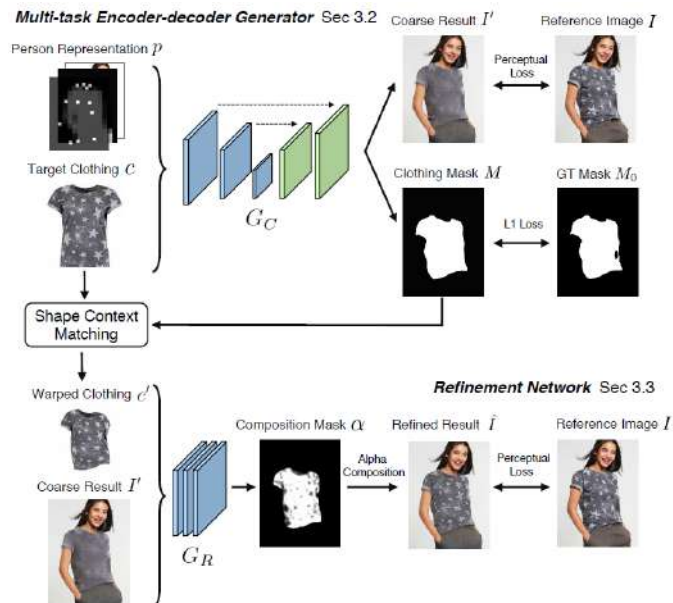


Figure 2-2 An overview of VITON (Han et al., 2018).

The subsequent CP-VTON (Wang et al., 2018b) was an improvement based on the VITON method, in which the authors first designed a new geometry matching module to make the TPS a fully learnable component, and secondly, the generation module was also refined to obtain better fitting results than the VITON method. Figure 2-3 shows the flowchart of CP-VTON. However, the CP-VTON method still has cases of failure when dealing with the following situations, such as, the shape information of target clothing is not properly preserved, rare posture and the inside and outside of the clothing cannot be distinguished.

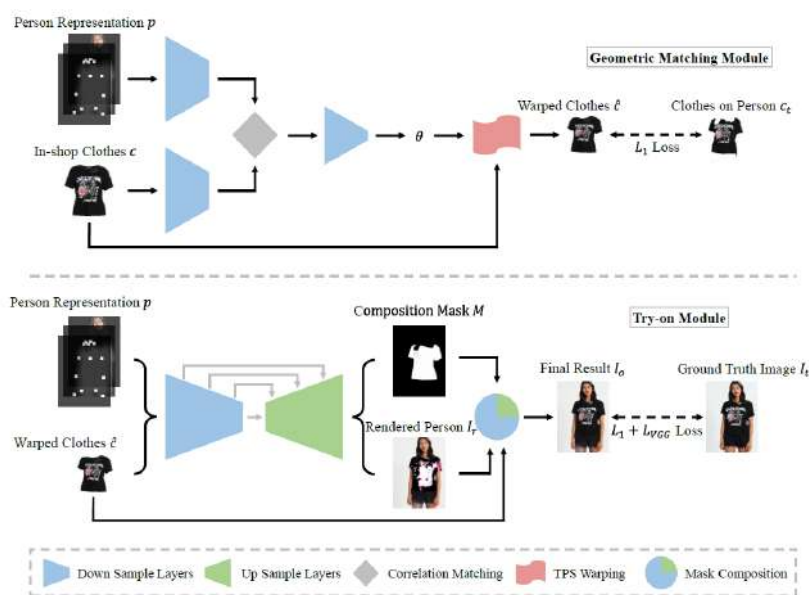


Figure 2-3 An overview of CP-VTON (Wang et al., 2018b).

In 2019, Dong et al. (2019) improved CP-VTON by training a geometric matching network to solve the clothing matching problem, effectively improving operational efficiency. But this method still fails in some cases, such as crossed arms, arms

obscuring clothing, and other specific gestures. And since the model cannot distinguish between the inside and outside of the clothing, some unreasonable fitting results will be generated.

2.1.1.2 2D VTON based on reference image and target clothing with model

The reference- and target clothing with model-based approach makes it have more application scenarios by changing the clothing product in reference image to the target clothing, because the target clothing with model image is more readily available than the clothing without model image and is less affected by the shooting conditions. In addition, the model image has implied the way and place of the target clothing, which helps to match the user's body shape and posture. Based on this type of input, the researchers have also proposed many methods to achieve the final try-on results. For example, Raj et al. (2018) proposed a network framework (SwarpNet) to achieve clothing virtual try-on for user's arbitrary body shape and posture. The authors decompose the task into two stages. For the first stage, the user's posture information and the model's clothing information were leveraged to predict the corresponding human parsing when the model's clothing was worn on the user, which is similar to FashionGAN (Zhu et al., 2017b). The second stage was to synthesize the image of user wearing the target clothing based on the predicted human parsing result and the texture information extracted from the model image. Figure 2-4 shows the flowchart of SwarpNet. While this technique effectively maintains alignment with the user's body

shape and posture, reliance on extracting texture details primarily from the model image means there is no assurance of consistency with the user's skin tone in the generated image.

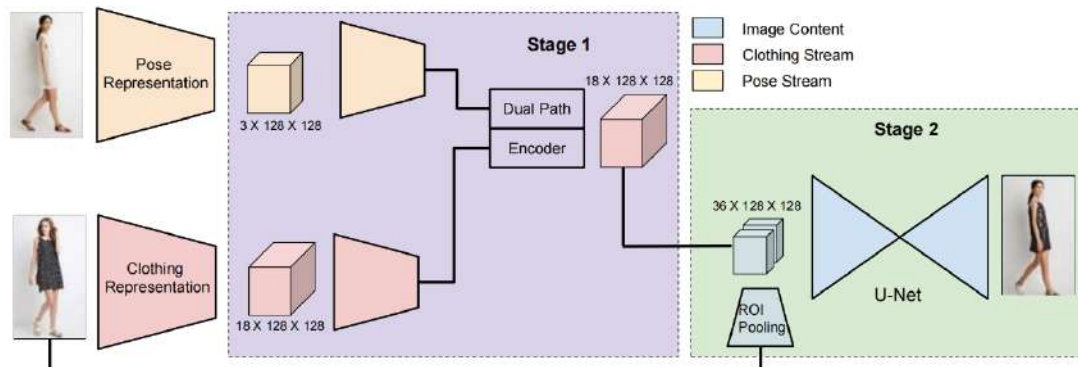


Figure 2-4 The overall pipeline of SwarpNet (Raj et al., 2018).

Moreover, the low resolution of the images generated by this method cannot be widely used. Wu et al. (2019) proposed a virtual fitting network (M2E-TON) based on dense posture. The task is converted into a posture transfer task. Based on the dense pose results of model images and user images, the authors first generated model images consistent with user postures by training a pose alignment network (PAN). Then, on this basis, the texture refinement network (TRN) underwent training to incorporate the modified clothing textures into the image aligned with the pose. Finally, the fitting network (FTN) was utilized to adapt the intended clothing to the original target individual image. However, the final result is the loss of some detail texture, such as the texture at the junction of the hair and collar. Figure 2-5 shows the flowchart of M2E-TON.

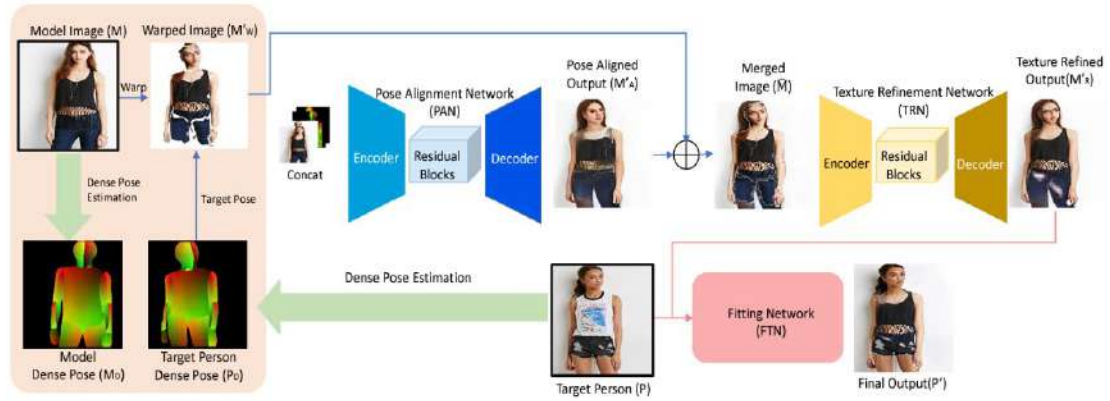


Figure 2-5 The overall pipeline of M2E-TON (Wu et al., 2019).

2.1.2 3D Virtual Try-on Technology

The core technology of the 3D VTON is to wear 3D garments on the 3D human body to achieve the effect of virtual try-on. In 2002, Lands' End has launched its first online VTON system that allows users to customize different parameters such as height, weight, circumference, skin tone, etc. to generate a virtual user body type, and then select predefined clothing and body postures to display (Cordier et al., 2003). In the later development, this VTON technology is also a widely used online clothing display method by major fashion brands have launched similar virtual fitting services. However, this online virtual fitting system provides a 3D Avatar that cannot accurately fit the user's body size simply by using the user's input of body size information. Moreover, studies have shown that the same human body measurement does not equal to the same body shape, which may lead to an invalid try-on result (Hu et al., 2018a). The styles of fashion to be tried on by these systems are limited to the fashion brand's own internal style, which limits the customer's choice to some extent.

Another solution for online fitting is the fitting robot represented by Fits. Me. Considering the impact of 3D human body data on virtual fitting, some companies and research institutes dedicated to designing humanoid assembled robots with sophisticated mechanical structures and advanced control systems to fit the body type of the consumer. For example, Abels and Kruusmaa (2013) developed by Fits.me company. Guo et al. (2016) designed a female fitting robot, Hong Kong polytechnic university (Peng, 2013), Zhejiang university (Li et al., 2019a; Xu et al., 2018), and other universities or institutes have also developed clothing fitting robots with their own characteristics. Although, the methods of using fitting robots eliminate clothing modeling and simulation, it also has disadvantages of terms of its poor body shape simulation effect, low accuracy, small deformation range, expensive price, etc.

To diversify clothing styles and achieve more new styles to try on, a CAD-based 3D VTON solution is proposed, which is where the user first designs 2D clothing patterns, and then uses physical simulations or iterative optimization of relevant parameters, stitching patterns and obtaining the required 3D garment. Also, Meng et al. (2010) added a 2 D/3D garment patterns design module to virtual fitting system to help users with interactive clothing design. Umetani et al. (2011) realized interactive two-way editing that allows seamless transitions between 2D patterns and 3D high-fidelity analog overlays. This method enables a continuous, interactive design process where both 2D and 3D representations are visible and can be edited concurrently in real time. Creators have the ability to dynamically modify 2D patterns and receive immediate

online feedback, facilitating swift prototyping and enhancing comprehension of intricate overhanging structures. In 2016, Bartle et al. (2016) designed a new operational framework enabling designers to directly implement the envisioned variations in a 3D environment.. Some of these proven solutions, such as Optitex 3D, Lecra, Browzwear, CLO3D and Marvelous Designer, etc., have been commercialized to enable the complete system of clothing styles design, the pattern-making, the grading, and the 3D virtual try-on. The 3D garments due to fabrics physical simulation can generate a drape effect similar to that of real-world clothing worn on a human body. However, the existing CAD-based 3D VTON can help the simulated fabric/clothing achieve realistic visual effects, but there is still a gap compared to the real fabric/clothing. And since the interactive nature of the software, professional operators, such as complex parameter tuning, are required to be involved in 3D garment modeling/simulation processes. This kind of 3D modeling to try-on solution cannot achieve rapid mass creation and is not a situation for mass consumption scenarios such as bulk of ready-to-wear online shopping.

Since 3D VTON's two necessities are 3D modeling of the human body and 3D modeling of clothing, the researchers focused on rapid reconstruction of 3D garment and 3D human body. Most recently, deep learning has achieved great success in the computer vision domain, demonstrating the powerful ability of deep neural networks in feature extraction. The researchers tried to recover 3D models quickly from the images based on DL to provide usable 3D data for virtual try-on technology. For

example, 3D human body reconstruction recovered or predicted a 3D human body from a single image, providing the core 3D human models for 3D VTON technology quickly. (Guler & Kokkinos, 2019; Kanazawa et al., 2018; Varol et al., 2018; Zheng et al., 2019).

3D reconstruction of clothing is relatively difficult due to several features of clothing, such as non-closed surfaces, complex shape topologies, complicated geometric details, etc., whose reconstruction depends on 3D datasets. Bhatnagar et al. (2019) stripped the 3D garment from the scanned clothed human body by predicting parametric garment geometry and layer it on top of the SMPL (Loper et al., 2015) model. Their MGN can dress in varying body shapes and poses because of the layered representation. Besides, Zhu et al. (2020) introduced a novel baseline method that combines the benefits of grid and implicit representation to faithfully reconstruct 3D garment from a single image, which can quickly provide a large amount of 3D garment data for 3D VTON. The extensive research on both 3D human body reconstruction and 3D garment reconstruction has significantly advanced 3D Virtual Try-On (VTON) technology. Armed with these 3D reconstruction techniques, in recent years 3D VTON has started to focus on ways to estimate the body's 3D pose and shape from the 2D image, and to generate textures for the body to visually generate the 3D model of wearing clothes. For example, Wang et al. (2019a) proposed an end-to-end learning strategy that can generate human body textures under the supervision of re-identification. The texture extracted by the author rendered the composite image and used the re-identification network as a perceptual metric to maximize the similarity between the rendered image and the input image. Figure 2-6 shows their workflow diagram and output.

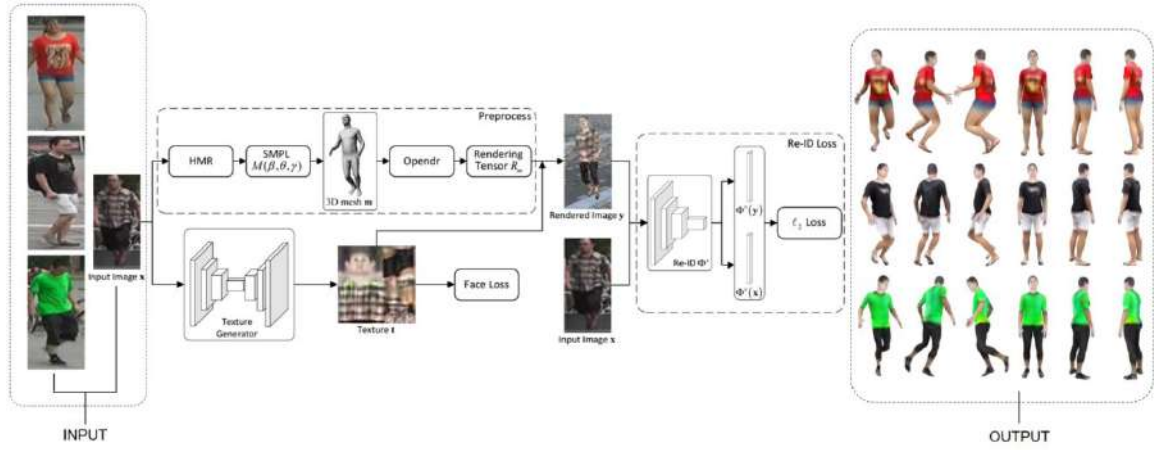


Figure 2-6 The Re-identification Network (Wang et al., 2019a).

In addition, Lazova et al. (2019) non-rigidly registered 3D-scanned SMPL models, efficiently encoding textures and geometries into corresponding images and then used the image-to-image conversion method to infer textures and geometries in the UV space of the SMPL model. The author turned a difficult 3D reasoning task into a simpler image conversion task, enabling the task of predicting a person's full 3D avatar from a single image. Figure 2-7 shows their overall pipeline and output.

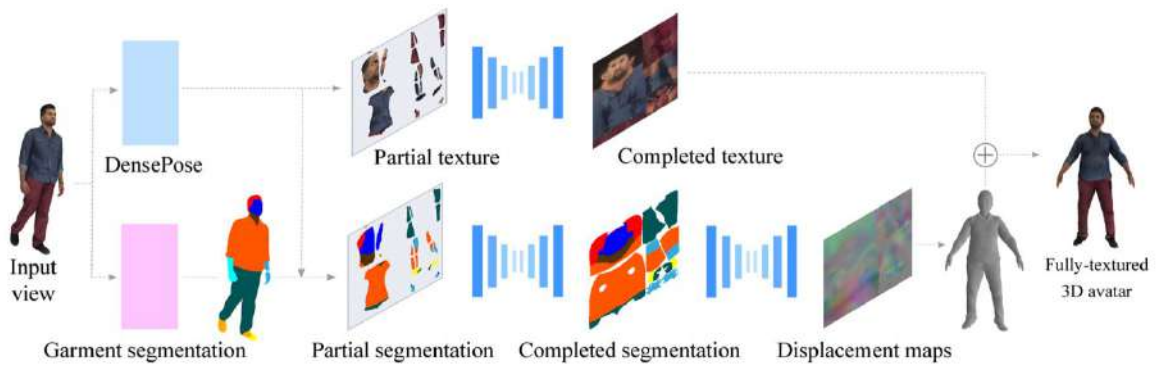


Figure 2-7 A fully textured 3D Avatar Generation of (Lazova et al., 2019).

Subsequently, in 2020, Mir et al. (2020) proposed a method that can automatically

transfer the texture of the clothing image (front and back) to the SMPL in real time to output a 3D model wearing the garment. The author's idea is to use only the shape information extracted from the clothing image and then to learn the dense correspondence of silhouettes to a 2D-UV map of a 3D garment surface. Figure 2-8 demonstrates the Pix2Surf model and results of their 3D VTON.

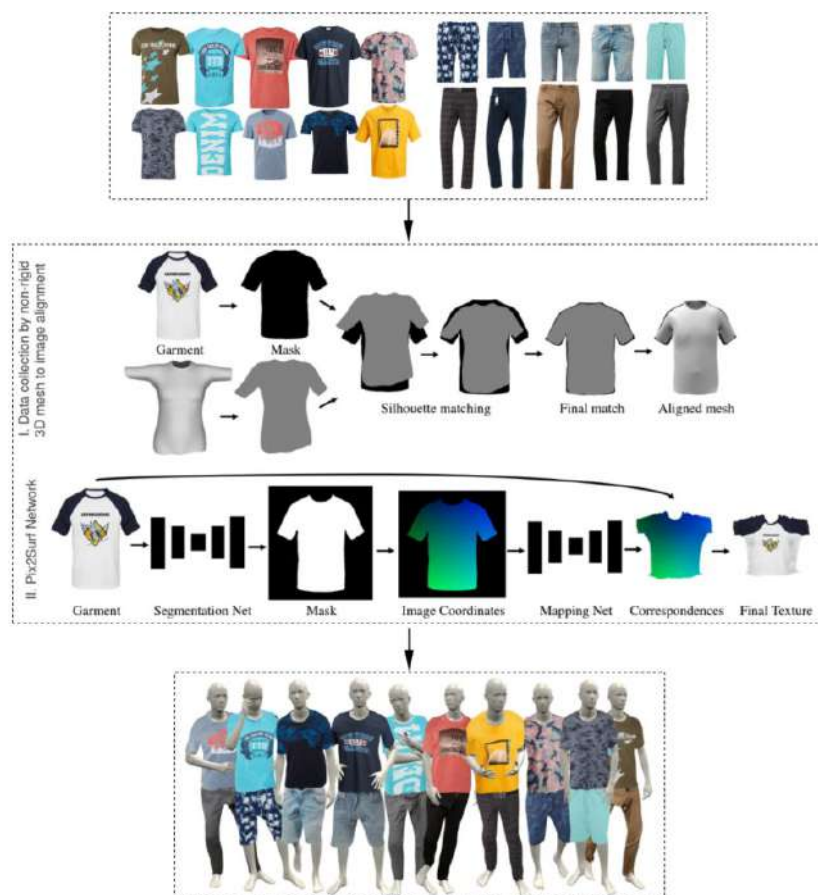


Figure 2-8 The Pix2Surf model and results of 3D VTON (Mir et al., 2020).

2.2 Fashion Image Understanding

Fashion image understanding, namely Fashion AI, focuses on using algorithms and techniques to automatically analyze and understand visual information related to fashion and clothing. It has gained high R&D fervor because it supports many human-

centered application processes, such as e-commerce, tailor-made styles, and smart fashion development. Recent fashion image understanding tasks are mainly based on deep learning methods, including clothing recognition and retrieval (Fu et al., 2013; Kalantidis et al., 2013; Liu et al., 2012; Yamaguchi et al., 2013), fashion semantic segmentation (Gong et al., 2019), fashion classification (Bossard et al., 2013; Kolisnik et al., 2021) and fashion trend analysis (Kiapour et al., 2014; Simo-Serra et al., 2015; Yamaguchi et al., 2014). To better extract the prior information of fashion items, previous work has explored bounding boxes (Huang et al., 2015), garment mask parsing (Belagiannis et al., 2015; Pfister et al., 2015) and fashion landmarks (Chen et al., 2019a; Ge et al., 2019; Guo et al., 2016; Kim et al., 2021; Yu et al., 2019c). Among these explorations, fashion landmark detection is an effective approach because it not only captures the bounding box of the clothing item but also localizes the functional region of the clothing.

2.2.1 Fashion Parsing

The primary goal of fashion parsing is to detect and categorize particular fashion items, such as various kinds of clothing (e.g., dresses, shirts, pants) or accessories (e.g., shoes, bags), as shown in Figure 2-9. Early fashion parsing was proposed by Yamaguchi et al. (2012), which utilized human pose estimation to predict the regions corresponding to clothing. Liu et al. (2013) utilized color category labels and integrated human pose estimation with pixel-level category classifiers for fashion parsing. Nonetheless, these methods, which rely on manually designed features, fail to effectively capture the

relationship between the appearance and structure of garments. After that, researchers look to deep learning-based methods. For example, Liang et al. (2015a) used two convolutional neural networks to predict the position of mask and shape parameters of each label mask, respectively. Liang et al. (2015b) integrated global image-level and local hyper-pixel in CNN network to enhance the fashion parsing precision. In addition, Zhao et al. (2018) proposed a mutually fine-grained integrated network (PGN) that can simultaneously process multiple people parsing without grouping. Ruan et al. (2019) introduced the CE2P framework, which utilizes a high-resolution embedding module, a global context embedding module, and an edge-aware module to obtain human parsing results with higher accuracy. Gong et al. (2019) and Wang et al. (2019b) offer more detailed parsing information for fashion images by considering the semantic hierarchical structure. The model accuracy of fashion parsing is increasing over the years. For example, Zhou et al. (2021) proposed a bottom-up mechanism to learn multi-person parsing and pose estimation by utilizing structural information (MGHR) at varying levels of detail for different individuals. They simplified the complex task of pixel grouping by converting it into an easier multi-person joint assembly task, thereby reducing the complexity of recognizing individual human instances. Zhang et al. (2022a) introduced a method called Part Perception Relationship Modeling (PRM) to tackle fashion parsing, which generates features with adaptable context, accommodating human parts of different sizes and shapes. EOPNet (He et al., 2023) integrates attention mechanisms and metric learning to tackle human parsing, providing a novel approach for fashion-related applications that lack predefined categories for human body parts.

Dou et al. (2024) introduced a Decouple Re-identification and human Parsing (DROP) approach to extract task-specific features tailored for both occluded person re-identification (ReID) and human parsing. The method separates the feature learning processes for person ReID and human parsing, incorporating only the location information of the person into the human parsing branch to aid the model in understanding semantic spatial context. Additionally, the ReID branch incorporates a part-aware compactness loss to strengthen the distinction of instance-level part features. Benefiting from this, the clothing masks can be quickly extracted as a priori for fashion landmarks. This also further ameliorates the localization of clothing and outline key points.



Figure 2-9 Example of fashion image parsing.

2.2.2 Pose Estimation

Human body pose estimation aims at the detection or localization of human body parts from an input image and the construction of a human body representation. Because of its broad range of applications across various domains (e.g., virtual fitting, motion analysis and AR, VR, etc.), it has gained more and more attention. Pose estimation using deep learning can be broadly summarized into two categories: regression-based

methods and heatmap-based methods, as shown in Figure 2-10. Regression methods mainly learn the mapping from input images to body joint positions or human model parameters (Toshev & Szegedy, 2014). For example, Toshev and Szegedy (2014) leveraged a convolutional neural network (CNN) to learn human joint points from images. Sun et al. (2017) reparametrized skeleton-based representations and proposed a composition-aware network that contains both human body information and pose structure. Luvizon et al. (2019) transformed the feature map into joint coordinates with soft-argmax function to perform pose estimation. Li et al. (2021a) introduced a transformer-based cascade network into this task to capture the human joints' spatial information by a self-attentive mechanism. To acquire richer pose features, Li et al. (2014) predict joints' coordinates from images using a regressor on the one hand and sliding window to detect body parts from image blocks on the other hand. Fan et al. (2015) modified the CNN model to simultaneously detect joints and localize them to improve the accuracy of pose estimation. Rather than directly estimating body-critical coordinate values, heatmap-based methods train the network to minimize the loss between the predicted heatmap and the target heatmap at each joint (Chen & Yuille, 2014; Newell et al., 2017). Wei et al. (2016) exploited a stage-by-stage optimized convolutional network sequential framework to predict the position of human joints. Newell et al. (2017) captured information at various scales utilizing an encoder-decoder network to improve the accuracy of joints prediction by repeating bottom-up and top-down processing. Sun et al. (2019) connected multi-resolution sub-networks in parallel to learn multi-scale fusion to predict more accurate joints' heat-map. Considering the

computational cost, Yu et al. (2021) designed conditional channel-weighted blocks to learn high-resolution representations, inherently exchanging information between channel and resolution.

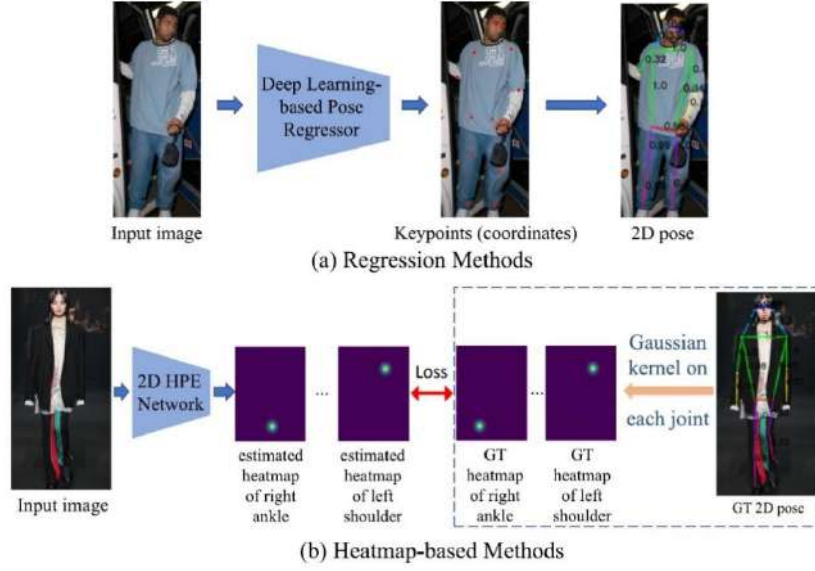


Figure 2-10 Illustration of pose estimation frameworks. (a) regression-based pose estimation. (b) heatmap-based pose estimation.

2.2.3 Fashion Landmark Detection

Fashion landmarks detection is necessary for understanding fashion images, and it supports many fashion applications in the field of computer vision domain, such as fashion classification, retrieval and recommendation. Fashion landmark detection not only locates functional key points on garments, but also implicitly captures the bounding boxes of garments' functional areas. The features extracted from these landmarks can distinguish between different garment categories, design styles, and garment textures. By analyzing fashion landmarks on the human body, the algorithm provides a better understanding of the user's body shape and size, offering accurate virtual fitting suggestions. Unlike the pose estimation task, fashion landmark detection

is fashion-centric, locating specific points on the human body that are critical to the fit and shape of a garment, such as the shoulder point, waistline, neckline, and sleeve length (Liu et al., 2016c). Figure 2-11 (a) and (b) show the difference between human joint estimation and fashion landmark detection. Because of the non-rigid deformation of garments and the diversity of styles, fashion landmark detection is more challenging. Figure 2-11 (c) and (d) are examples of fashion landmarks in the deformed garment and in self-occluded dressed human, respectively.



Figure 2-11 Example of fashion landmarks detection (Liu et al., 2016c).

If the input fashion image is denoted as I , the formal definition of landmark detection is outlined in Eq.(2-2).

$$L = \{L_n: n = 1, \dots, N\}, L_n \in R^2 \quad (2-2)$$

L includes all identified landmarks, and N is the total number of landmarks. After reviewing a large amount of literature on fashion landmark detection, fashion landmark detection tasks can be roughly divided into the following categories according to the deep learning models they use. They are regression-based fashion landmark detection, pose estimation-based methods, attention model-based fashion landmark detection, and constraint-based methods.

2.2.3.1 Regression Based Models

Regression-based approaches consider landmark detection primarily as a regression problem. Through training a deep neural network and utilizing a regression function, these methods aim to map an image to a set of landmarks by minimizing the mean square error across all instances. The learning goal can be described as:

$$f^* = \min_f \frac{1}{N} \sum_{n=1}^N \|f(I^n; \theta) - L^n\|_2 \quad (2-3)$$

Where N represents the number of samples, f is the regression function.

Liu et al. (2016a) were the pioneers in introducing the concept of fashion landmark detection and utilizing FashionNet to predict clothing attributes. They also released a comprehensively annotated dataset named DeepFashion to facilitate the supervised learning of multiple fashion tasks. In Figure 2-12 orange, green, and blue parts represent the three branches of the FashionNet model respectively.

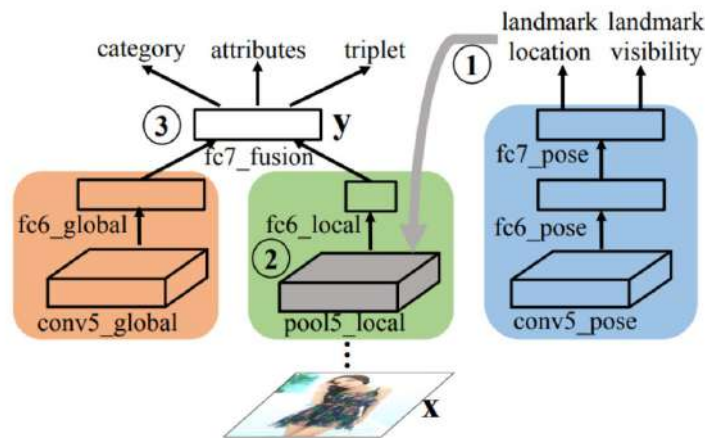


Figure 2-12 The FashionNet model (Liu et al., 2016a).

In the same year, Liu et al. (2016c) developed a cascaded network framework to enhance the precision of fashion landmark estimation. Initially, they utilized VGG-16 (Simonyan & Zisserman, 2014) to predict landmarks, followed by estimating the fashion landmark offsets, and ultimately outputting the contextual landmark offsets at each stage.

2.2.3.2 Pose Estimation Models

Some scholars argue that the pose estimation task is akin to fashion landmark detection, which is to find key feature points that can represent the person or clothing in the image. A fashion posture machine (FPM) for fashion landmark detection was proposed in (Hu et al., 2018b). FPM conducts research on tasks based on 2D pose estimation, with the aim of effectively locating fashion landmarks without prior knowledge of body structures, especially when the clothes in the image are affected by the model's pose.

Figure 2-13 shows the overall pipeline of FPM.

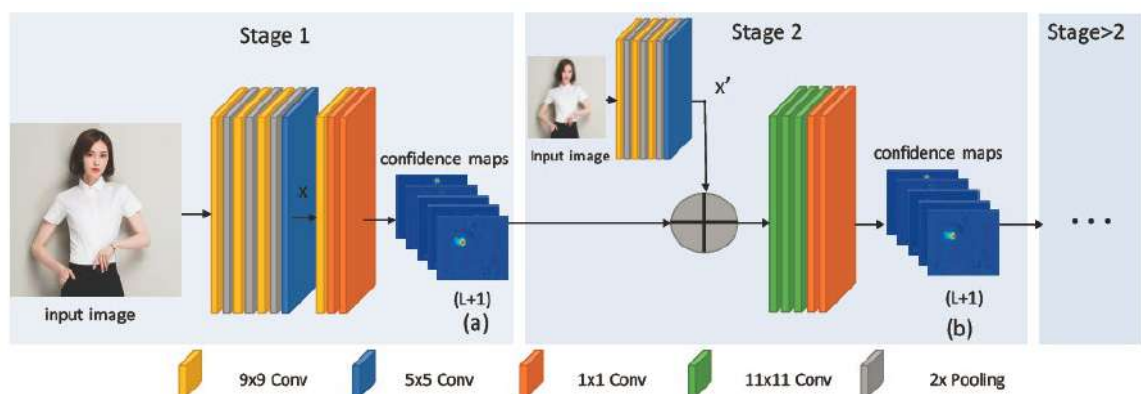


Figure 2-13 Illustration of the FPM (Hu et al., 2018b).

The authors first take the image features obtained by the CNN as input in the FPM task, and then pass the first stage obtained the image features and confidence maps to the next stage and serve as inputs to that stage. Subsequently, convolutional networks allow multiple classifiers to group predictable features based on contextual information to eliminate the influence of human pose on feature detection.

2.2.3.3 Attentive Models

The landmark detection method based on the attention mechanism is also one of the widely used methods. For instance, Liu and Lu (2018) proposed a fashion analysis network that can accurately locate fashion landmarks by increasing the resolution and up-sampling of heat maps. The authors used the L_2 loss to train the landmark localization branch, the formula is as follows:

$$l_{landmarks} = \frac{1}{N} \|M_{ijk} - M'_{ijk}\|_2^2 \quad (2-4)$$

Where M is the ground truth of the landmark heatmap, $M' \in R^{224 * 224 * 8}$, N is the total number of array elements. The method used landmark heat maps as a reference to generate uniform attention so that the network can have enough information to enhance or reduce features. Figure 2-14 visualizes the attention maps.



Figure 2-14 The attention maps learned by the Attention Branch (Li et al., 2019c).

Most recently, Li et al. (2019c) accurately detected fashion landmarks because of a spatial-aware nonlocal model they devised. The model is acceptable for high-level semantic knowledge utilizing spatial attention maps to encode a priori knowledge capability. The global-local embedding model introduced by Lee et al. (2019b) contains both global and convolutional procedures. The model exploits long distance dependencies and local information in images to detect fashion landmarks. Figure 2-15 shows the overall overview method of global-local embedding.

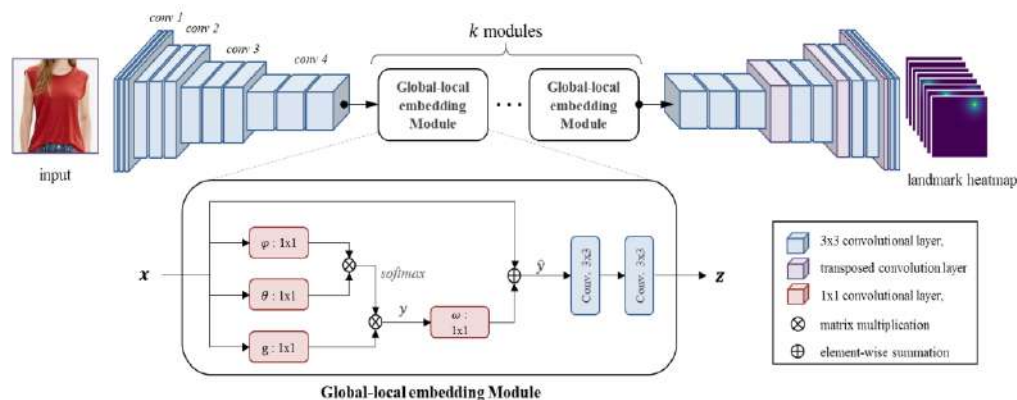


Figure 2-15 The overview method of the global-local embedding (Lee et al., 2019b).

2.2.3.4 Constraints Based Methods

In addition to the above models, it has also been suggested that conditions such as layout constraints or semantic constraints such as spatial relationships should be considered to improve the performance of fashion landmarks detection. Therefore, semantic constraints are proposed to incorporate into its model. For example, Yu et al. (2019c) attempted to explore the layout relationships among fashion landmarks through

layout graph inference. Initially, the model maps image features to nodes within the graph. Next, graph inference is employed to determine global layout relationships. Finally, the graph nodes are mapped back to enhance the features. Furthermore, Wang et al. (2018d) designed the BCRNN model to receive and process high-level semantic knowledge, which utilizes global syntactic constraints to output the layout of fashionable landmarks. Figure 2-16 is the method overview of the BCRNN.

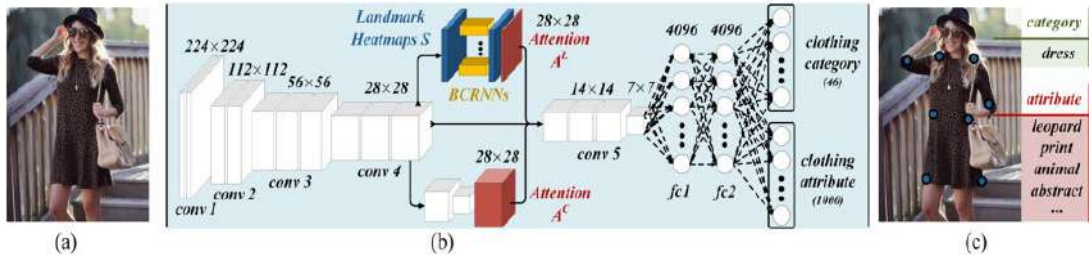


Figure 2-16 The method overall of the BCRNN model (Wang et al., 2018d).

2.2.4 Image deformation

Since the *model ii* of 2D VTON aims to transfer clothing texture without quality loss and appropriately deform the relevant areas of the image. This section will focus on the texture transfer and deformation methods. Texture transfer technology was first proposed by (Catmull, 1974) and is one of the important technologies for generating images based on geometric models in computer graphics. This technology is to cover any plane image on the surface of the object, so as to generate a texture with real color feeling. Technology can fully reflect the various structural details and features contained in the surface of the object by using the texture and can also judge the original shape of the object by observing the deformation degree of the texture. Because this

technology is prone to problems such as seams and distortions in the process of texture tiling, many improved algorithms have been proposed since (Catmull, 1974). Busto et al. (2010) used mathematical functions to complete the transformation and generation of texture images, and this image has a 3D feel. Ma et al. (2014) devised a constrained technique for texture transfer and mapping that enables interactive manipulation of the mapping region. This method facilitates challenging tasks within the realm of surface texture transfer and mapping. Considering the body-aware 2D try-on effect, a powerful deformation algorithm is utilized to ensure that the texture remains high-definition during the normal deformation of the garment. The proposed body-aware 2D VTON method is based on thin plate spline (TPS) for texture transfer. This is a popular basis function for depicting coordinate mappings from a 2D space to another 2D space. Let v_i represent the target function values at points (x_i, y_i) on the plane, where $i=1, 2, \dots, p$. Specifically, v_i will be assigned to the target coordinates (x_i', y_i') in a sequential manner to create a continuous transformation for each coordinate. Assume that the points (x_i, y_i) are all distinct and not in a straight line. The TPS interpolant $f(x, y)$ minimizes the bending energy.

$$I_f = \iint_{R^2} (f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2) dx dy \quad (2-5)$$

And has the form as follows:

$$f(x, y) = a_1 + a_x x + a_y y + \sum_{i=1}^p \omega_i U(\|(x_i, y_i) - (x, y)\|) \quad (2-6)$$

Where U is the radial basis function:

$$U_{(x)} = \gamma^2 \log r \quad (2-7)$$

The solution form is generalized, and the following matrix is used to represent the previously mentioned values, there are:

$$\mathbf{P} = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & x_n & y_n \end{bmatrix} \quad (2-8)$$

where $\gamma_{ij} = \|p_i - p_j\|$ represents the distance between two control points. Let the matrix $\mathbf{L}$ be:

$$\mathbf{L} = \begin{bmatrix} \mathbf{K} & \mathbf{P} \\ \mathbf{P}^T & \mathbf{0} \end{bmatrix} \in R^{(N+3) \times (N+3)} \quad (2-9)$$

Where $\mathbf{K}$ is the height matrix:

$$\mathbf{K} = \begin{bmatrix} U_{(r_{11})} & U_{(r_{12})} & I \\ U_{(r_{21})} & U_{(r_{22})} & I \\ \dots & \dots & U_{(r_{NN})} \end{bmatrix} \quad (2-10)$$

From the $\phi(p_i) = v_i$, we can know that:

$$\mathbf{Y} = \mathbf{L}(\Omega|m_0, m_1, m_2)^T \quad (2-11)$$

That is:

$$(\Omega|m_0, m_1, m_2)^T = \mathbf{L}^{-1} \mathbf{Y} \quad (2-12)$$

Once the parameter group of the formula is calculated, then the interpolation function $\phi(p)$ is known, as long as any point on the given plane can be interpolated to the target plane through the interpolation function. Figure 2-17 shows the example of a TPS-based fish warps.

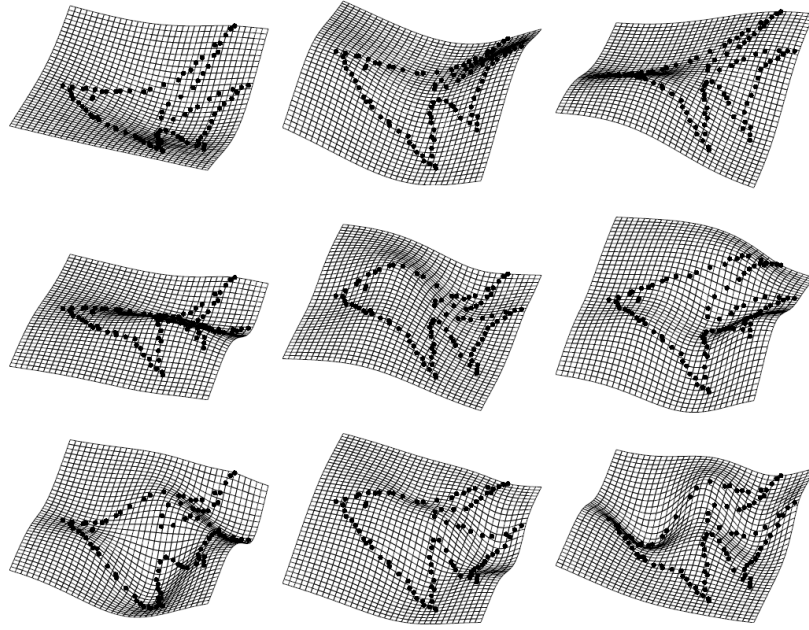


Figure 2-17 The example of a TPS-based fish warps (Ma et al., 2014).

Utilizing the TPS method for texture transfer and deformation allows for the preservation of intricate high-frequency texture details during the process. According to the selected image features, the image deformation processing techniques can be categorized into three types, i.e., grid-based image deformation, line segment-based image deformation and control point-based image deformation. For the grid-based deformation technique, Smythe (1990) proposed a deformation method that utilizes the idea of spatial transformation for grid distortion. Subsequently, Sederberg and Parry (1986) proposed grid-based free deformation (FFD) algorithm by moving grid points to increase the speed of deformation. Kho and Garland (2005) suggested deformation strategies for irregular polygonal meshes to control the deformation of the whole image. Blanco and Oliveira (2008) presented deformation techniques to reduce the mesh and handle it using feature curves to achieve fast mesh deformation. Yang and Feng (2009a) manipulated image deformation by topology-aware rigid grid. However, although grid-

based image deformation algorithms can realize image deformation, the deformation effect is usually uncontrollable. For this reason, subsequent researchers came up with line segment-based image deformation approaches. They defined the characteristic line segments in the original image as different characteristic coordinates and deformed them smoothly based on Sharp interpolation (Shepard, 1968) through the use of well-established alignment relations. This deformation technique usually has high computational complexity without being able to be generalized to a large number of practical applications. To simplify the computational process, Bookstein (1989) proposed a simple deformation method based on the interpolation of discrete features in the image on the basis of the thin plate spline function. Yang and Feng (2009b) processed image distortion through automatic feature matching and hierarchical interpolation. Igarashi and Igarashi (2009) present control points-based image deformation algorithms. They achieve very effective deformation by solving a set of joint linear Eq.s and minimizing a quadratic error metric. Weng et al. (2006) based on (Igarashi & Igarashi, 2009) introduced the use of nonlinear least squares optimization for image deformation. Schaefer et al. (2006) based on the least squares method (Levin, 1998) proposed image deformation algorithm based on moving least squares (MLS). Specifically, firstly, select the points of some feature elements as control points, then move the control points to get the mapping function of the image, and finally use the mapping function to realize the image deformation.

2.3 Generative Adversarial Networks

Traditional methods for image synthesis (e.g., those based on process generation, rule-based systems, and early machine learning techniques) have several limitations. They usually struggle to cope with the complexity of real-world images because of the amount of manual work required to define rules and parameters. Secondly poor generalization and limited creativity due to the inability of traditional methods to generalize well to new or unseen data. Advances in deep learning (LeCun et al., 2015), particularly generative adversarial networks (Goodfellow et al., 2014) and diffusion models (Ho et al., 2020), have addressed many of these limitations, leading to more realistic, scalable and flexible image synthesis.

2.3.1 Overview of GAN

Initially proposed by Goodfellow et al. (2014), Generative Adversarial Networks (GANs) are a type of deep learning framework created to tackle generative tasks by leveraging gradient descent optimization. In contrast to traditional end-to-end deep learning models, GANs consist of two networks: the Generator, which produces images, and the Discriminator, which distinguishes between generated images and those from the actual data distribution. The training process of GANs involves an adversarial dynamic. In Figure 2-18, the foundational GAN model is depicted as taking random noise inputs.

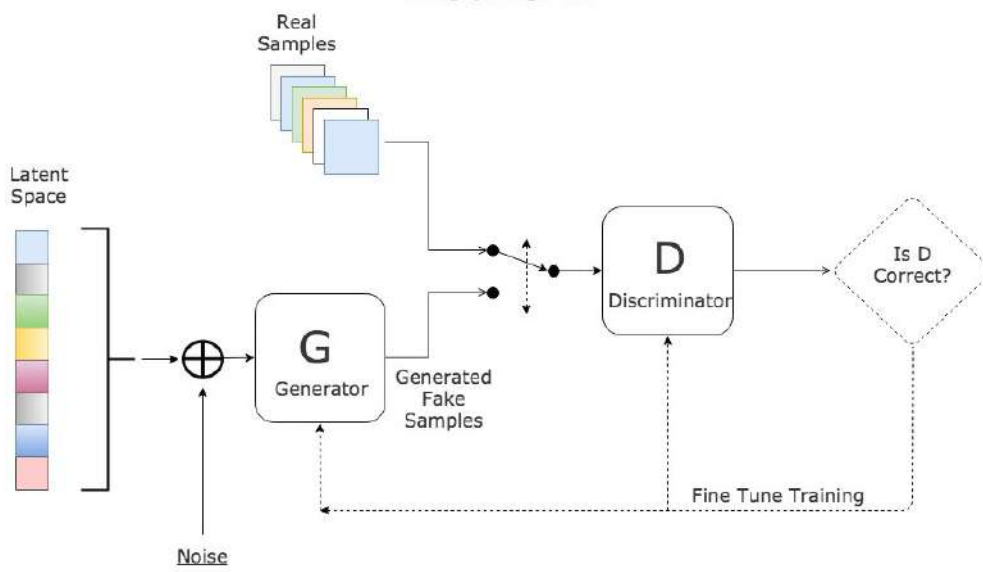


Figure 2-18 The Prototype of GAN.

In the context of GANs, where $G(\cdot, \theta_g)$ represents the generator and $D(\cdot, \theta_d)$ represents the discriminator with corresponding parameters θ_g and θ_d , the optimization objective can be expressed as:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim \mathbb{P}_d(x)} [\log(D(x))] + \mathbb{E}_{z \sim \mathbb{P}_z(z)} [\log(1 - D(G(z)))] \quad (2-13)$$

Here, x is sampled from real images, and z is sampled from a random noise distribution.

An intuitive interpretation of Eq. (2-13) is that the discriminator acts as a parameterized loss function for the generator. With fixed discriminator parameters, it operates akin to a binary classifier, distinguishing between real and generated images.

The optimization goal for the discriminator is to maximize the terms $\log(D(x))$ and $\log(1 - D(G(z)))$ initially. On the other hand, for the generator, the aim is to produce realistic images to deceive the discriminator.

2.3.2 Optimality of GAN

Goodfellow et al. (2014) demonstrated the optimality of the discriminator by expanding the expected value formulation of Eq. (2-13).

$$\begin{aligned} V(G, D) &= \int_x p_d(x) \log(D(x)) dx + \int_z p_g(z) \log(1 - D(G(z))) dz \\ &= \int_x p_d(x) \log(D(x)) + p_g(x) \log(1 - D(x)) dx \end{aligned} \quad (2-14)$$

Let $\frac{\partial V}{\partial D} = 0$, the discriminator D that optimizes the GAN framework is:

$$D_G^*(x) = \frac{p_d(x)}{p_d(x) + p_g(x)}. \quad (2-15)$$

By substituting the discriminator term in Eq. (2-13) with that in Eq. (2-15), the optimization process for the generator can be expressed as follows:

$$\begin{aligned} C(G) &= \max_D V(G, D) \\ &= \mathbb{E}_{x \sim p_d} \left[\log \frac{p_d(x)}{p_d(x) + p_g(x)} \right] + \mathbb{E}_{x \sim p_g} \left[\log \left(1 - \frac{p_d(x)}{p_d(x) + p_g(x)} \right) \right] \\ &= \mathbb{E}_{x \sim p_d} \left[\log \frac{p_d(x)}{p_d(x) + p_g(x)} \right] + \mathbb{E}_{x \sim p_g} \left[\log \frac{p_g(x)}{p_d(x) + p_g(x)} \right] \end{aligned} \quad (2-16)$$

This introduces the Kullback-Leibler Divergence (KLD), a commonly utilized metric for optimizing classification tasks:

$$D_{KL}(P||Q) = \mathbb{E}_{x \sim p_x} \log \frac{P(x)}{Q(x)}. \quad (2-17)$$

Thus, by substituting the term in Eq. (2-17) with the Kullback-Leibler Divergence (KLD), the resulting min-max game in the equation effectively becomes the optimization of the Jensen-Shannon Divergence (JSD):

$$\begin{aligned} C(G) &= D_{KL}(p_d \parallel p_d + p_g) + D_{KL}(p_g \parallel p_d + p_g) \\ &= -\log(4) + D_{KL}\left(p_d \parallel \frac{p_d + p_g}{2}\right) + D_{KL}\left(p_g \parallel \frac{p_d + p_g}{2}\right) \\ &= -\log(4) + 2 \cdot D_{JS}(p_d \parallel p_g) \end{aligned} \quad (2-18)$$

The Jensen-Shannon Divergence (JSD) ranges from 0 to 1. When the two probability

distributions are exactly the same, the JSD attains its minimum value of $-\log(4)$. Therefore, the lowest possible loss value for the generator is 0. If the discriminator and generator reach Nash Equilibrium, their loss values will converge and fluctuate around $-\log(4)$.

Despite various enhancements in training techniques, they have not addressed the fundamental issue with GANs: the optimization goal based on JSD is inadequate for measuring two completely distinct distributions. Arjovsky et al. (2017) highlighted this limitation and proposed a new loss function, the Wasserstein Distance:

$$W(P_d, P_g) = \inf_{\gamma \in \Pi(P_d, P_g)} \mathbb{E}_{(x,y) \sim \gamma} [\|x - y\|]. \quad (2-19)$$

The set $\Pi(P_d, P_g)$ comprises all joint distributions $\gamma(x, y)$ with marginal distributions being P_d and P_g . Yet, calculating the infimum in Eq.(2-19) is computationally challenging. By applying Kantorovich-Rubinstein duality, Eq. (2-19) can be reformulated.

$$W(P_d, P_g) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim P_d} [f(x)] - \mathbb{E}_{x \sim P_g} [f(x)]. \quad (2-20)$$

To fulfill this condition, the function f must be a 1-Lipschitz function, meaning that the discriminator needs to be Lipschitz Continuous. Consequently, the optimization objective for W-GAN is:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim P_d(x)} D(x) - \mathbb{E}_{z \sim P_z(z)} D(G(z)). \quad (2-21)$$

For a function f to be Lipschitz Continuous, there exists a constant L such that:

$$\|f(x) - f(y)\| \leq L\|x - y\| \Rightarrow \frac{\|f(x) - f(y)\|}{\|x - y\|} \leq L. \quad (2-22)$$

In a straightforward sense, the Lipschitz Constant L limits the steepness of f . Given that the discriminator is restricted, Eq. (2-22) becomes amenable to optimization. Driven by this rationale, researchers have employed various strategies to ensure Lipschitz Continuity.

Weight Clipping. In Arjovsky et al. (2017)'s suggestion, they introduced weight clipping as a solution. Their core concept revolved around bounding the model's parameters to constrain the output. Following each weight update, this approach involved limiting the updated weights to the range of $[-0.01, 0.01]$. However, they observed that momentum-based optimizers such as Adam were ineffective in this context. This inefficacy stemmed from the fact that weight clipping led to a concentration of weights around ± 0.01 . Despite its computational efficiency, weight clipping proved to be an unstable training method.

Gradient Penalty. Subsequently, Gulrajani et al. (2017) introduced gradient constraints to enforce Lipschitz Continuity. They achieved this by incorporating a gradient penalty term into Eq.(2-23).

$$\min_G \max_D V(G, D) = \mathbb{E}_x[D(x)] - \mathbb{E}_z[D(G(z))] + \lambda \mathbb{E}_{\hat{x}}[\|\nabla D(\hat{x})\|]. \quad (2-23)$$

Nevertheless, obtaining samples from the entire set of x can be challenging. To address this issue, they introduced an interpolation technique to simulate a sample $\hat{x}$:

$$\hat{x} = \alpha x_d + (1 - \alpha)x_g. \quad (2-24)$$

Randomly selecting x_d from P_d and x_g from P_g , they interpolated these samples using

α sampled from a uniform distribution $U(0, 1)$. Subsequently, they enforced gradient constraints on the discriminator by means of mean squared error, thus implementing the gradient penalty term. This method notably enhanced the training stability of GANs and could be optimized effectively using Adam.

Spectral Normalization. Faragallah et al. (2020) proposed the spectral normalization for GANs. Building upon the fundamental principles of Wasserstein GANs, they incorporated Lipschitz Constraints through spectral normalization, eliminating the need for input interpolation. They emphasized that the Lipschitz constant of function f is equivalent to its spectral norm when function f represents a network applying a linear transformation with weights W to an input hidden layer h .

$$\|f\|_{Lip} = \sup_h \sigma(\nabla f(h)) = \sup_h \sigma(W) = \sigma(W). \quad (2-25)$$

The spectral norm of a weight matrix W that transforms $h_{in} \rightarrow h_{out}$ is represented by $\sigma(W)$.

$$\sigma(W) := \max_{h: h \neq 0} \frac{\|Wh\|_2}{\|h\|_2} = \max_{\|h\|_2 \leq 1} \|Wh\|_2. \quad (2-26)$$

This implies that the spectral norm of W is equivalent to its largest singular value. By implementing spectral normalization on every layer within the discriminator:

$$W_{SN}(W) = \frac{W}{\sigma(W)}. \quad (2-27)$$

Consequently, the Lipschitz Constant will be constrained to one.

2.3.3 GAN for Generated Image/Texture Quality

While GANs have demonstrated promise in image generation, the quality of the images and textures they produce has been a matter of debate. The introduction of Wasserstein GAN in 2017 addressed the challenge of optimizing deep generators. But the task of generating high-quality images from noise vectors remained challenging. In 2018, Progressive GAN introduced a novel approach that involved cascading generators and discriminators of varying resolutions (Karras et al., 2018). The training process begins by training the GAN at lower resolutions until convergence, gradually integrating higher-resolution modules with lower-resolution ones.

Following this, StyleGAN introduced an innovative technique to translate input random noise vectors into understandable style codes, enhancing the detail in generated images (Karras et al., 2019), shown in Figure 2-19. They utilize a series of fully connected layers to decode the noise into a style code space denoted as w . Additionally, they incorporate various scales of noise, which introduce diverse random variations across different aspects.

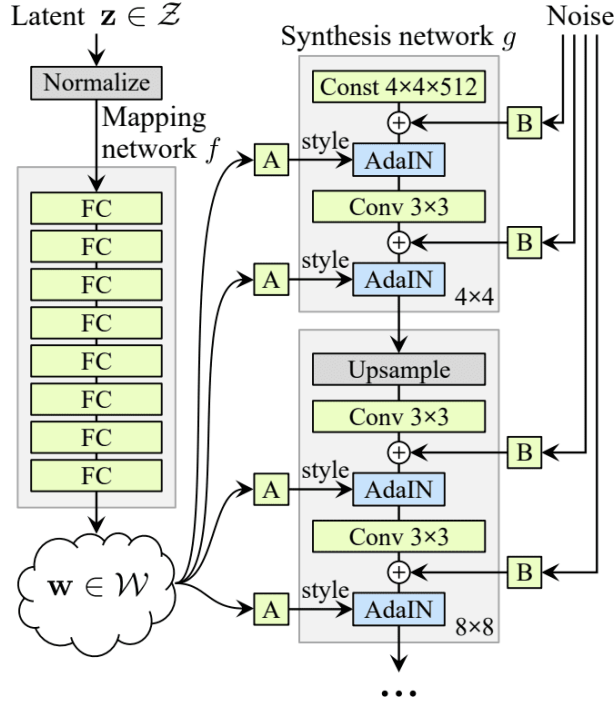


Figure 2-19 The StyleGAN framework (Karras et al., 2019).

Due to the powerful decoupling ability of StyleGAN, many researchers have introduced it to the texture domain. For example, Lin et al. (2023) integrated a multiscale texture broadcast module into the StyleGAN-V2 framework to enhance texture synthesis by injecting appropriate perceptual biases into the synthesis process, as shown in Figure 2-20.

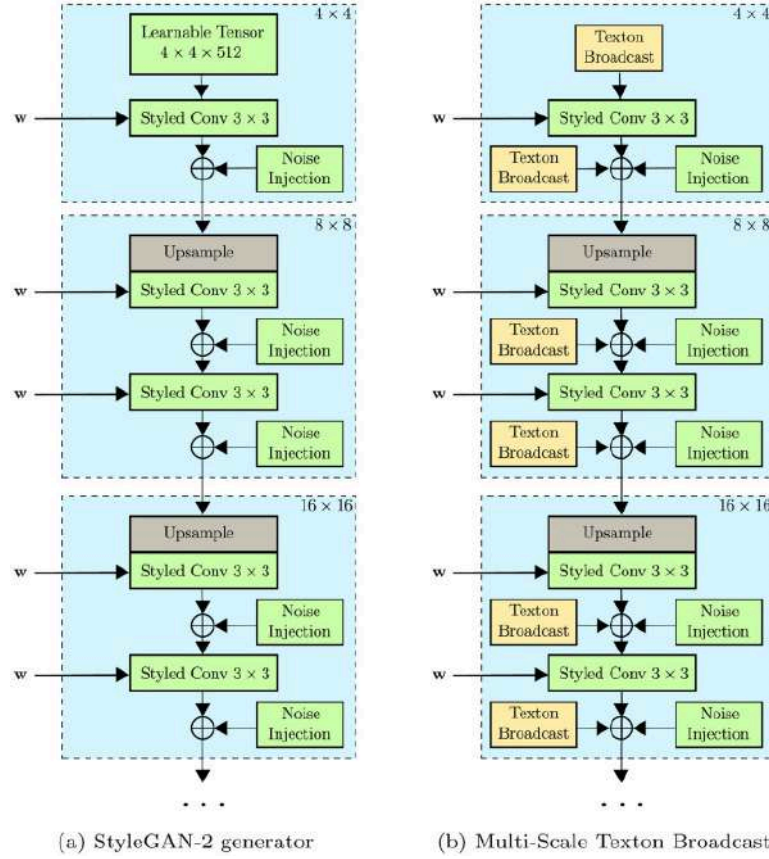


Figure 2-20 The texture generation framework is based on StyleGAN (Lin et al., 2023).

Aside from enhancing the generator, adjustments are made to the discriminator to enhance image quality. Isola et al. (2017) introduce PatchGAN in the image-to-image transformation task. Unlike a standard discriminator that predicts a single scalar to classify realness, PatchGAN focuses on analyzing image patches to assess authenticity, for example, Isola et al. (2017) set their patch size to 70×70 .

2.4 Augmented Reality

2.4.1 Development of Augmented Reality

AR technology has experienced rapid technological evolution and application expansion since it was proposed in the 1990s and has gradually become a hot topic in

cross-field research. Augmented Reality (AR) technology is the use of computers and other hardware devices to generate a realistic visual, hearing, force, touch and motion and other senses of the virtual environment, through the sensing device to make the experienter into the virtual environment (Cho & Schwarz, 2010), is the digital world superimposed on the real world direct interactive experience. Early research focused on the military and aviation fields, and then gradually expanded to a variety of industries such as healthcare, education, entertainment, and retail (Azuma, 1997). According to (Azuma, 1997), augmented reality should include three key components: the integration of real and virtual elements, 3D registration, and real-time interaction. In recent years, improvements in hardware and algorithm optimization have led to significant advancements in AR technology, particularly in image recognition, real-time rendering, and interactive experiences (Billinghurst et al., 2015). The popularity of mobile devices and the promotion of 5G networks provide infrastructure support for the widespread application of AR technology, enabling its rapid penetration in the consumer market (Carmigniani et al., 2011). At the technical level, the combination of deep learning and computer vision significantly improves the perception ability and user experience of AR systems. At the same time, the integration of AR with technologies such as the Internet of Things (IoT) and artificial intelligence (AI) further expands its application scenarios, such as smart homes, industrial manufacturing, and smart cities (Rauschnabel et al., 2022). Although AR technology has developed rapidly, research indicates that it still encounters challenges related to hardware costs, user privacy, and data security (Craig, 2013). Overall, AR technology combines information from both

the real and virtual worlds, enhancing users' perception and experience in a direct and immersive manner through the interaction between users and virtual images within real-world settings. Its cross-field application potential is huge, and it is expected to make greater breakthroughs in technological innovation and commercialization in the future.

2.4.2 Augmented Reality application for fashion industry

Augmented reality (AR) and virtual reality (VR) applications are growing rapidly and are increasingly being used in the fashion industry (Javornik, 2016). Its technology is characterized by narrowing the boundary between the real world and the virtual world, stimulating consumers' senses, enabling the imagination of the virtual world to place consumers in a new interactive environment through the combination of the real world and the digital world. The impact of AR technology in the fashion industry covers design, production, marketing, consumption and sustainable development. For instance, AR has greatly enhanced the shopping experience by offering virtual try-ons and personalized recommendations, which have helped decrease return rates and support the integration of online and offline retail models (Kim et al., 2023). In the design phase, AR allows designers to refine their creations through virtual modeling and real-time adjustments, thereby lowering the costs associated with physical samples and boosting design efficiency (Lee et al., 2020). From a marketing perspective, AR-based advertising and virtual fashion shows offer brands a more engaging and immersive promotional approach, increasing consumer interaction and brand loyalty (Hilken et al., 2022). As a supplement to traditional fashion e-commerce, the application of AR can improve consumers' perceived ownership on e-commerce platforms and build a bridge

between offline and online retail (Papagiannidis et al., 2014). Since 2000, AR Virtual Try-On has been available providing an exciting impressions to fashion industry from the utilization of this technology (Rigby, 2013). Through AR camera scanning and tracking the customer's body shape and pose, they can have the fitting experience of trying on clothes in the physical fitting room (Kang, 2014). Since the AR camera acts as a virtual mirror, users can quickly try on clothing and decide whether it is appropriate for them in terms of color, size, style. As a result, customers can try on more clothes in less time (Erra & Colonnese, 2015). In addition, AR based virtual mirrors capture photos of their outfits from any angle, compare them with other clothing, or share their try-on effects with others through social applications.(Berman, 2019; Pantano & Gandini, 2017). Virtual fashion has fundamentally reshaped the way fashion is created, consumed and experienced (Santos et al., 2020). This digital evolution has been welcomed by well-known brands and innovative startups. Notable examples include Nike and Adidas, which have utilized AR to enhance the shopping experience, thereby boosting consumer satisfaction (Yim & Park, 2019). AR-compatible Gucci Virtual 25 sneaker released in 2021 and Burberry's Roblox version of the Lola handbag in 2022, which featured unconventional materials such as clouds, water, and wild leaves (Loureiro, 2023). These initiatives mark a significant shift towards a digital economy, with numerous startups leading innovative approaches. For instance, The Fabricant, established in 2018 as the first digital fashion house, showcases the potential for a fully virtual fashion supply chain that delivers dematerialized, personalized digital experiences (Casciani et al., 2022; Särämäkari, 2021). Similarly, RTFKT has created a

robust digital ecosystem that involves NFTs, metaverse games and virtual fashion collectibles. Dress X enhances the digital fashion experience by offering digital try-ons, which not only support online self-expression but also provide a sustainable alternative to frequent wardrobe changes (Hollensen et al., 2022; Roberts-Islam, 2020).

Moreover, AR technology contributes significantly to the sustainable development of the fashion sector by minimizing resource waste and extending product lifespans, aiding the industry's shift towards environmental sustainability (Kim et al., 2023). Research also highlights AR's expanding role in education and training, social interactions, and luxury goods, providing new platforms for learning and experiences for both designers and consumers (Rauschnabel et al., 2019). Figure 2-21 is an example of an haute couture virtual try-on based on AR technology. Pantano (2015) believes that the latest technological innovations can bring more profits to the fashion industry by improving the shopping experience of consumers and encourage companies and fashion brands to introduce new technologies in their online or offline stores to improve the shopping experience and upgrade the store environment. According to a 2023 Nielsen Global report, more than 50 percent of respondents are prepared to use this technology to help retail. Overall, AR technology is significantly altering how the fashion industry operates. It enhances efficiency and innovation, while also offering consumers a more personalized, convenient, and immersive virtual try-on experience. This transformation is a key driver in advancing the industry's digital evolution.



Figure 2-21 An example of an haute couture virtual try-on based on AR technology.

2.5 Chapter Summary

This chapter systematically classifies virtual try-on technologies according to their different inputs, then conducts an in-depth exploration and discussion of existing 2D VTON and 3D VTON methods, and summarizes the research gaps of the two try-on schemes (refer to Section 1.2.1 and Section 1.2.2 for details). This lays a solid foundation for the new try-on scheme proposed in this study.

Considering that 2D VTON is widely regarded as a computer vision task in the deep learning community, this chapter defines the task of fashion image understanding and reviews the related methods of fashion image detection and image deformation. The goal of fashion image detection is to recognize and pinpoint clothing and accessories within an image. Utilizing computer vision methods such as object detection and segmentation can help pinpoint critical areas. Accurate detection ensures that the virtual try-on system synthesizes clothing realistically onto the user's image. This chapter elaborates on common fashion image detection methods and analyzes the limitations of

these aspects, preparing for the subsequent proposed **fashion landmark localization model** (*Model i*). The image deformation task is to adjust the detected garments to fit the users body shape and posture. The technique involves algorithms that manipulate garment images to match the user's size, maintaining the natural appearance of the garment. This chapter delves into traditional fashion image deformation algorithms. Drawing inspiration from these algorithms, a **body-aware garment deformation model** (*Model ii*) will be developed, offering an efficient solution for a more realistic and customized virtual try-on experience.

In addition, the implementation of 3D VTON method includes two core technologies: 3D Avatar and 3D garment reconstruction. This chapter first gives a comprehensive review of the previously implemented methods. Secondly, the 3D VTON methods based on physical simulation and sewing-pattern reconstruction are reviewed, because these methods can simulate the realistic draping effect of 3D garment on the body according to the user's body shape. This study draws on these methods to propose a **cloth digitization pipeline** (*Model iv*) for 3D VTON. Then, considering that the existing 3D VTON methods mainly focus on creating 3D avatars or 3D garments with realistic shapes and geometries while ignoring the reconstruction of garment surface textures, this chapter also reviews the generative adversarial network and its progress in output image quality. This provides a detailed method introduction for the subsequent newly proposed **tileable texture generation** (*Model iii*).

Finally, AR technology is also reviewed to prepare for augmented reality (AR) 3D VTON fashion applications. This trajectory extends to immersive online retail supported by AR technology, bridging the gap between offline and online shopping by offering virtual try-on options for shoes, accessories, clothing, and jewelry.

Chapter 3. Fashion Landmark Localization

3.1 Introduction of fashion landmark

Locating fashion items on any input images is often a preliminary or initial step that supports various other vision tasks, such as clothing image retrieval (Kuang et al., 2019; Wang et al.), clothing category classification (He et al., 2016), attribute recognition (Fu et al., 2013; Sun et al., 2015), clothing segmentation (Chen et al., 2016; Li et al., 2021b), and clothing generation (Nakamura & Goto, 2018).

These image-based vision tasks in fashion domain are also referred to as 'fashion image understanding', an active research field in recent years. Understanding fashion images has extended practical advantages and can be applied in enormous applications in fashion and other domains. The study of fashion image understanding has widely benefited from the rapid advancements in deep learning-based new models and the availability of large datasets covering real-world fashion images. However, the real value of these datasets is in doubt, because different annotation strategies have been adopted in these datasets, resulting in different landmarks for different clothing styles and low compatibility of these datasets. For example, Ge et al. (2019) only created a data-rich dataset DeepFashion2 (Ge et al., 2019) to solve problems, after he refutes the existing problems, such as single clothing-item per image, sparse landmarks (4~8 only), and no per-pixel masks, in DeepFashion (Liu et al., 2016a). Data annotation is

incompatible between these two datasets serving similar problems, which results in a drain on financial resources and computing power.

This chapter proposed an automatic method to locate key points of fashion items on fashion images by taking advantage of clear correspondence between clothing and the human body, called fashion landmark localization model (the proposed *Model i*). *Model i* has a powerful ability to locate landmarks in functional areas of clothing, such as the neckline and sleeve landmarks. Figure 3-1 and Table 3-1 show the definition and naming of *Model i* fashion landmarks.



Index	Landmark Name
0	neckline_left
1	neckline_center
2	neckline_right
3	shoulder_left
4	shoulder_right
5	armpit_left
6	armpit_right
7	waistline_left
8	waistline_right
9	hem_left
10	hem_center
11	hem_right
12	sleeve_left_out
13	sleeve_right_out
14	sleeve_left_in
15	sleeve_right_in

Figure 3-1
Illustration of the proposed landmark's definition.

Table 3-1
Naming of landmarks.

Model i consists of the initial mask extraction, the clothing mask completion, coarse clothing key-point sampling and a refined clothing key-point detector. Figure 3-2 shows the workflow of the proposed *Model i*'s framework.

Specifically, since most clothing key points are along edges/outline of the garment, coarse corners detector and morphology algorithm are developed to localize possible corner points on the clothing mask based on human parsing results. However, the initial points detected may not be valid corners along clothing boundary but points inside the clothing mask because of possible occlusions, the clothing mask completer (CMC) is therefore designed to help automatically complete occluded regions in the phase of the clothing mask completion. Finally, the graph-based clothing feature lines, human pose information as a priori and inherent clothing-body location constraint are used in conjunction to refine the location of detected clothing key points. Coarse points are divided into 6 groups in this study; they are the collar part points, the shoulder part points, the armpit part points, the waist points, the clothing hem points and the sleeve points.

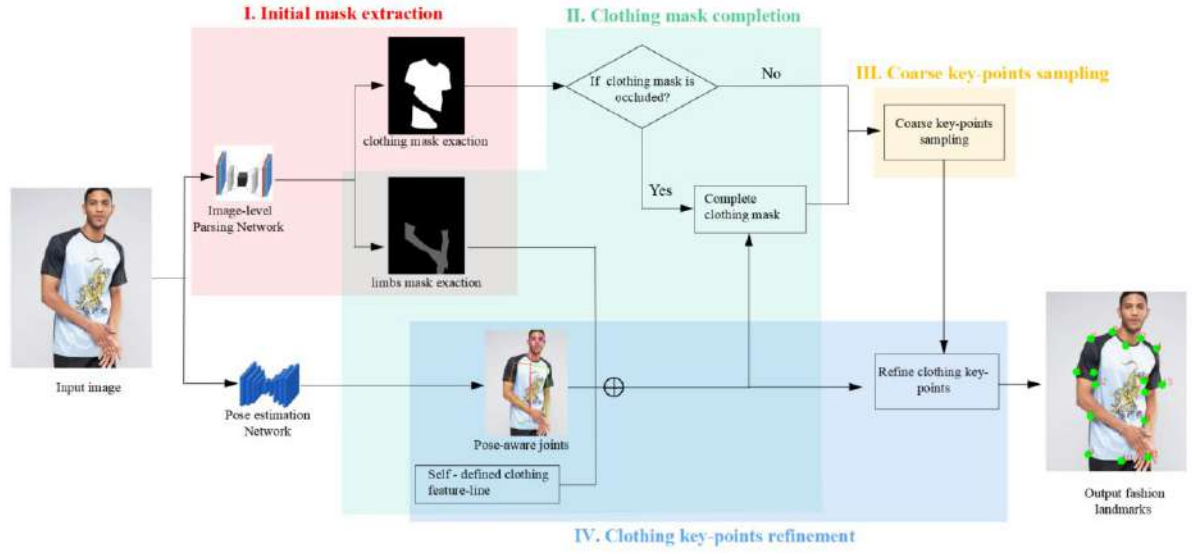


Figure 3-2 The overall framework of the fashion landmarks localization module.

3.2 Related Work

The purpose of landmarks is to locate the neckline point, shoulder point and other functional areas of clothing, helping to effectively identify the layout and extensive attributes of fashion items. Recent years, thanks to the public release of the large dataset (Ge et al., 2019; Liu et al., 2016a) for accelerating the research on landmark detection task. Different methods are proposed by researchers in succession based on task dependencies. Fashion landmark detection tasks can be roughly divided into the following categories based on the deep learning model they use. They are regression-based fashion landmark detection, attention model-based fashion landmark detection, constraint-based methods and pose estimation-based fashion landmark detection methods

Regression-based fashion landmark detection. Regression-based methods address the task of locating fashion landmarks by considering it as a regression problem. It involves training deep neural networks that utilize regression functions to associate an image with a set of landmarks. The objective is to reduce the mean square error across all instances. For example, first proposed FashionNet to predict clothing attributes and landmarks to understand fashion images, and they introduced a richly annotated dataset named DeepFashion (Liu et al., 2016a) as a benchmark dataset for multiple tasks. Following FashionNet, Liu et al. (2016c) proposed a framework of cascading to improve fashion landmarks estimation accuracy while reducing the amount of computation. It's a three-stage structure using VGG-16 (Simonyan & Zisserman, 2014) in each stage to predict landmarks, estimate fashion landmark offsets, and output contextual landmark offsets respectively. These methods treat Fashion Landmarks Localization as the regression problems by training deep neural networks and using regression functions to minimize the mean square error of all N samples.

Fashion landmark detection based on attention mechanism. Besides the regression-based model, attention mechanism is a widely used method in landmarks detection. For example, Liu and Lu (2018) introduced an attention mechanism, called 'landmark heatmaps' to produce a unified attention. Their proposed network adds two branches based on VGG-16 to capture attention and localize landmarks. Li et al. (2019c) proposed a Spatial-Aware Non-Local model to detect fashion landmarks. This model

can accept high-level semantic knowledge in the way of encoding the prior knowledge by the spatial attention map. Following this, Lee et al. (2019b) detected landmarks by utilizing a global-local embedding model that contains a global procedure and a convolutional procedure to obtain long-range dependencies and local information in fashion images respectively.

Constraint-based fashion landmark detection. Furthermore, it is advisable to explore constraint-based methods to enhance the effectiveness of fashion landmark detection. For example, Yu et al. (2019c) attempted to explore the layout relationships among fashion landmarks through layout graph inference. Initially, the model maps image features to nodes within the graph. Next, graph inference is employed to determine global layout relationships. Finally, the graph nodes are mapped back to enhance the features. Furthermore, Wang et al. (2018d) designed the BCRNN model to receive and process high-level semantic knowledge, which utilizes global syntactic constraints to output the layout of fashionable landmarks.

Fashion landmark detection based on pose estimation. Besides the aforementioned models, the pose estimation task is akin to fashion landmark detection, as both involve identifying key feature points in an image that can represent a person or a garment. For instance, Hu et al. (2018b) introduced a fashion attitude machine (FPM) designed for

fashion landmark detection. It seeks to accurately identify fashion landmarks without relying on prior knowledge of body structure, even when the clothing in the image is affected by the model's pose.

The above methods are deep learning based on a large number of data-driven ways for feature extraction. The learned feature interpretability is weak, and some unusual landmarks are predicted to be incomprehensible although their overall performance is very good. It is argued that the task of human pose estimation is similar to fashion key points localization but differs in the location and number of points. Unlike the above methods, a pose-aware segmentation-based method is proposed in this study to locate fashion landmarks by taking advantage of the clear correspondence between clothing and the human body, as inspired by (Cao et al., 2017; Hu et al., 2018b). The segmented image is leveraged to extract the clothing mask, and the joint points of human pose estimation are used as prior knowledge in the proposed methods to roughly locate the different functional areas of clothing.

3.3 Method

3.3.1 Initial Mask Extraction

Since the fashion landmarks are along the edge/outline of the garment, it is an important step to quickly and accurately detect the edge lines of each fashion item for the initial key points sampling. Considering the mixture of foreground and background of the

input image, and each fashion item is difficult to be automatically distinguished, the clothing and torso areas are thus extracted separately based on the segmentation method. The performance of human parsing for 12 state-of-the-arts on the LIP (Liang et al., 2018) validation set was reported in Table 3-2, to help us pick out the best method to extract the accuracy initial clothing masks. The LIP dataset was chosen because it contains the most training images and annotates 19 types of labels. Based on these advantages, LIP has become the gold standard benchmark for human parsing. Mean Intersection over Union (mIoU) was used as a standard metric for image semantic segmentation.

Table 3-2 Comparison of human parsing performance with 12 SOTA on the LIP validation set.

State-of-the-arts Methods	mIoU (%)
DeepLab † (Chen et al., 2017)	44.80
SSL † (Gong et al., 2017)	46.19
MMAN (Luo et al., 2018)	46.81
JPPNet † (Liang et al., 2018)	51.37
DeepLabv3+ (Chen et al., 2018)	52.09
CE2P (Ruan et al., 2019)	53.10
CorrPM (Zhang et al., 2020b)	55.33
PCNet (Zhang et al., 2020a)	57.03
OCR (Yuan et al., 2020)	55.60
SLRS (Li et al., 2020a)	56.34
CNIF † (Wang et al., 2019b)	57.74

SCHP † (Li et al., 2021b)	58.62
SCHP	59.36

*The symbol † marks the multiple scales testing results.

As we can see, SCHP achieves the best results with 58.62% mIoU, and 59.36% mIoU when using data augmentation techniques for all the validation data. For pixel-level annotations in human parsing, there are many incorrect annotations in the ground truth masks, especially the fine-grained LIP annotations (there are 19 annotations). The SCHP is finally used because this model can iteratively aggregate the current learned model and the previous optimal model in an online manner, thereby inferring more reliable pseudo masks, and the corresponding corrected labels can further improve the performance of the model. However, the predicted segmented area is not the 100% ground truth because of the blurred boundaries and similar pixel values between different semantic parts, which can cause confusion between the semantics of the image as well as the unexpected noise that can be generated in the predicted mask. So, the initial mask was constructed by the morphological closing operator (as in Eq. (3-1)) based on the preservation of their critical geometry and topology. The closing of A by B is obtained by the dilation of A by B, followed by erosion of the resulting structure by B:

$$A \bullet B = (A \oplus B) \ominus B \quad (3-1)$$

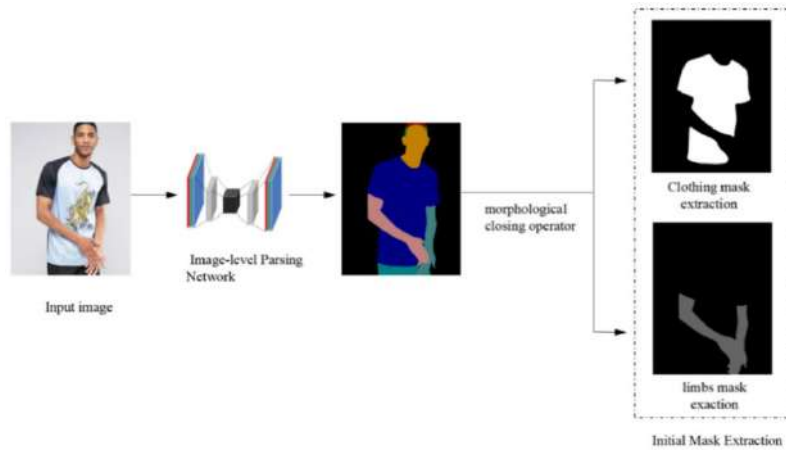


Figure 3-3 The process of the initial mask extraction.

3.3.2 Clothing Mask Completion

The problem of occlusion is often encountered in the fashion landmarks detection. If the clothing mask is occluded, the contour points inside the clothing will be recognized as edge points. A discriminator is then designed to determine if the initial masks are occluded or not. If the mask is not occluded, the proposed system will continue with the process of initial coarse key-points sampling (Section 3.3.3). Conversely, Clothing Mask Completer (CMC) is designed to automatically locate and complete occluded regions. It includes three modules (1). Pose-aware joints information, (2). Clothing feature line self-defined according to human body proportion and clothing structure, (3). The new mask was created in the occluded area.

Discriminator. The discriminator is designed to determine if the clothing mask obtained is occluded. The number of connected components in the image is calculated and defined as n , $n \in \mathbb{Z}^+$. Figure 3-4 shows the number of connected components in different fashion poses.

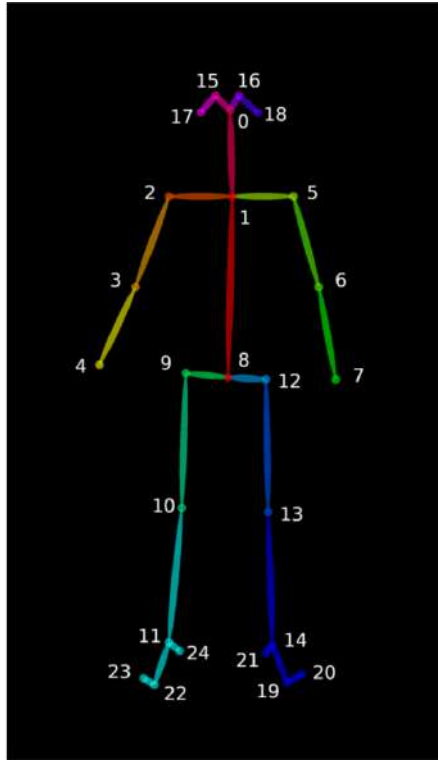


Figure 3-4 Examples of connected components with different poses: $n=1$ in (a) and (b), $n=2$ in (c), $n=3$ in (d).

If $N > 1$, the clothing mask is occluded. At $N = 1$, the minimum external rectangle of the clothing mask was calculated, and the position coordinates of the midpoints M and N in the width direction of the rectangle were found. The line LMN divides the rectangle into A and B . Calculate the area difference of A and B , $SC = |SA - SB|$, and then compare SC with the threshold T . If $SC > T$ the clothing mask is occluded, otherwise, it is not.

Clothing Mask Completer (CMC). CMC is designed to complete these covered areas, which includes 3 modules.

Module1. Pose-aware joints information. Pose gives better global information in the form of connecting joints that can provide a reference standard for locating fashion landmarks. A state-of-the-art pose estimation method (Cao et al., 2017) is used in this module of CMC to obtain joint information. Figure 3-5 shows the output format of anatomical key-points or joints and their definitions.



label	Definition
0	Nose
1	FNP
2	R_shoulder
3	R_elbow
4	R_wrist
5	L_shoulder
6	L_elbow
7	L_wrist
8	Center_hip
9	R_hip
10	R_knee
11	R_ankle
12	L_hip
13	L_knee
14	L_ankle
15	R_eye
16	L_eye
17	R_ear
18	L_ear
19	L_foot_toe
20	L_foot_left_toe
21	L_foot_heel
22	R_foot_toe
23	R_foot_right_toe
24	R_foot_heel

Figure 3-5 Joints definition of pose estimation (Cao et al., 2017).

Module2. Clothing feature line self-defined according to human body proportion and clothing structure. Based on the results of pose estimation, define the shoulder-line, hipline by connecting point R_shoulder, point L_shoulder and point R_hip, L_hip respectively. The position of the bustline and waistline is then estimated based on the ‘head-to-body ratio’. The distance between the point FNP and point Center_hip is defined as D_{sh} . If the head length is a , $D_{sp} = 2.5 a$. The approximate position of the shoulder-line, bustline, waistline, and hipline is shown in Figure 3-6. In addition, the width of the clothing is constrained according to the shoulder-waist ratio (SWR), bust-waist ratio (BWR) and the waist-hip ratio (WHR).

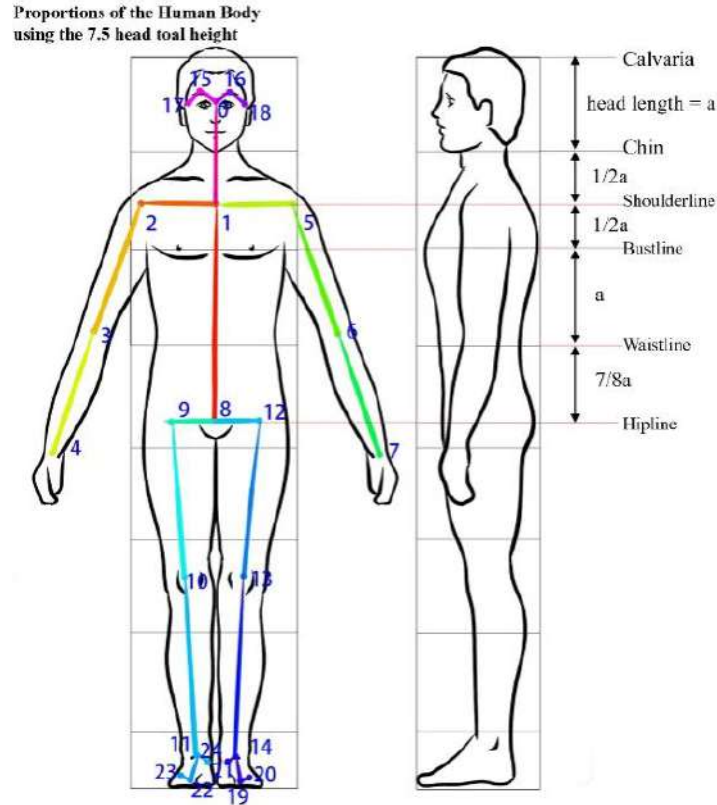


Figure 3-6 Illustration of estimated joints and self-defined clothing feature line.

Module3. The new mask created in the occluded area. According to the common fashion poses, the clothing mask patches (occluded areas) are distributed in the area of hair, limbs and decorations (e.g., bags). Hair usually covers the collar and shoulder area of the clothing, and limbs and ornaments usually cover the sleeves and large body area of the clothing. Module 3 of CMC is used to complete these occluded areas, including the collar area, the sleeve area, and the main torso area.

Collar Area. B-spline curves and V-shaped polylines are used to fit the collar arc according to the shape of the collar (round neck, V-neck). These curves or lines connect the left and right shoulder joints to create a mask called a collar patch mask, shown in

Figure 3-7 (a).

Sleeve Area. Create a new rectangular mask at the sleeve to complete the occluded clothing sleeve area. The length and width of the rectangular mask are constrained by the length of the clothing sleeves and the width of the human arm respectively, shown in the blue area of Figure 3-7 (b).

Main Torso Area. The S1, S2, S3, and S4 areas shown in Figure 3-7 (c) are used to complete the occluded clothing main torso area. First, the European distance between the left and right shoulder joints is calculated, defined as b . The SWR, BWR, WHR are used to roughly constrain the width of the clothing. The ‘head-to-body ratio’ is used to constrain the length of the clothing.

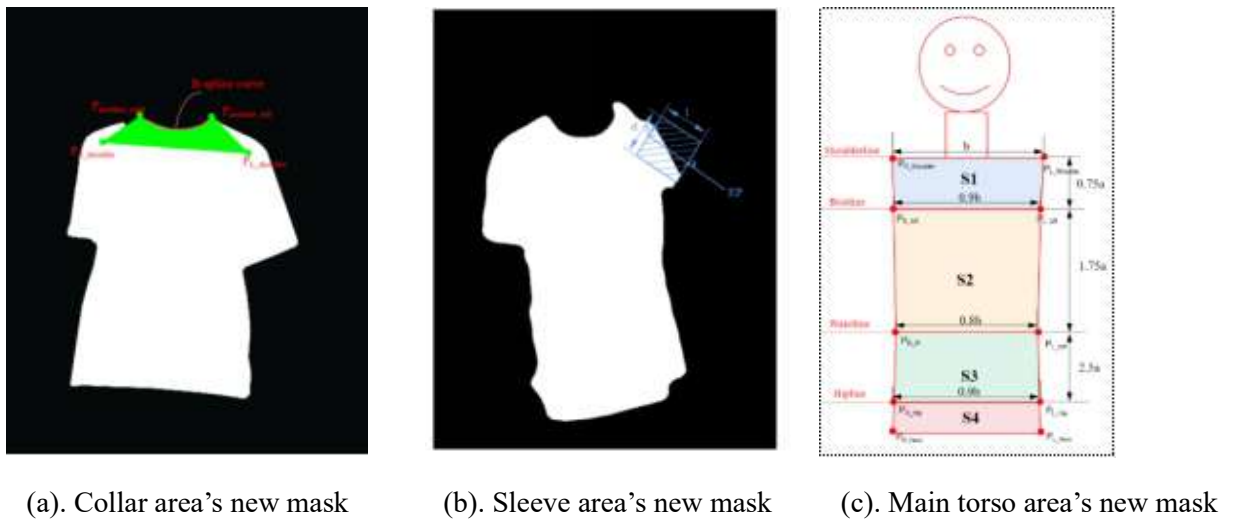


Figure 3-7 Illustration of the new mask created in the occluded area.

3.3.3 Coarse Clothing Key-points Sampling

In this section, the purpose is to sample key-points based on clothing masks obtained

in Section 3.3.2. The required clothing key-points are distributed on the contour line of the clothing mask. Considering that one of the properties of ‘corners’ is that the movement in any direction is large, the Corner Detection Algorithm is the priority. After comparing several classical corner detection algorithms, Shi-Tomasi corner detection (Kenney et al., 2005) is used for the following reasons: 1. Our task needs to return the fixed N best corners and set the Euclidean distance between the N corners. 2. The Shi-Tomasi corner detection operator is an improvement of the Harris corner detection operator, which is suitable for processing cloth masks in terms of stability, anti-noise capability and complexity. The preliminary experimental results were unsatisfactory when these garment outline points were used, because shape-related matching methods are not garment-specific, too many sampling points will usually lead to cross-matching and partial shifts (see Figure 3-8). Given small errors in the matching can cause significant distortion when deforming the image, which further impacts negatively on the subsequent experimental results. Therefore, functional points representing key parts of the garment are selected based on the initial sampling points. For initial sampling points, define S_{GOP} as the set of n points, where $S_{GOP} = \{s_1, s_2, \dots, s_n\}$, $s_i = (x, y)$.

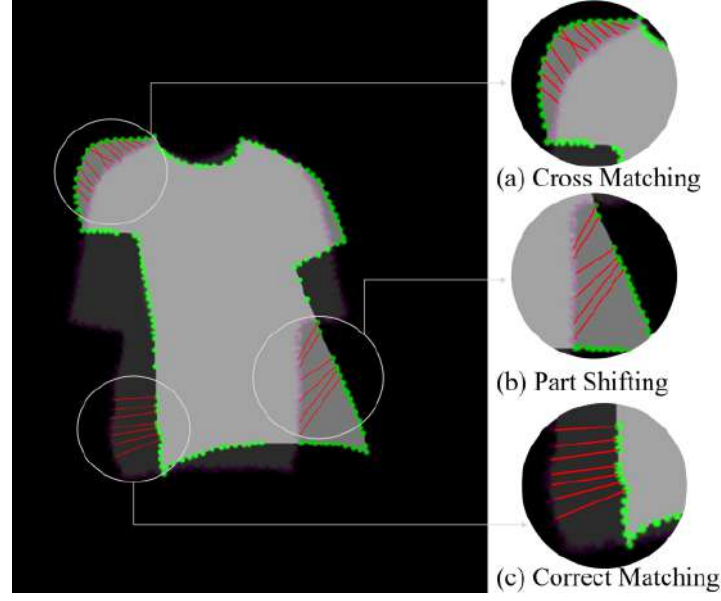


Figure 3-8 Matching results when using garment outline points.

3.3.4 Clothing Key-points Refinement

Recent works study on pose estimation provided the human joint points as their output, which can provide a vital positional prior. Inspired by them, a pose-aware approach is adopted to group and refine garment outline points into fashion landmarks by explicitly modeling the close relationship between the human body and garment. The pose estimation method introduced by (Cao et al., 2017) is adopted to detect the human joints. For each input image, 25 joint points J were obtained first, as shown in Figure 3-5. J containing head, neck, shoulder, elbow, wrist, hip, knee and ankle joints. In addition, the clothing feature lines were defined according to the human body proportions and clothing structure. The shoulder-line, hipline by connecting right shoulder point j_2 , left shoulder point j_5 and right hip point j_9 , left hip point j_{12} , respectively. The positions of the bustline and waistline are then estimated based on the ‘head-to-body ratio’. The distance between the front neck point j_1 and center hip point j_8 is defined as $\|j_1j_8\|$. If

the head length is a , $\|j_1 j_8\| = \frac{19}{8}a$. The width of the clothing is constrained according to the shoulder waist ratio (SWR), bust-waist ratio (BWR) and the waist-hip ratio (WHR). Figure 3-6 shows the approximate positions of the shoulder-line, bustline, waistline, and hipline. Taking the 25 joints, J as the reference points and the clothing feature lines as the reference lines, the garment outline points S_{GOP} are divided into 6 subsets according to the human joints and clothing feature lines to which they belong. Specifically, we first define 4 subsets, denoted by S_{collar} , $S_{shoulder}$, S_{sleeve} , and $S_{hem} \in S_{GOP}$, which are located around the j_1, j_2 and j_5, j_3 and j_6, j_9 and j_{12} . S_{collar} , $S_{shoulder}$, S_{sleeve} , and S_{hem} represent the set of outline points in the collar N_{cp} , shoulder N_{shp} , sleeve N_{slp} , and hem region N_{hp} respectively. In addition, another 2 subsets denoted by S_{bust} and $S_{waist} \in S_{GOP}$, which are defined located around the clothing bustline, and waistline respectively. S_{bust} and S_{waist} represent the set of coarse points in bust N_{bp} and waist region N_{wp} .

Collar Landmark Localization.

First, S_{GOP} is constrained by means of $\|j_2 j_5\|$ to identify S_{collar} , with P_m and P_n defined as the left and right endpoints of S_{collar} . Then, an angle regularization of the garment mask is performed by rotating θ to horizontalize the shoulder line.

Next, the collar center landmark is determined by calculating the minimum Euclidean distance between S_{collar} and j_1 . The collar left landmark, or collar right landmark is determined by calculating the minimum value of $s_i(y)$ in the S_{collar} . Figure 3-9 visualizes collar landmark localization, and the corresponding algorithm is listed in

Algorithm 3-1

Algorithm 3-1 Collar landmark localization.

Input: set $S_{GOP} = \{s_1, s_2, \dots, s_n\}$, $s_i = (x, y)$ of n garment outline points,
left endpoint of S_{collar} : P_n and left shoulder joint: j_5

Output: center, left, and right landmarks of collar L_m , L_l , and L_r

- 1 Determine set S_{GOP} to represent collar region points
- 2 **for** $s_i \in S_{CKP}$ **do**
- 3 **if** $\|s_i - P_n\| \leq \|j_5 - P_n\|$ **then**
- 4 $s_i \in S_{collr}$
- 5 **end**
- 6 **end**
- 7 Determine the middle point of collar: $L_m \leftarrow \min_{s_i} \|s_i - P_n\|$, $s_i \in S_{collar}$
- 8 Determine the rotation angle: $\theta \leftarrow -\arccos \frac{\langle \vec{j_5}, \vec{P_n} \rangle \cdot (1,0)}{\|\vec{j_5}, \vec{P_n}\| \times \|(1,0)\|}$
- 9 Apply rotation transformation to S_{collar}
- 10 **for** $s_i \in S_{collar}$ **do**
- 11 $s_i \leftarrow \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} S_i^T$
- 12 **end**
- 13 Split S_{collar} to left and right subset $S_{collarL}$ and $S_{collarR}$
- 14 **for** $s_i \in S_{collar}$ **do**
- 15 **if** $s_i(x) \leq P_n(x)$ **then**
- 16 $s_i \in S_{collarL}$
- 17 **else**
- 18 $s_i \in S_{collarR}$
- 19 **end**
- 20 **end**
- 21 Determine the left collar point: $L_l \leftarrow \min_{s_i} s_i(y)$, $s_i \in S_{collarL}$
- 22 Determine the right collar point: $L_r \leftarrow \min_{s_i} s_i(y)$, $s_i \in S_{collarR}$

Shoulder Landmark Localization. $D_{ss}(i)$ is defined as the Euclidean distance from

$S_{shoulder}$ to j_2 and j_5 , where i is the i -th outline point in the $S_{shoulder}$. The left and right shoulder landmarks are located by calculating the minimum $D_{ss}(i)$ as follows:

$$D_{ss}(i) = \|S_{shoulder}(i) - j_2\|, \forall i \in \{1, 2, \dots, N_{shp}\} \quad (3-2)$$

$$D_{ss}(i) = \|S_{shoulder}(i) - j_5\|, \forall i \in \{1, 2, \dots, N_{shp}\} \quad (3-3)$$

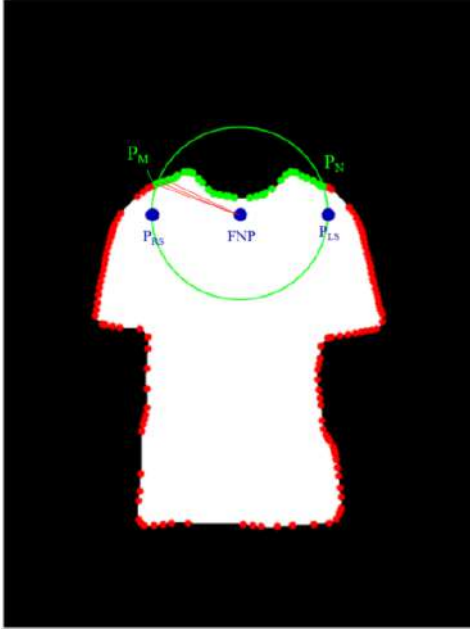


Figure 3-9 Collar landmarks localization. Figure 3-10 Shoulder landmarks localization.

Hem & Sleeve Landmark Localization. The Ramer-Douglas-Peucker Algorithm (Douglas & Peucker, 1973) was used to effectively reduce the number of points required to represent the garment outline curve while preserving the garment geometry. Given the basic structure of the garment, with clear corners at sleeves and hem, applying this algorithm to the garment mask, the range of S_{hem} and S_{sleeve} will be reduced but not eliminated. Then j_9, j_{12} and j_3, j_6 are used as spatial constraint prior to further narrowing the ranges of the S_{hem} and S_{sleeve} . Define the reduced S_{hem} as $S'_{hem} = \{s'_1, s'_2, \dots, s'_m\}$, $s'_i = (x, y)$ is m garment outline points ($m \leq 10$ in the

experiment). Select two points s'_1 and s'_m to be connected as a straight-line L_{hem} . The final hem landmark (e.g. left hem landmark) satisfies any point s'_i of S'_{hem} that is the maximum distance to L_{hem} . In the same way, the remaining hem landmarks and sleeve landmarks are automatically obtained.

Armpit & lower bust landmark location. The armpit and lower bust landmarks are distributed on both sides of the clothing bustline $\pm 0.25a$. And then the BWR of the human body is used as a geometric constraint to locate the specific locations of armpit and lower bust landmarks.

Higher waist & Waist Landmark location. The higher waist and waist landmarks are distributed on along $0.25a$ above the clothing waistline and on the waistline respectively. The clothing waistline is translated up $0.25a$ and extended, with its intersection with the garment mask side seam representing the left and right higher waist landmarks, respectively. The intersection of the extended clothing waistline and the garment mask side seam represents the left and right waist landmarks respectively. Furthermore, to avoid the intersection points being calculated as part of the sleeve area, BWR and WHR are used as geometric constraints in the width direction. In the experiment, the male's BWR is set to 0.7 and WHR to 0.9, while the female's BWR is set to 1.3 and WHR to 0.8 (Cloud & Perilloux, 2015).

3.4 Experiments

Since the motivation for researching this task is different from the method of (Ge et al., 2019) method based on deep learning. The proposed method is relatively more biased towards traditional computer vision tasks. The purpose of the experiment is not to beat DL-based method in performance, but to verify the rationality of landmarks in the follow-up task (interactive annotation tool). Qualitative and quantitative results of the proposed algorithm will be presented and compared to the results of (Ge et al., 2019) in Section 3.4.3 to verify the validity of the *Model i*.

3.4.1 Datasets and Implementation Details

The proposed fashion landmark localization model can be extended to a variety of categories of clothing by changing the relative position of the clothing key points guided by human joints. In this study, one of the categories (e.g., T-shirts) was picked for experimentation. The dataset contains 1000 images from the DeepFashion2 dataset (Ge et al., 2019), which comprises 491K images of 13 clothing categories and 801K clothing items. Each item in an image is labeled with fashion landmarks annotations that can be the ground truth to examine the effectiveness of the proposed method. The fashion landmarks are predicted in DeepFashion2 with the KGDet model having 25 points in upper clothing. However, it is suspected that some of these 25 points are redundant and do not reflect clothing functionality. The *Model i* thus only used 16 points, as shown in Figure 3-1, to do qualitative and quantitative analysis.

3.4.2 Evaluation metrics

The normalized error (NE) metric (Liu et al., 2016b) is utilized as the criterion for evaluation of the proposed *Model i* and DeepFashion2 method. NE is the Euclidean distance between predicted and ground-truth landmarks, which formulated as:

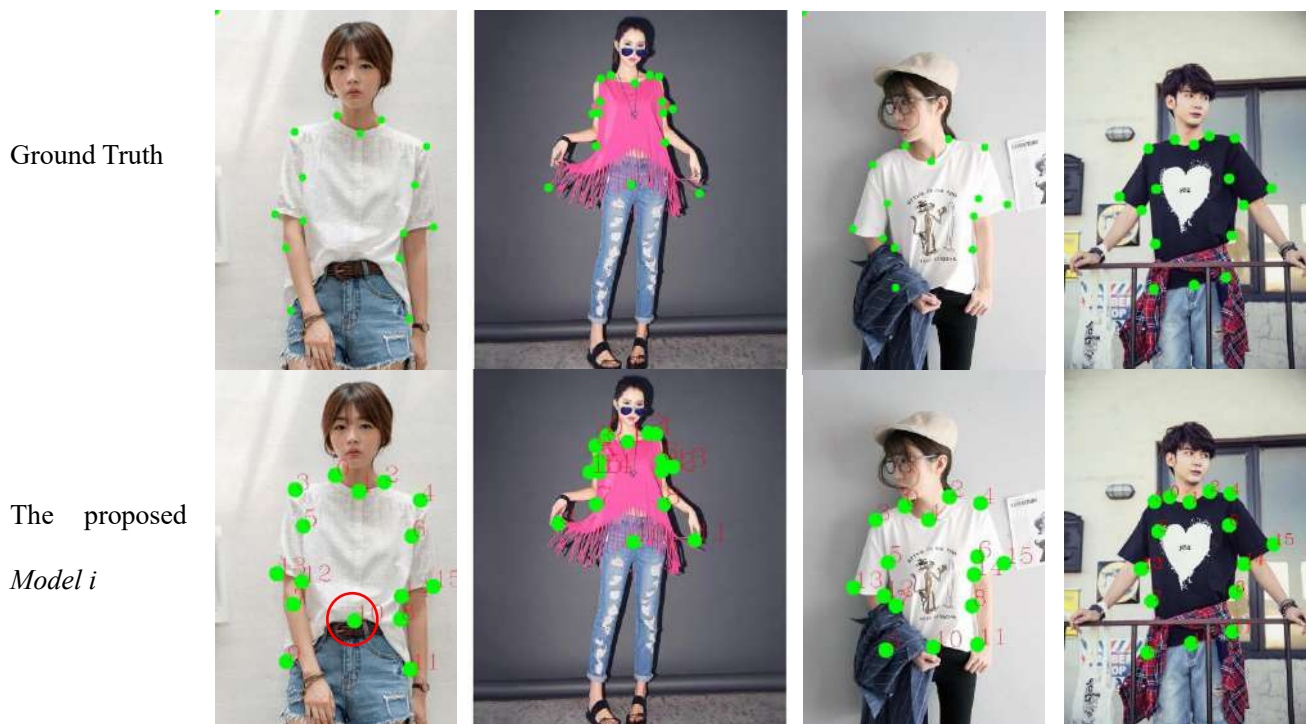
$$NE = \frac{1}{m} \sum_{i=1}^m \frac{|x_i - \hat{x}_i|_2}{h_i w_i} \quad (3-4)$$

where m denotes the number of points, x_i and $\hat{x}_i$ are the ground-truth and estimated landmarks respectively and h_i, w_i are the height and width of i -th point, $\frac{|x_i - \hat{x}_i|_2}{h_i w_i}$ is used to normalize the ground truth landmarks in coordinate space.

3.4.3 Results and Discussions

Experiments are conducted on the dataset in Section 3.4.1 to compare the fashion landmarks located by the proposed *Model i* with the landmarks predicted by the method of (Ge et al., 2019). The comparison results were shown in Table 3-3 (Hem Landmarks are not compared because of different definitions). From the table, the accuracy of the proposed method was not as good as deepfashion2, the key-points in the collar and shoulder area are located more accurately. The purpose of the experiment is not to beat DL-based method in performance, but to verify the rationality of landmarks in the follow-up task (automatic annotation tool). Instead, it was surprising to find that the NE values of shoulder landmarks were 0.0127 and 0.0195, which were less than 0.0231 and 0.0249 of the method by (Ge et al., 2019), as shown in Table 3-3. The landmark localization results are also visualized in Figure 3-11. The proposed *Model i* can effectively eliminate the interference of other types of clothing textures because of the geometric constraints on clothing and body proportions, as shown in the second row of

the fourth column of Figure 3-11, the prediction of point 10, Point 11, Point 12 are less accurate than the proposed method. In addition, since Ge et al. (2019)’s method has a different definition of the hem landmarks from ours, it is considered that the hem landmarks should be along the edge of the upper clothing. But they believe that the hem landmarks should be invisible in this case shown in the first and third row of Figure 3-11, so there are no annotations for invisible hem landmarks. From the visualized results, even so, the proposed method can still locate the reasonable hem landmarks. For example, at point 10 in the third row of Figure 3-11, Ge et al. (2019)’s method predicts on the trousers (unreasonable) and the proposed *Model i* predicts on the upper clothing edge (reasonable).



(Ge et al.,
2019)
DeepFashion2



Figure 3-11 Examples of Fashion Landmarks Localization Module qualitative results.

Table 3-3 Quantitative Results of Fashion Landmark Localization Model.

	(Ge et al., 2019) DeepFashion2	The proposed <i>Model i</i>
neckline_right	0.031	0.0261
neckline_center	0.0291	0.0253
neckline_left	0.027	0.0264
shoulder_right	0.0231	0.0127
shoulder_left	0.0249	0.0195
armpit_right	0.0364	0.0564
armpit_left	0.0347	0.0541
waistline_right	0.0227	0.0347
waistline_left	0.0203	0.0331
sleeve_right_out	0.0316	0.0584
sleeve_right_in	0.0271	0.0632
sleeve_left_out	0.0261	0.0568
sleeve_left_in	0.0297	0.0694
Avg.	0.028	0.0412

3.5 Limitation

3.5.1 Limitations of dataset selection

We carried out our experiments using subdatasets from DeepFashion2 (Ge et al., 2019), and this choice of dataset comes with both benefits and limitations. Unlike the supervised learning method, which requires learning the pixel positions of fashion landmarks from 491,000 labeled images, our method benefits from the relatively small size of the sub-dataset, consisting of only 1,000 images. This significantly enhances the efficiency of the experiments and accelerates the evaluation of our fashion landmark localization model (*Model i*). Moreover, sub-datasets allow us to concentrate on particular challenges and enable a detailed analysis of fashion landmarks by examining the modeling and positioning relationships between clothing and the human body.

However, there are also limitations to this way of selecting dataset. Firstly, the sub-dataset of DeepFashion2 (Ge et al., 2019) might not adequately capture the distribution of the original dataset, which can lead to reduced generalization capability of our model. Consequently, experimental results may not be applicable to a broader range of data or real-world scenarios. For instance, our algorithm struggles to accurately identify fashion landmarks of a garment from a side view. Secondly, choosing sub-dataset may introduce bias, potentially resulting in less stable experimental outcomes, particularly when dealing with self-occlusion issues or complex poses.

In the future, we intend to integrate our fashion landmark localization model (*Model i*),

which focuses on clothing and human body modeling, into a deep learning network (*Deep Model i*). We will employ transfer learning to fine-tune models that have been pre-trained on extensive datasets such as DeepFashion2 (Ge et al., 2019), aiming to enhance the performance of our *Deep Model i* on smaller dataset.

3.5.2 Failure Cases Discussions

The proposed fashion landmark localization model (*Model i*), as indicated in Section 3.3, extracts the initial masks, then completes their occluded regions, and then identifies the edge points of the completed clothing masks as coarse clothing key points. By grouping and refining the coarse clothing key points, the final fashion landmarks are located. Clothing mask completion in sleeve regions with slightly bad effects and weak robustness due to strong self-occlusion challenges induced by the large human stance. Since the *Model i* employs a less accurate clothing mask to locate the coarse sleeve key points, it is less accurate in locating the final fashion landmarks in the sleeve regions. Nonetheless, the sleeve fashion landmarks located by the proposed *Model i* are interpretable, which are near the clothing sleeve regions. Figure 3-12 (yellow circle) shows some failure sleeve fashion landmarks cases.



Figure 3-12 Examples of failure cases.

3.6 Chapter Summary

This chapter introduced a pose-aware segmentation-based method that can automatically locate key points even without Ground Truth. The proposed method can be applied to annotate fashion images and reannotate existing datasets for cross dataset learning. From the experimental results, the proposed method has certain advantages compared with the state-of-the art method in locating clothing shoulder and collar landmarks. The algorithm will be refined to improve the overall accuracy of fashion landmark prediction.

Chapter 4. Body-aware 2D Virtual Try-on

4.1 Introduction of 2D virtual try-on

Virtual Try-On systems (VTON) require creating a digital representation of a fashion item that can not only faithfully reflect the appearance of the real item but also fit to a given human. A practical VTON system is not only to present high-quality and realistic clothed images, but also to enable users to make more informed purchasing decisions in terms of the appropriate fit, shape, and size of the clothing (Bartle et al., 2016; Lin & Wang, 2016; Shin & Baytar, 2014). In recent years, numerous researchers (Ge et al., 2021a; He et al., 2022; Kim et al., 2024; Xie et al., 2023b) have dedicated significant efforts to develop image-based VTONs in an attempt to obtain more natural and accurate virtual try-on results in 2D form. The pioneering model of virtual try-on, VITON (Han et al., 2018), and its variants CP-VTON (Wang et al., 2018a), LA-VTON (Lee et al., 2019a), VTNFP (Yu et al., 2019b), SP-VTON (Song et al., 2020), VITON-HD (Choi et al., 2021) and GP-VTON (Xie et al., 2023b) use generative adversarial networks (GANs) (Goodfellow et al., 2014) to implicitly learn image generation. Recently, several works such as LaDI-VTON, OOTDiffusion, TryOnDiffusion (Gou et al., 2023; Kim et al., 2024; Zhu et al., 2023) consider implementing the virtual try-on through latent diffusion models (Rombach et al., 2022). The diffusion-based models significantly improve the image quality by iteratively denoising latent features (Ho et al., 2020; Rombach et al., 2022). Although the previous methods have made efforts in achieving superior image fidelity, they mainly focus on reconstructing the cloth appearance instead of faithfully reflecting the cloth fitting results. These methods

utilized the idea of image inpainting to reconstruct the clothed region. The clothed region is determined through human parsing techniques. However, this technique directly masks the upper body and does not provide hints on exact human joint ratios. Therefore, neither the GAN-based nor LDM-based VTON methods consider that a particular garment (will or can) fit differently on persons with different body types. For example, a T-shirt that fits the regular person might be looser in the chest and mid-section areas and have longer sleeves if worn by a skinny person. Furthermore, they ignore the position of a specific try-on garment. The body can vary depending on the person’s height and body proportions and shape. For example, a T-shirt that fits a teenager may be too short for an adult. Therefore, it is essential to account for the garments’ looseness and wearing position based on human body size and shape. The goal is to develop a practically useful image-based VTON method that closely resembles a real-life try-on experience.

To address the above-mentioned issues, the 3D VTON systems (Gao et al., 2024; Li et al., 2024b; Liu et al., 2023a) may fulfill such fitting requirements. However, the developed pipelines of these systems introduce additional setup/hardware constraints and extra computation, limiting applicability. As a step towards this goal, this chapter introduce a novel human **body size** dependent fashion **landmark transformation** VTON method, termed BSLT-VTON. The proposed BSLT-VTON is different from the previous generation approaches in that it synthesizes garment images on different-sized target persons, which is more true-to-life. To achieve this, it is considered that body-

size based garments deformation could be a valuable prior to a VTON technique. the relationship between human body and garment is explicitly modeled. It is achieved by employing a 2-step development strategy in BSLT-VTON to deform, scale, and adapt the garment to different people in real-time. Figure 4-1 shows a selection of try-on examples obtained by BSLT-VTON. The proposed method provides a more realistic and accurate representation of how the garment would actually look and fit on the target person.

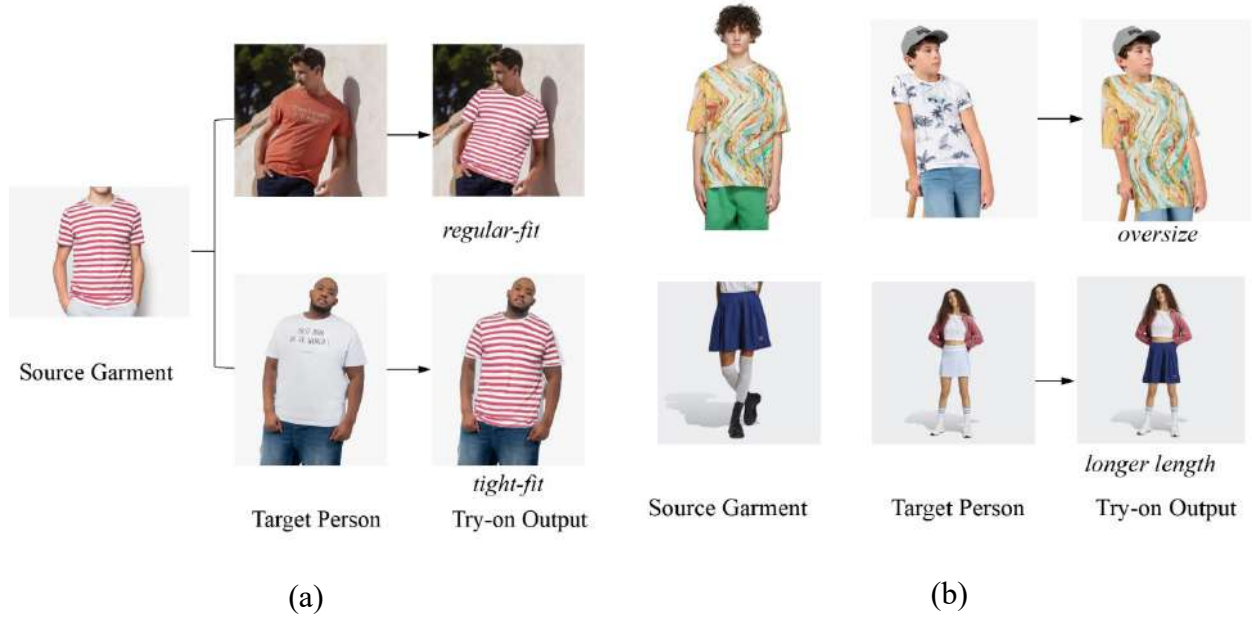


Figure 4-1 BSLT-VTON introduces a body-aware geometric deformation optimization algorithm that enables realistic virtual try-on. (a) An example of BSLT-VTON's garment looseness during wear. (a) Top: T-shirt try-on for a mesomorph shows regular fit. Bottom: Shirt try-on for an endomorph shows tight fit. (b) Illustration of BSLT-VTON's garment position during wear. Top: T-shirt try-on for a teenager shows oversize. Bottom: skirt try-on for a girl with a short and long skirt respectively.

The 2-step development strategy in BSLT-VTON covers (i) fashion landmark localization module, and (ii) a body-aware garment deformation model. In the first stage, a fashion landmark localization model (*Model i*) is developed to locate a set of key points (e.g., collar points, sleeve points, etc.) of the garment, aiming to provide meaningful control points for garment deformation.

In the first stage, a fashion landmark localization model (*Model i*) is developed to locate a set of key points (e.g. collar points, sleeve points, etc.) of garment, aiming to provide meaningful control points for garment deformation. In the second stage, a body-aware garment deformation model (*Model ii*) is designed. This aims to guide the source garment to deform onto a target human while being responsive to the target human's body shape, height/pose. In summary, the proposed BSLT-VTON method does not require training a deep learning model or the reconstruction of 3D garments and avatars to address the issue of simulating clothing fitting. It achieves realistic and functional try-on effects by modeling in a 2D visual form, transferring the wearable looseness and wearing position between the target person and the source garment. The main contributions are as follows:

- This chapter proposes a shape-faithful VTON method based on body size dependent fashion landmark transformation that can effectively transfer the wearable looseness and wearing position between the source garment and the target person, benefiting from a fashion landmark localization and human body-aware garment deformation model.

- The fashion landmarks localization model takes advantage of clear relationship between clothing and human body. This clearly defined modeling approach can locate fashion landmarks and flexibly extend the number and location of landmarks for controllable garment deformation.
- The human body-aware garment deformation model is designed to deform and scale the garment appropriately according to different body shapes. Its effectiveness has been demonstrated in terms of garment height and length (is overall fit) effects for people of different body shapes and heights/ proportions.
- Extensive experiments demonstrate that the proposed BSLT-VTON is satisfactory in terms of garment deformation and outperforms some state-of-the-art VTON methods in preserving garment details.

4.2 Related Work

As a comprehensive topic of image processing, VTON has been studied for many years. To establish a new method, this chapter mainly recaps the related VTON work and the research fields involved in the proposed method, including fashion landmark detection and fashion image deformation.

4.2.1 Virtual Try-on

Virtual Try-On related technologies have improved considerably in recent years. For example, VTON function can be performed in either 2D or 3D way. 3D VTON models or technologies, (Bartle et al., 2016; Pumarola et al., 2019), rely on 3D human body

models and physically-based cloth simulation to realistically simulate garment drape when worn. Although some VTON-related software products (e.g., Optitex3D, CLO, Browzwear and *etc.*) have been commercialized to as a complete system of clothing style design, patternmaking, grading, and virtual fitting, they still face certain difficulties of insufficient workforce of the 3D Art-Tech, expensive hardware installation (e.g., infrared scanning, point cloud equipment), computation capability *etc.* More recent, deep learning-based 3D VTON attempt to efficiently reconstruct the 3D human (Guler & Kokkinos, 2019; Kanazawa et al., 2018; Zheng et al., 2019) and 3D garments (Bhatnagar et al., 2019; Zhu et al., 2020) as a core technical support for their application. However, the lack of easily accessible annotated 3D data and the inferior quality of reconstructed 3D models and texture details still hamper the widespread deployment of 3D VTON models. As a result of the 3D VTON’s limitations, more efficient 2D VTON models have been developed, which reformulate VTON as a conditional image generation problem without 3D data being required. Such an image-based VTON solution entails comprehensive image processing that combines image segmentation (Chen et al., 2017; Gong et al., 2017; Liang et al., 2015a), deformation (Chen et al., 2019b; Schaefer et al., 2006), and generation (Han et al., 2019b; Zhu et al., 2019) to enable learning for cross-domain transfer between the clothing from in-shop fashion images and the target/user human body. The common 2D VTON architecture is mostly based on the GAN model (Goodfellow et al., 2014) to generate virtual fitting images that are conditioned on the input signal to minimize the adversarial loss, making the image generated by the generator indistinguishable from the real image determined

by the discriminator to train the network. Different VTON frameworks have been proposed by various researchers in succession based on task dependencies. For example, VITON (Han et al., 2018), a pioneering virtual try-on model adopts the coarse-to-fine strategy to transfer and deform the desired clothing to the corresponding areas of the person. Variants of VITON, such as CP-VTON (Wang et al., 2018a), SP-VTON (Song et al., 2020), LA-VTON (Lee et al., 2019a), VTNFP (Yu et al., 2019b), have also been continuously optimized for improving the virtual visual appearance. Furthermore, GarmentGAN (Raffee & Sollami, 2021) retains the clothing details by separating the clothing shape and appearance during in the network training, using segmentation maps and human pose information. In the other hand, ACGPN (Yang et al., 2020) maintain the features of the target person and optimize the details of the source garments with due consideration of the semantic layout. Nevertheless, since these VTONs are generally given a pair of images (source garment and target person) as input, that of the garment only or alone is heavily influenced by the target posture of the person, resulting in poor try-on results for complex poses. To overcome their problems, SwapNet (Raj et al., 2018) operates in image space and enables the transfer of clothing images between images of people with arbitrary body poses, shapes and clothes. MG-VTON (Dong et al., 2019) and Fit-Me (Hsieh et al., 2019) lead to multiple pose dependent virtual try-on, which provide a realistic solution for complex pose fitting by parsing the target clothing and estimating human pose. M2E-TON (Wu et al., 2019) produces try-on outcomes featuring coordinated poses and accurately retained textures through the integration of a dense pose alignment network, a texture refinement network, and a

fitting network to merge the transferred clothes into the target person image. PF-AFN (Ge et al., 2021b) realizes the dependence of the VTON model on human parsing and employs knowledge distillation to find dense correspondences between person images and clothing images, thereby producing high-quality results. The above methods all (or at least) utilize human parsing, pose estimation and texture refinement for virtual try-on operations, ignoring the impact of different body shape differences in clothing transfer. Contrary to the above, the specific modeling method effectively considers the close relationship between the body and clothing, producing a more realistic and reasonable fitting effect.

4.2.2 Fashion landmark detection

Fashion landmark detection can be used for various applications, such as establishing the relationship between clothing and human body for virtual try-on purposes, embracing the accuracy of sizing and fit recommendations, and optimizing the entire design and manufacturing processes. It employs algorithms to detect or locate specific landmark points, such as necklines, shoulders, and other key design related elements that are important for creating a detailed and accurate digital representation of the garment from a human body respective.

In recent years, thanks to the public release of a large dataset (Ge et al., 2019; Guo et al., 2016) accelerated the research on landmark detection. For example, FashionNet (Guo et al., 2016) was the first to use clothing attributes and fashion landmarks to

understand fashion images. They proposed a three-level cascade framework to more accurately estimate fashion landmarks, using VGG-16 (Simonyan & Zisserman, 2014) at each stage to predict landmark location and visibility, estimate local features on fashion landmarks, and capture global features of the entire garment item, respectively. Besides regression-based models, attention related mechanisms are widely used for landmark detection. ‘landmark heatmaps’ have been proposed (Liu & Lu, 2018) for capturing attention to and localizing landmarks. Moreover, Li et al. (2019c) proposed a spatial-aware non-local landmark detection model that accepts high-level semantic knowledge in encoding the prior knowledge by means of a spatial attention map. Lee et al. (2019a) detected landmarks by utilizing a global-local embedding model which contains a global procedure and a convolutional procedure to obtain long-range relationships and localized information in fashion images respectively.

A key requirement on the above landmark detection methods is the availability of large, annotated training data, which provide a true or more realistic base and a simpler design of loss terms. A significant drawback that impedes its broader applications lies in the variance of annotation strategies across datasets, leading to diverse landmarks for varying clothing styles and limited dataset compatibility. While pose estimation shares similarities with fashion landmark localization, the distinctions lie in the specifics of landmark location and quantity. Unlike the above methods, a pose-aware segmentation-based method is proposed to locate fashion landmarks by explicitly modelling the relationship between clothing and the human body. The input image is segmented to

extract the source garment, and which the human pose joints are used as prior knowledge to locate the different functional areas of the garment. The landmarks located or identified by the proposed method can be extended and provide the flexibility to re-annotate existing datasets.

4.2.3 Clothing deformation

As a field-specific deformation algorithm, the clothing deformation algorithm is widely used in VTON because it can transfer texture well during the process of deforming clothing. There are two main types: TPS (Choi et al., 2021; Fincato et al., 2022; Fincato et al., 2021; Minar et al., 2020) and Appearance Flow (Ge et al., 2021b; Han et al., 2019a). For example, VITON (Han et al., 2018) first tries to deform clothing masks extracted from both the clothing foreground and the person by calculating the shape context TPS warping, achieving improved results for virtual try-on. CP-VTON (Wang et al., 2018a) employs to automatically learn the TPS deformation parameters in their dual-input network to generate deformation results. ACGPN (Yang et al., 2020) incorporates second-order constraints that enables the network to achieve improved fidelity in maintaining the visual appearance of the clothing throughout the deformation process. However, the methods mentioned above mainly concentrate on directly mapping pixels at the spatial level of the image, deforming the input clothing to fit the ground truth (GT) clothing area without considering variations in clothing size and human body shape. The proposed BSLT-VTON incorporates fashion landmarks and estimated human joints to effectively deform clothing based on various body shapes

and heights/proportions.

4.3 Method

The aim of the proposed method is the virtual try-on of source garments on target persons who have different body shapes and sizes, which requires body shape information and content-aware image deformation. Different from other VTON methods which utilize a network to directly generate a try-on image, the proposed human body size dependent fashion landmark transformation VTON method (BSLT-VTON) decomposes VTON pipeline into (i) a fashion landmark localization model (as described in Chapter 3) and (ii) a body-aware garment deformation model (as described in Section 4.3.2), as shown in Figure 4-2 (Symbols refer to Table 4-1). In order to guide the body-aware deformation to synthesize a shape-faithful fitting result, the fashion landmarks are first located (Figure 4-2 (a) (b) and (c)) by utilizing the close relationship between the human body and clothing, as it could reflect the new wearer's shape. Then, the relevant proportion coefficients of the human body and garment are calculated to relocate the fashion landmarks on the target garment so as to distinguish between them and those on the source garment (Figure 4-2 (e)). In addition, fashion landmarks, human parsing and pose estimation are introduced as body prior information in the deformation process (Figure 4-2 (d) and (f)) so as to effectively transfer the garment looseness and position on the source person to the target person.

4.3.1 Mathematical notation

Table 4-1 lists contain the mathematical notation used for garment, mask, self-perceived size, fashion landmark and human person symbols used. Some notations that will be used throughout this chapter are introduced.

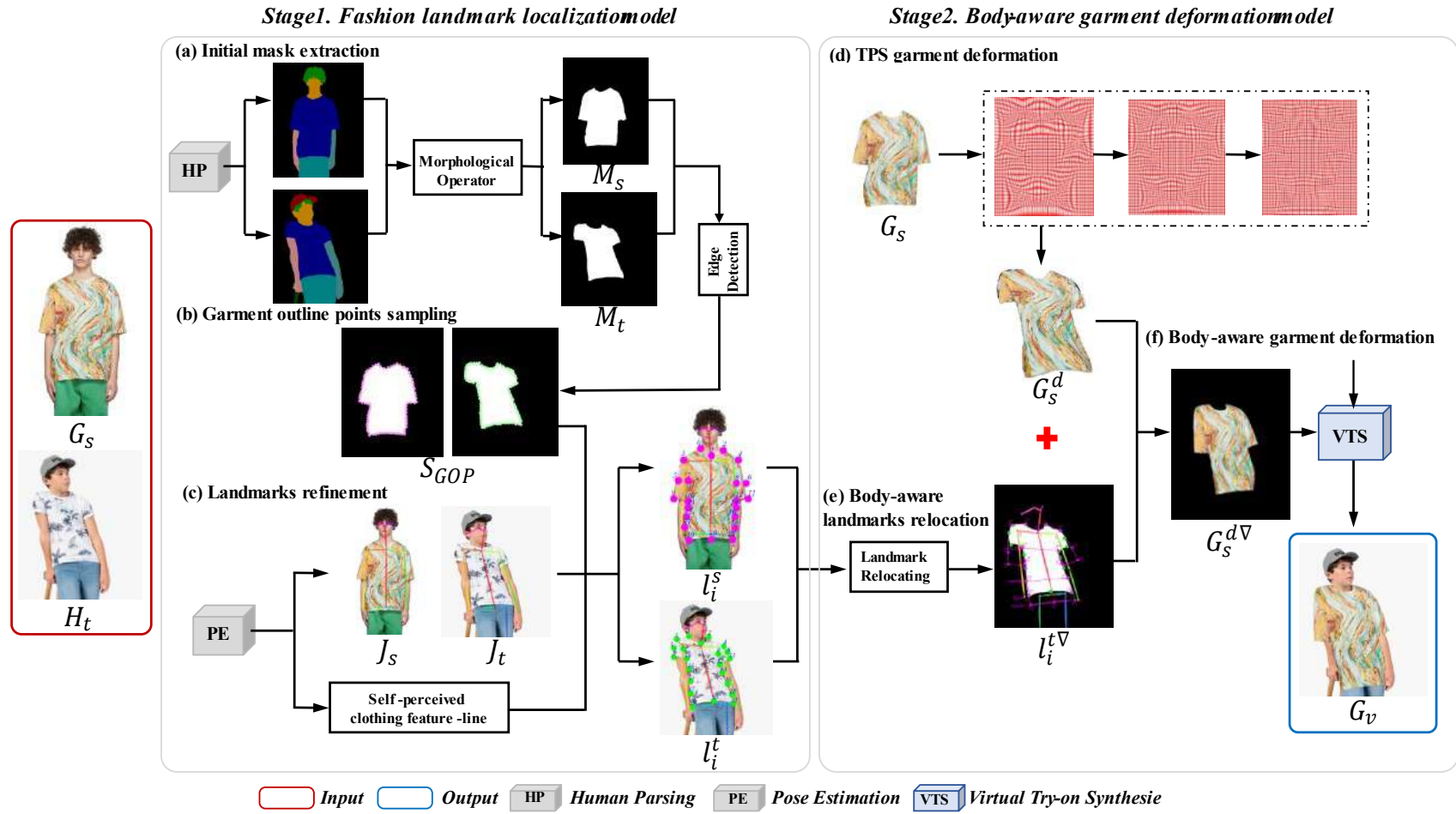


Figure 4-2 Illustration of the human body size dependent fashion landmark transformation VTON method.

Table 4-1 The mathematical notation used for different variables.

Garment	Mask	Self-perceived size	Fashion landmarks	Human person
Source garment: G_s	Source garment mask: M_s	Source garment size: S_s	Source garment landmark: l_i^s	Source human: H_s
Target garment: G_t	Target garment mask: M_t	Source body size: J_s	Target garment landmark: l_i^t	Target human: H_t
TPS-deformed source garment: G_s^d	TPS-deformed garment mask: M_s^d	Target garment size: S_t	Body-aware target garment landmark: $l_i^{t\forall}$	
Body-aware deformed garment: $G_s^{d\forall}$	Body-aware target garment mask: $M_t^\forall$	Target body size: J_t		
Body-aware target garment: $G_t^\forall$		Body-aware target garment size: $S_t^\forall$		

4.3.2 Human body-aware garment deformation model

Whereas the most common VTON method utilizes the target garment mask M_t to guide the source garment mask M_s to deform, generating the source garment G_s texture in the appropriate clothing area of the target human H_t . Such a VTON will squeeze G_s of different styles (tight-fit, regular-fit, over-fit) into the corresponding target garment area G_t , which cannot reflect the fitting effect of the H_t trying on different style of G_s . The proposed BSLT-VTON can solve this problem through body-aware landmark relocation (Section 4.3.2.2) and landmark-based garment deformation (Section 4.3.2.3).

4.3.2.1 Definition the size of garment and human body

Garment size is very important for creating a properly fitting garment. To ensure that the range of the garment is sized appropriately for the intended wearer, the overall size of G_s and G_t is first defined. The correspondence between the garment and the various regions of the wearer then is modeled explicitly. The overall size of the garment is expressed in terms of width W and length L .

Top garment. The width measurements include collar width, shoulder width, bust width, lower bust width, upper waist width, waist width, hem width and sleeve width. In addition, collar-to-hemline length, hipline-to-hemline length and sleeve length are also described as length measurements. *Bottom garment.* The width measurements include waistband width, hip width and leg width. The length measurements include waist-to-

crotch length and waist-to-hem length. Human body size is calculated based on human joints J , shown in Figure 3-6 with the j_i as defined the j -th joint of the human body. The human body J_s is the Euclidean distance between two defined joints. l_i is defined as the i -th fashion landmark of clothing located in the *Model i*, as shown in Figure 4-3. The clothing width W and length L are the Euclidean distances between correspond two landmarks respectively. The detailed descriptions and relationships are summarized in Table 4-2.

Table 4-2 The size (length & width) of different parts of the garment and the corresponding human body size

Garment category	Name of garment length & width of garment parts	Clothing overall size	Human body size
Top (a,b)	Collar width: W_{collar}	$\ l_0l_2\ $	$\ j_2j_5\ $
	Shoulder width: $W_{shoulder}$	$\ l_3l_4\ $	
	Bust width: W_{bust}	$\ l_7l_8\ $	
	Lower bust width: W_{lbust}	$\ l_{13}l_{14}\ $	
	Higher (Upper) waist width: W_{hwaist}	$\ l_{15}l_{16}\ $	$\ j_9j_{12}\ $
	Waist width: W_{waist}	$\ l_{17}l_{18}\ $	
	Hem width: W_{hem}	$\ l_{19}l_{21}\ $	
	Sleeve width: W_{sleeve}	$\ l_9l_{10}\ , \ l_{11}l_{12}\ $	$\ j_2j_1\ , \ j_1j_5\ $
	Collar-to-hemline length: $L_{collar \rightarrow hem}$	$\ l_0l_{19}\ , \ l_2l_{21}\ $	$\ j_1j_8\ $
	Waistline-to-hemline length: $L_{waist \rightarrow hem}$	$\ l_{17}l_{19}\ , \ l_{18}l_{21}\ $	
	Sleeve length: L_{sleeve}	$\ l_3l_5\ + \ l_5l_9\ ,$	$\ j_2j_3\ , \ j_2j_3\ + \ j_3j_4\ $

		$\ l_4l_6\ + \ l_6l_{11}\ $	$\ j_5j_6\ , \ j_5j_6\ + \ j_6j_7\ $
Bottom (c,d)	Waistband width: $W_{waistband}$	$\ l_0l_2\ $	$\ j_9j_{12}\ $
	Hip width: $Whip$	$\ l_3l_4\ $	
	Leg width: W_{leg}	$\ l_{10}l_{11}\ , \ l_{12}l_{13}\ $	$\ j_9j_8\ , \ j_8j_{12}\ $
	Waist-to-crotch length: $L_{waist \rightarrow crotch}$	$\ l_1l_5\ $	$(\ j_9j_{10}\ + \ j_{12}j_{13}\)/2$
	Waist-to-hem length: $L_{waist \rightarrow hem}$	$\ l_0l_6\ + \ l_6l_{10}\ ,$ $\ l_2l_9\ + \ l_9l_{13}\ $	$\ j_9j_{10}\ + \ j_{10}j_{11}\ ,$ $\ j_{12}j_{13}\ + \ j_{13}j_{14}\ $

4.3.2.2 *Body-aware try-on garment landmark relocation*

People have different body shapes and heights, even the same clothes will correspondingly look different when tried on them. It is valid and reasonable to employ the body frame as a basis for estimating variable the correspondingly different clothing sizes and establishing the corresponding ratio relationships between them, since the basic structure of the human body is similar for all individuals, regardless of their body shape. Hence, the looseness and position of the clothing fitted on the H_t are adjusted according to the ratio relationships between the garment size (S_s and S_t) and human frame size (J_t).

Firstly, the corresponding S_s and J_s are calculated for the i^{th} fashion landmark ($l_{i,x}^s, l_{i,y}^s$) in the source garment. Secondly, the position information of the target human joints ($j_{i,x}^t, j_{i,y}^t$) and J_t to are used to calculate $l_i^{t\forall}$, shown as the rosy dot in Figure 4-3 (b). Based on the assumption that the basic structure of the human frame is similar for individuals of different sizes, the relative width J_s is approximately equal to the J_t . Three criteria for landmark relocation, orientation, length, and width, are employed. The orientation criterion is to compare what direction the newly positioned $l_i^{t\forall}$ is on the l_i^t of target garment. The width orientation criterion estimates the range of the variation between the $G_t^\forall$ (shown as the rosy polygon area in Figure 4-3 (b)) and G_t in the direction of the width of the different parts displayed in Table 4-2. By estimating the range of movement in the width direction, it is possible to ensure that the source garment try-on fits and looks natural on the target person. The length orientation

criterion is calculated in a similar way to the width orientation criterion, which is to ensure the proper try-on position of the source garment on the target person. The horizontal movement of $l_i^{t\nabla}$ is calculated as in Eq. (4-1) as follows:

$$l_i^{t\nabla} = (j_{i,x}^t + \frac{S_s - J_s}{2}, j_{i,y}^t) \text{ or } (j_{i,x}^t - \frac{S_s - J_s}{2}, j_{i,y}^t) \quad (4-1)$$

The vertical movement is treated in a similar way. The only difference is that the landmarks in the target image are calculated as in Eq. (4-2) as follows:

$$l_i^{t\nabla} = (j_{i,x}^t + \frac{S_s - J_s}{\lambda}, j_{i,y}^t) \text{ or } (j_{i,x}^t - \frac{S_s - J_s}{\lambda}, j_{i,y}^t) \quad (4-2)$$

where λ is the scaler defined as $\frac{\|j_2^s j_5^s\|}{\|j_2^t j_5^t\|}$.

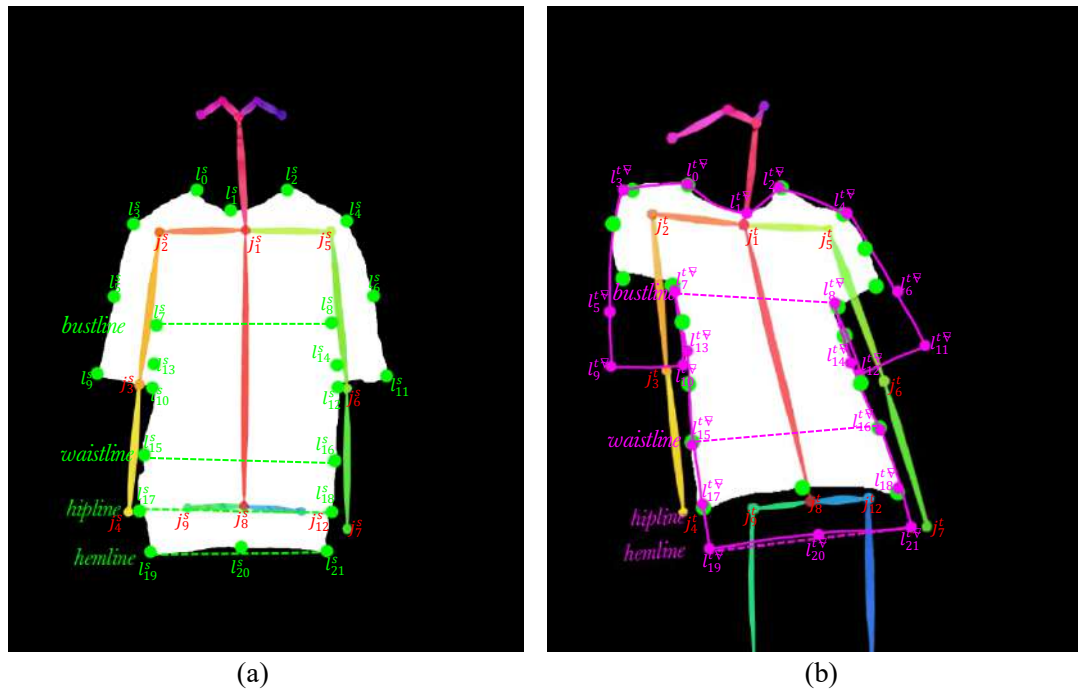


Figure 4-3 Body-aware target garment fashion landmark relocation, where illustrates source garment landmark l_i^s (green dots in (a)), target garment landmark l_i^t (green dots in (b)), and body-aware target garment landmark $l_i^{t\nabla}$ (rosy dots in (b)).

4.3.2.3 Garment deformation via fashion landmark

The proposed *Model ii* describes a 2-step garment deformation strategy to accommodate different body shapes in virtual try-on applications. In the first step of the deformation strategy, TPS (Bookstein, 1989) is used to deform the $G_s \rightarrow G_t$ by aligning $l_i^s \rightarrow l_i^t$. In the second step, Moving Regularized Least Squares (MRLS) deformation algorithm (Ma et al., 2013) is used to deform $G_s^d \rightarrow G_s^{d\forall}$ by aligning $l_i^t \rightarrow l_i^{t\forall}$. To prevent excessive deformation of the source garment, a small incremental deformation steps approach is adopted in *Model ii*. In each iterative deformation approach, TPS parameters are used to describe the smooth translation of the fashion landmarks. First, the source garment mask M_s and target garment mask M_t are downsampled with the same size patch of 3 x 3 pixels to obtain the related control matrices C_s and $C_t \in \mathbb{R}^{h \times w}$, with the difference field ΔC used to compare C_s and C_t . The goal is to make M_s and M_t gradually tend to be similar during the deforming process, that is $\Delta C \rightarrow 0$. The pixel displacement from C_s to C_t is recorded, and each pixel $G \in \mathbb{R}^{(h-2) \times (w-2) \times 2}$ in the difference between C_s and C_t is calculated. Then the G can be defined as follows:

$$G(p_g^i) = \dot{S} \circ \Gamma \circ \left(\sum_{k=-1}^1 (p_{\Delta C}^{k \times h+i-1} - p_{\Delta C}^{k \times h+i+1}), \sum_{j=-1}^1 (p_{\Delta C}^{j+i-1} - p_{\Delta C}^{j+i+1}) \right) \quad (4-3)$$

$$\dot{S} = (\dot{s}_x, \dot{s}_y) \quad (4-4)$$

$$\Gamma = (\gamma_k, \gamma_j) \quad (4-5)$$

where p_g^i is the i^{th} pixel of G , $\dot{S}$ is the incremental difference, Γ is the difference delta weight, which describes the influence of each adjacent pixel on the deformation with $\circ$

is being the Hadamard product.

The TPS parameters are represented as:

$$\delta = \begin{bmatrix} K & X \\ X^T & 0 \end{bmatrix}^{-1} \cdot Y \quad (4-6)$$

We let the initial position of C_s be the source control matrix $X \in R^{h \times w \times 3}$, and K is the correlation matrix between the control points of C_s . The expression of target control matrix Y is expressed as follows:

$$Y = C_s + G(p_g^i), Y \in R^{(h \times w + 3) \times 2} \quad (4-7)$$

Substituting Eq. (4-7) into Eq.(4-6), gives the following:

$$\delta = \begin{bmatrix} K & C_s \\ C_s^T & 0 \end{bmatrix}^{-1} \cdot (C_s + G(p_g^i)) \quad (4-8)$$

According to Eq. (4-8) that the sample size of the control matrix and the incremental difference between the two masks determine the deformation parameters of TPS. We set a defined of difference scale coefficients $N = \{1 \times, 2 \times, \dots, 2^n \times\}$ to iterate ξ times to reduce the difference between M_s and M_t . In each iteration, the TPS-deformed source garment masks $\{M_s^d(0), M_s^d(1), \dots, M_s^d(n)\}$ are generated by the TPS parameters $\{\delta(0), \delta(1), \dots, \delta(n)\}$. After several iterations, the shapes of the two garment masks are changed step by step until they are the same (i.e. $M_s \equiv M_t$). The iteration number ξ , corresponding to $\delta(n)$ is adopted to deform the G_s^d in order to retain the topological consistency between G_s^d and G_t .

Nevertheless, the G_s^d obtained by the above TPS deformation is similar in shape to the G_t . The looseness and position of the clothing fitted on the H_t failed to reflect.

Therefore, in the second step of the deformation strategy, the $l_i^{t^v}$ is used as control points to guide the G_s^d further deform action. After reviewing several deformation systems, the MRLS algorithm is finally adopted. For the traditional moving least squares (MLS) image deformation method (Schaefer et al., 2006), the deformation function $l_v(p)$ is defined as a rigid transformation the purpose of which is to obtain the least squares error as follow:

$$\arg \min_{l_v} \sum_i w_i \|l_v(p_i) - q_i\|^2 \quad (4-9)$$

Nevertheless, this rigid deformation involves an affine transformation that includes inhomogeneous mismatch and scaling, leading to unsatisfactory results when applied to varying clothing images. Hence, the MRLS algorithm is adopted to solve the non-rigid object deformation problem. We use $g_v(p)$ to replace $l_v(p)$, that is, to minimize the least squares error by means of the following regularization:

$$\arg \min_{g_v} \sum_{i=1}^n w_i \|p_i + g_v(p_i) - q_i\|^2 + \lambda \Phi(g_v) \quad (4-10)$$

It can be seen that the algorithm is similar in form to MLS except for the regularization term $\lambda \Phi(g_v)$, where $\lambda > 0$, w_i is the weight coefficients. For each coordinate point v of the image, the displacement function $g_v(p)$ obeys the Hilbert space $\mathcal{H}$ (Reproducing Kernel Hilbert Space, RKHS) (Micchelli & Pontil, 2005), KHS can be obtained by means of $\mathbf{F}$ (Zhao et al., 2011), hence $g_v(p)$ can be expressed as in Eq. (4-11) as follow:

$$g_v(p) = \sum_{i=1}^n \mathbf{F}(P, P_i) c_i \quad (4-11)$$

where the coefficient c_i is 2×1 -dimensional vector.

The mapping function f is expressed as Eq.(4-11), with $w_i(p)$ tending to infinity when point v approaches P_i , at which point the function f satisfies interpolation, $f(p_i) = q_i$.

Also, for any i , $g_v(p) = 0$ and $f(v) = v$ when $p_i = q_i$.

$$f(v) = v + g_v(v) \quad (4-12)$$

The regulating term $\lambda\Phi(g_v) = \|g_v\|_F^2 = \langle g_v, g_v \rangle_F$, using $\Gamma(P_i, P_j) = e^{-\|p_i - p_j\|^2 / \beta^2} I$ (β determines the width of the interpolation) to simplify Eq. (4-10) as follow:

$$\left\| W^{\frac{1}{2}}(P + \Gamma C - Q) \right\|_F^2 + \lambda \text{tr}(C^T \Gamma C) \quad (4-13)$$

where Γ is the Gamma matrix $\Gamma_{ij} = e^{-\|p_i - p_j\|^2 / \beta^2}$. W is a diagonal matrix with $w_i(p)$ diagonal elements. P and Q are the original and target set of control points respectively, and $C = (C_1, \dots, C_n)^T$ is a coefficient matrix $n \times 2$ in size. By taking the partial derivative of Eq. (4-13) and making it equal to 0, C can be calculated as follows:

$$C = (\Gamma + \lambda W^{-1})^{-1}(Q - P) \quad (4-14)$$

Substituting Eq. (4-14) into Eq.(4-12), leads to the mapping function f as follow:

$$f(v) = v + [\Gamma_v(\Gamma + \lambda W^{-1})^{-1}(Q - P)]^T \quad (4-15)$$

In fact, the transformed l_i^s have the same coordinates as the target garment landmarks l_i^t . Hence, l_i^t is used as the original set of control points and body-aware garment landmarks $l_i^{t\vee}$ as the target set of control points. The size of the initial grid is the same as that of the TPS-deformed source garment image G_s^d . Algorithm 4-1 details the body-aware garment deformation process.

Algorithm 4-1 Body-aware garment MRLS deformation

Input: the set of target garment landmarks: $\{l_i^t\}_{i=1}^n$

the set of body-aware target garment landmarks: $\{l_i^{t\mathcal{V}}\}_{i=1}^n$

the TPS-deformed source garment: G_s^d

the initial grid: G_{grid}

Output: the body-aware deformed garment (BLST): $G_s^{d\mathcal{V}}$

1. **Calculate** the Gamma Matrix Γ , $\Gamma_{ij} = e^{-\|p_i - p_j\|^2 / \beta^2}$
2. **For** $v_{i,j} \in G_{grid}$ **do**
3. **Calculate the** weight function W and vector Γ_v ,
4. **Calculate** the mapping function $f(v)$ by mean of Eq.(4-15) .
5. Pixel resampling from the deformed grid,
6. **Return** $G_s^{d\mathcal{V}}$.

4.3.3 The virtual try-on synthesis

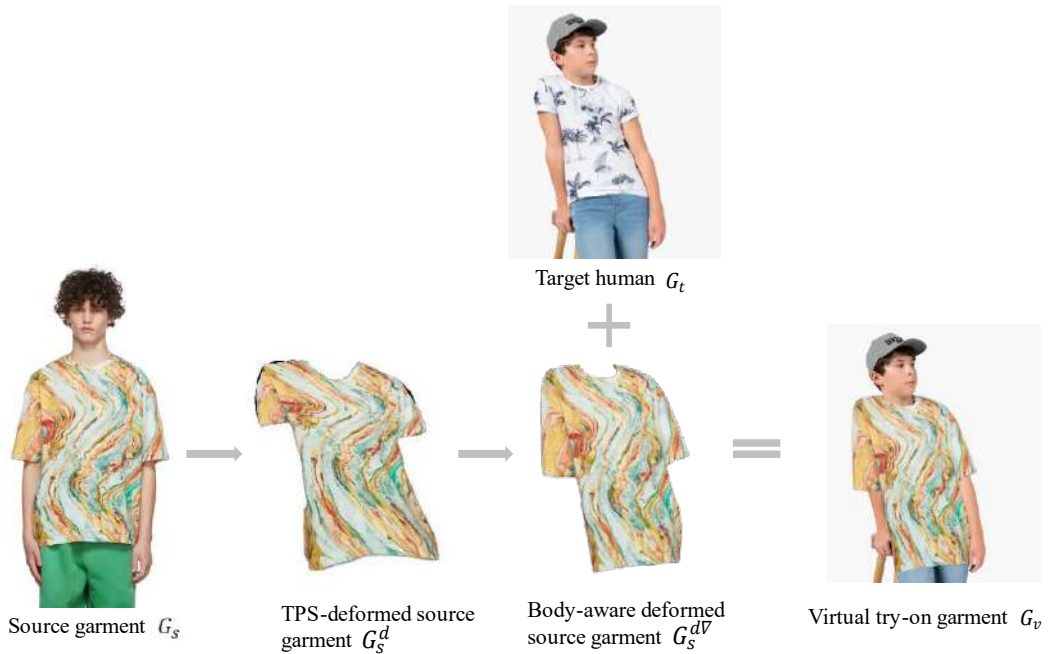


Figure 4-4 The illustration of the virtual try-on image synthesis.

Blending images to largely reduce and eliminate the influence of artifacts, a variant of the Poisson blending (Pérez et al., 2003) is utilized to synthesize the G_s^{dV} to the target human H_t . This is achieved by using built-in function of Opencv cv2.seamlessClone executed. The virtual try-on image synthesis is illustrated in Figure 4-4.

4.4 Experiments

4.4.1 Experimental Setup

Dataset. Since the Zalando Dataset (Han et al., 2018) used for the virtual try-on experiments is no longer publicly available, a human body size dependent fashion landmark transformation (BLST) dataset was composed of a sub-dataset of Deepfashion2 (Ge et al., 2019) and garment images from several e-commerce websites (e.g. Adidas, SSENSE). The dataset has a total of 2000 garment images. Two images are randomly sampled from BLST dataset, one is the source garment, and the other is the target garment image and are available for the input of BSLT-VTON. The BLST dataset is not large compared to other datasets used for VTON but is adequate to serve as a test set to demonstrate faithfulness and the superior performance in the visual quality of the results generated by the proposed BSLT-VTON.

Compared Methods. On the BLST dataset, the proposed BSLT-VTON is compared with multiple state-of-the-art VTON competitors, including PFAFN (Ge et al., 2021b), DiOr (Dubey, 2021), Flow-style VTON (He et al., 2022), MCG VITON (Fang et al., 2022), LaDI-VTON (Morelli et al., 2023) and OOTDiffusion (Xu et al., 2024). These

methods, aside from MCG VITON, utilize generative models (Goodfellow et al., 2014; Ho et al., 2020) to facilitate the try-on process. Their goal is to maintain the texture and details of input source garments while also preserving the physical attributes of the target human as much as possible. MCG VITON is to reconstruct a 3D Avatar based on the target person image and render it into a virtual human body image. Then the input garment is deformed via TPS algorithm and worn on the virtual human body to preview the try-on effect.

Evaluation Metrics. As per VTON standard, all evaluation metrics are implemented following the previous work (Detlefsen et al., 2022; Parmar et al., 2022). Although the proposed BSLT-VTON supports higher-resolution virtual try-on, all the experiments are conducted at the resolution of 512×512 for fair comparison with previous VTON methods. We use structural similarity (SSIM) (Wang et al., 2004) and LPIPS (Zhang et al., 2018) to measure the quality of the generated image in terms of restoring the input images. And we employ KID (Bińkowski et al., 2018) and FID (Heusel et al., 2017) for realism and fidelity assessment.

4.4.2 Universality

4.4.2.1 The Try-on results for different body shapes

The proposed BSLT-VTON is harmonize to the human body shape and aims to accurately represents the user’s body shape and display how a particular garment would

appear and fit on. As a preliminary trial, we randomly selected a loose T-shirt as a test garment to the try-on effect on four different body types: thin, fat, tall and short. As illustrated in Figure 4-5, the T-shirt originally worn on the source person has a different fit (looseness) on the persons with different body types (Figure 4-5 (a) and (b)). The T-shirt was fitting loosely on the thinner person, showing loose-fit effect (Figure 4-5 (e)), mainly because the G_s is larger than the T-shirt's in the width direction (Figure 4-5 (a)). Figure 4-5 (f) reflects the opposite effect, where the T-shirt appears too tightly on the target person, even to the point where the target person's own white T-shirt is visible beneath the black T-shirt. The differences in the fit of the T-shirt in Figure 4-5 (c) and (d) are eventually due to differences in their heights and body proportions, one target person (Figure 4-5 (c)) is being taller than the other one (Figure 4-5 (d)). Hence, even though the garment is the same, the length fit is different for them. For the person in Figure 4-5 (c), this is an oversized (i.e. too long) T-shirt, while for the human in Figure 4-5 (d), it is just right. In addition, it was found from the experimental results that the deformation ratio of the source garment image differed in the sleeve, bust, waist, and hip areas, but remained almost the same in the collar width and shoulder area.

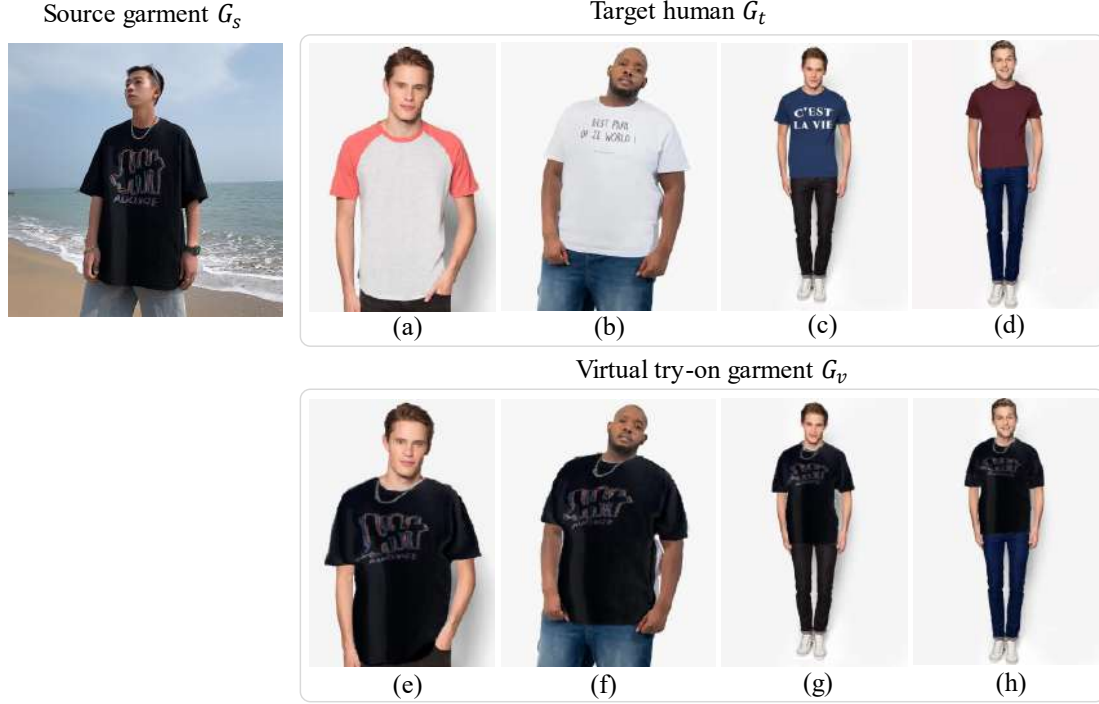


Figure 4-5 The virtual try-on results for different body shapes.

4.4.2.2 Try-on results for different garment types

To verify the correct and effectiveness of BSLT-VTON on different types of garment, T-shirts, long-sleeved shirts, shorts and long pants were investigated. The upper garments are characterized by 22 fashion landmarks. These landmarks serve as control points to guide the deformation of the garments in the collar, shoulder, sleeve, waist, and hem area, which remain the texture and style unchanged. For example, in Figure 4-6 (a), the T-shirt G_v shows that the source garment is a t-shirt with shorter sleeves. In Figure 4-6 (b), the striped T-shirt is shown as worn on a target person of similar body shape, and the deformation degree of the G_v striped texture is kept as similar as possible to that of the source garment. The silhouette of the long-sleeve top G_v is having reasonable variable, such as a raised arm curve due to large postural changes. The right

sleeve is kept a garment to the body, which cannot be dealt with in terms of deformation (shown in Figure 4-6 (c)). The lower garments are defined by means of 14 fashion landmarks, which control the garment deformation at the waist, hip, crotch, knee and ankle area respectively. The form fit looseness/tightness of shorts or long pants is mainly determined by the waist-to-hip ratio of the garment and waist-to-hip ratio of body frame. For example, in Figure 4-6 (g), the trouser legs of the baggy trousers are wider than those of the target garment, and the crotch parts are deformed downward.



Figure 4-6 Application of BSLT-VTON capability to try-on garment for various categories using fashion landmarks to guide the garment deformation.

4.4.3 Ablation study

4.4.3.1 Overview of try-on influences

As described in Section 4.2.1, the common VTON methods cannot reflect the actual

and faithful fitting effect (wearable looseness and wearing position) when H_t tries different styles of G_s . The BSLT-VTON solves this problem mainly through fashion landmarks localization model and human body-aware garment deformation model. Fashion landmarks (control points) are particularly important in terms of garment image deformation. The influence of different positions and numbers of fashion landmarks on deformation will be discussed in Section 4.3.2.2. In addition, 2-step deformation strategies operate or function together on the images to properly deform the G_s and prevent excessive deformation while transferring the garment fit (wearable looseness and wearing position) of the H_s onto the H_t . The effect of different deformation algorithms on the garment deformation will be discussed in Section 4.4.3.3.

4.4.3.2 The influence of fashion landmarks on garment deformation

Fashion landmarks represent the clothing shape and boundaries and serve as control points to guide garment deformation. In this section, the differences between the landmarks located by the proposed *Model i* and two other landmark detection methods in controlling clothing deformation are compared. Fashion landmarks detection methods have benefited widely from the rapid advancements in deep learning-based models. In (Guo et al., 2016), only 4~8 landmarks are learned from the sparse annotated dataset despite high detection accuracy. Due to sparse landmarks, the deformation of important parts of the garment (such as sleeves, chest, waist, bottom hem etc.) cannot be properly controlled, thus affecting the try-on effect. The results show in rows (i) to

(iii) of column (e) of Figure 4-7, reflect this problem in particular. In (Ge et al., 2019), fashion landmarks are densely labeled, but the relationship between garment and body is not considered, resulting in garment deformation that is not body-aware related. As shown in Figure 4-7, the source garment is deformed according to the landmarks without considering the layout of the garment in each part of the target person (chest, sleeve part, waist, *etc.*). In addition, since this work considers the problem of self-occlusion or other objects occluding clothes. Garment areas are polygonal closed areas formed by connected landmarks, which contain occluded texture noise. These texture noises can negatively impact visual fitting. As shown in Figure 4-7, rows (iii) (iv) and column (h), the VTON results have texture interference beyond the try-on field (e.g., skin, belt and jeans texture, *etc.*). The proposed *Model i* takes clothing segmentation and garment-body relationship as necessary prior to locate fashion landmarks controlling garment deformation.



Figure 4-7 Virtual try-on effects of different fashion landmark detection methods: columns (a), and (b) show the source and target garments; columns (c), (f), and (i) show the fashion landmarks using the DFA framework, Match-RNN and the proposed Model i column (d), (g), (j) show segmented garment using these methods; column (e), (h), (k) show virtual try-on effects based on landmarks and segmented garment.

4.4.3.3 The deformation scales of different deformation algorithms

Two deformation strategies (TPS deformation and body-aware garment deformation) work together on images after locating the relevant fashion landmarks. This section compares the effects of the TPS algorithm, MRLS algorithm and TPS+MRLS algorithms on garment deformation in virtual try-on, as shown in Figure 4-8. For all the images, the TPS algorithm cannot control the internal image deformation well with the same control points and the same deformation scale. For example, the text on the T-shirt in the first row of Figure 4-8, the two fabric splices at the bottom hem of the garment in the second row, and the pattern deformation of the plaid T-shirt in the third row. In contrast to this, the MRLS algorithm makes reasonably constrained the internal pattern deformation by controlling the mesh deformation. Nevertheless, the large deformation at the hem based on the MRLS algorithm will produce unsmooth curves when comparing the garment sizes with huge differences. When TPS+MRLS work together on the source garment, they not only prevent excessive deformation of texture details within the garment, but also produce the natural, non-rigid deformation of the garment's outline.

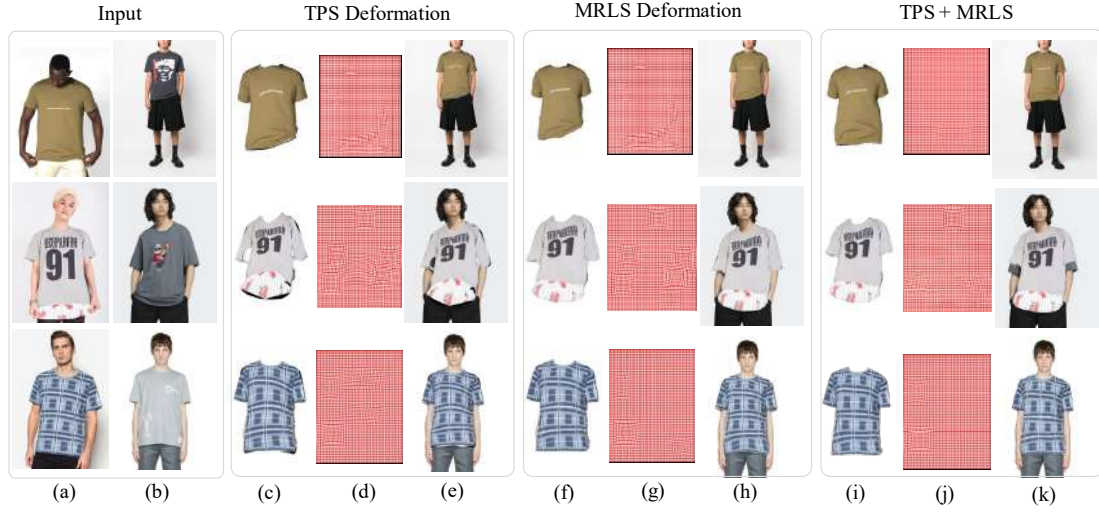


Figure 4-8 Virtual try-on effects of different deformation algorithms: columns (a), and (b) show the source garment and target garments; columns (c), (f), (i) show garment deformation using the methods of TPS, MRLS and TPS+MRLS; column (d), (g), (j) show the clothing deformation scales of; columns (e), (h), (k) show virtual try-on effects.

4.4.4 Experimental Results

4.4.4.1 Qualitative results

Figure 4-9 visually shows some examples of the proposed BSLT-VTON and other VTON methods. It is observed that, compared with other methods, the proposed method consistently achieves the best try-on effects for various top garments. More specifically, GAN-based methods like PFAFN (Ge et al., 2021b), Flow-style VTON (He et al., 2022) often fail to generate fitted garments for users when faced with target images with complex poses and backgrounds, especially for the shoulder and sleeve areas. Although DiOr (Dubey, 2021) is a bit better than the first two in terms of preserving the characteristics and textures of the clothing, it fails to generate realistic target human

bodies. While other LDM-based methods including LaDI-VTON (Morelli et al., 2023) and OOTDiffusion (Xu et al., 2024). LaDI-VTON (Morelli et al., 2023) tends to lose some garment details such as complicated patterns, logos or color fidelity, making the outfitted images look unrealistic. The overall performance of OOTDiffusion (Xu et al., 2024) is satisfactory but it suffers from detail issues, such as weird limbs (5th and 8th rows of (h)) and unnatural-looking garments (6th and 9th rows of (h)). It is noted that MCG VITON (Fang et al., 2022) and the proposed BSLT-VTON are based on the deformation method to achieve virtual try-on, which better retains the clothing texture details. However, in MCG VITON (Fang et al., 2022), the body and garment are often mismatched, causing the virtual avatar to either be inadequately covered or excessively covered by the deformed garments. This method estimates the target human body shape by reconstructing 3D Avatar based on SMPL (Loper et al., 2015), which cannot guarantee the accuracy of virtual human in complex poses (2nd, 6th and 8th rows of (f)). In contrast, the BSLT-VTON is superior in (1) fine-grained garment details reservation, and (2) controllability of garment deformation for various trial wearers. Since the proposed *Model ii* uses a small incremental clothing deformation method to prevent excessive deformation of clothing while conveying texture well. In contrast to generative model-based VTON methods that struggle to eliminate texture artifacts in the output, the deformed clothing in this study shows minimal visual discrepancies in texture compared to the input garment. In addition, the proposed *Model i* estimates different clothing size variables based on body frame and creates corresponding proportional relationships, which can synthesize faithful and suitable fitting suggestions

for users of various body types.

4.4.4.2 Quantitative results

To ensure a fair comparison, these pretrained (publicly released) models are applied to the BLST dataset in all cases. The same configuration is set for the different methods. Figure 4-9 presents the quantitative evaluation results by comparing SSIM, LPIPS, FID and KID scores. It is observed that deformation-based methods such as MCG VITON (Fang et al., 2022) and GAN-based models like PFAFN (Ge et al., 2021b) achieve relatively favorable SSIM scores, suggesting they effectively preserve some structural information of the input images. However, the images they generate lack detailed fidelity, resulting in lower LPIPS scores compared to ours. The previous GAN-based model like DiOr (Dubey, 2021) along with LDM-based methods like LaDI-VTON (Morelli et al., 2023) and OOTDiffusion (Xu et al., 2024) generate more realistic images that reflect their lower FID and KID scores. However, they fall short in accurately restoring the detailed features of the clothing or the target human.

In comparison, the BSLT-VTON can deform new garments toward the user shape while transferring the garment fit in terms of looseness and length of the source garments while keeping their textures content unchanged. And thus, substantially outperforms other methods on the other three metrics (SSIM, LPIPS and FID) while obtaining comparable KID scores to the OOTDiffusion (Xu et al., 2024). These results are attributed to the realistic fashion landmarks deformation model, which incorporates the

relationship between clothing and the human body, as well as the accurate garment geometric deformation model utilized in the proposed method.

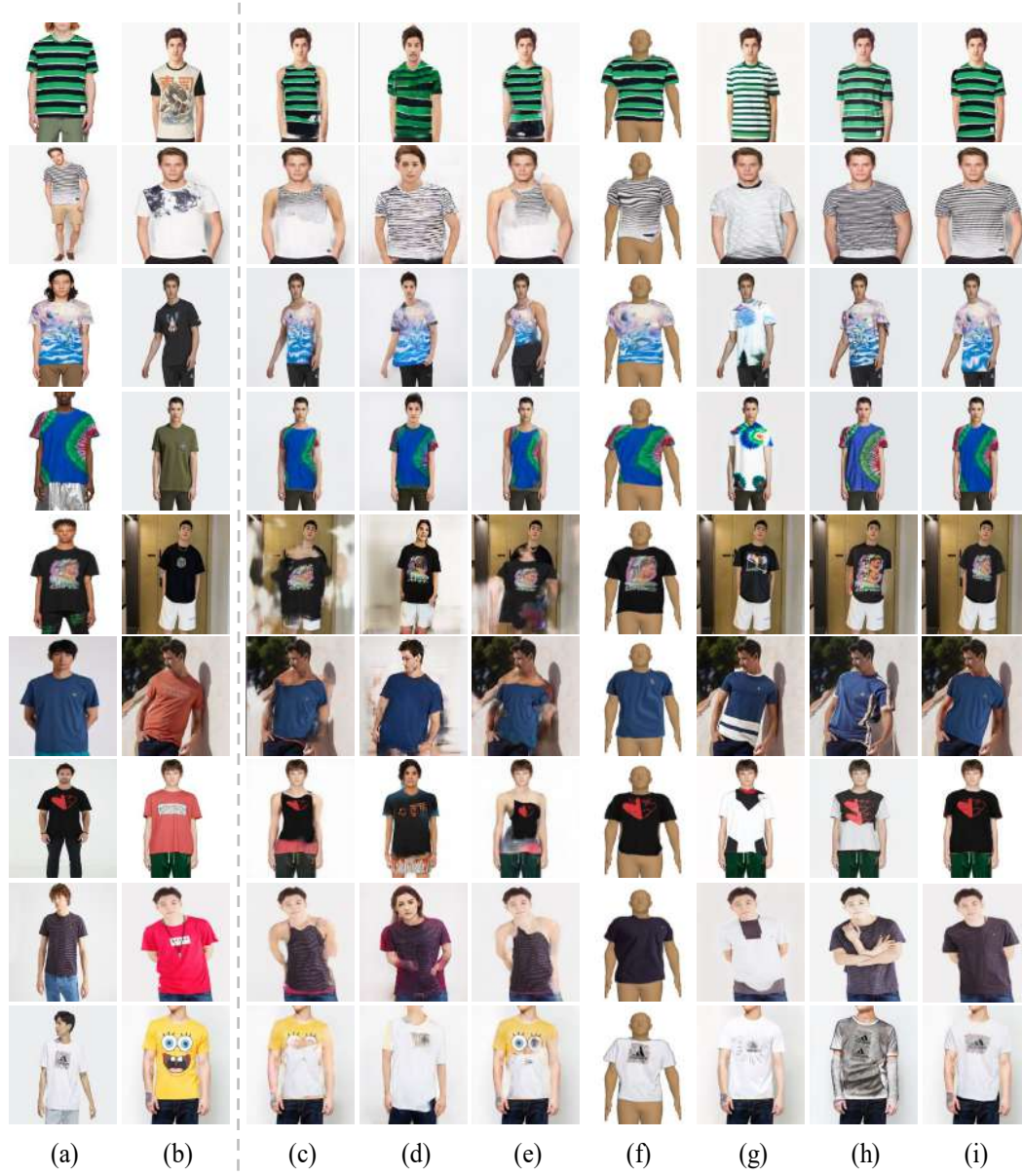


Figure 4-9 Qualitative comparison of SOAT VTON methods along the different column as follows of (c) PFAFN's, (d) DiOr's, (e) Flow-style VTON's, (f) MCG VITON, (g) LaDI-VTON, (h) OOTDiffusion and (i) the proposed BSLT-VTON, respectively. Columns (a) and (b) are the input source garment and target human, respectively.

Table 4-3 Quantitative comparison the performance of different virtual Try On method.

	PFAFN (Ge et al., 2021b)	DiOr (Dubey, 2021)	Flow-style VTON (He et al., 2022)	MCG VITON (Fang et al., 2022)	LaDI-VTON (Morelli et al., 2023)	OOTDiffusion (Xu et al., 2024)	The proposed BSLT-VTON
↑SSIM	0.5643	0.5428	0.5612	<u>0.5727</u>	0.5532	0.5422	0.5734
↓LPIPS	0.4559	0.4922	0.4647	0.5118	0.4761	<u>0.4463</u>	0.4444
↓FID	68.0113	37.6642	76.1348	156.62	39.7674	<u>23.2647</u>	19.5742
↓KID	0.070	<u>0.016</u>	0.085	0.222	0.023	0.003	0.003

4.4.5 Limitation

Although the BSLT-VTON proposed in this chapter faithfully and appropriately transforms different styles of input clothing (tight-fitting, regular fitting, over-fitting) onto the target person's body, reflecting the try-on effect of different styles. Nevertheless, this model exhibits limitations concerning human body shapes, particularly evident in the following areas. First, the data we collect tends to favor conventional body forms, lacking representation for extreme or specific body shapes like excessively slim, overweight individuals, or those with disproportionate body sizes. This data bias could compromise the model's efficacy in producing accurate results for non-standard body shapes. For instance, garments may not conform well to the body, appearing misaligned or distorted. Second, 2D virtual try-on rely solely on flat images and fail to capture the 3D geometric nuances (such as depth and curvature) of the human body, hindering their ability to replicate the interaction between clothing and body contours. For example, when dealing with form-fitting attire, 2D VTON techniques struggle to authentically simulate the effects of stretching and folding on diverse body shapes, leading to an unnatural appearance. These technical limitations may result in the generated fitting outcomes lacking realism, diminishing user trust and satisfaction with the system. Consequently, users may notice disparities between their virtual try-ons and actual clothing fits, potentially reducing system usage frequency.

To overcome these limitations, we proposed a 3D visual try-on system in Chapter 5 as a complement to the limitations of the 2D virtual try-on system. This enhancement aims

to boost the overall efficacy and versatility of the virtual try-on system, catering more effectively to user requirements.

4.5 Chapter Summary

This chapter proposed a body size dependent fashion landmark transformation virtual try-on method that effectively transfers the wearable looseness and wearing position between the source garment and the target person benefiting from fashion landmark localization model (proposed in Chapter 3) and human body-aware garment deformation model. Compared to the previous works, the proposed BSLT-VTON faithfully and appropriately deforms input garments of different styles (tight-fitting, regular fitting, over-fitting) onto the target person's body, reflecting the try-on effect of different styles. This is mainly because the fashion landmark localization model can explicitly model the relationship between clothing and the human body. It achieves this by estimating different clothing size variables based on human body frame and establishing corresponding ratio relationships. In addition, the visual results are superior in preserving texture details even for complex texture patterns, as it uses a 2-step small-increment clothing deformation strategy to prevent excessive deformation of clothing during the try-on process. Detailed analysis of the newly collected BSLT dataset that BSLT-VTON is realistic and rational in terms of garment deformation with attention demonstrated to the differences in body size and shape.

Chapter 5. Tileable Texture Generation for 3D Visual Try-on

5.1 Introduction of tileable texture generation

Seamless textures are ubiquitous in a variety of applications (Li et al., 2024a; Liu et al., 2020), such as in digital content creation for 3D modeling, animation, video games, AR/VR experiences, and other visual media. They effectively enrich the overall aesthetics of digital objects and enhance the realism of virtual environments. A seamless texture is a texture that the visual elements are systematically arranged into a semantically continuous pattern within the boundary/seam area of the texture. An important property of seamless textures is that they can be tiled or repeated across a larger surface or background without any visible seams or edges between individual tiles. Seamless textures are, therefore, widely used in cloth digitization, which serves as a key component in the metaverse. Figure 5-1 shows various digital 3D garments with textures generated by the proposed pipeline.

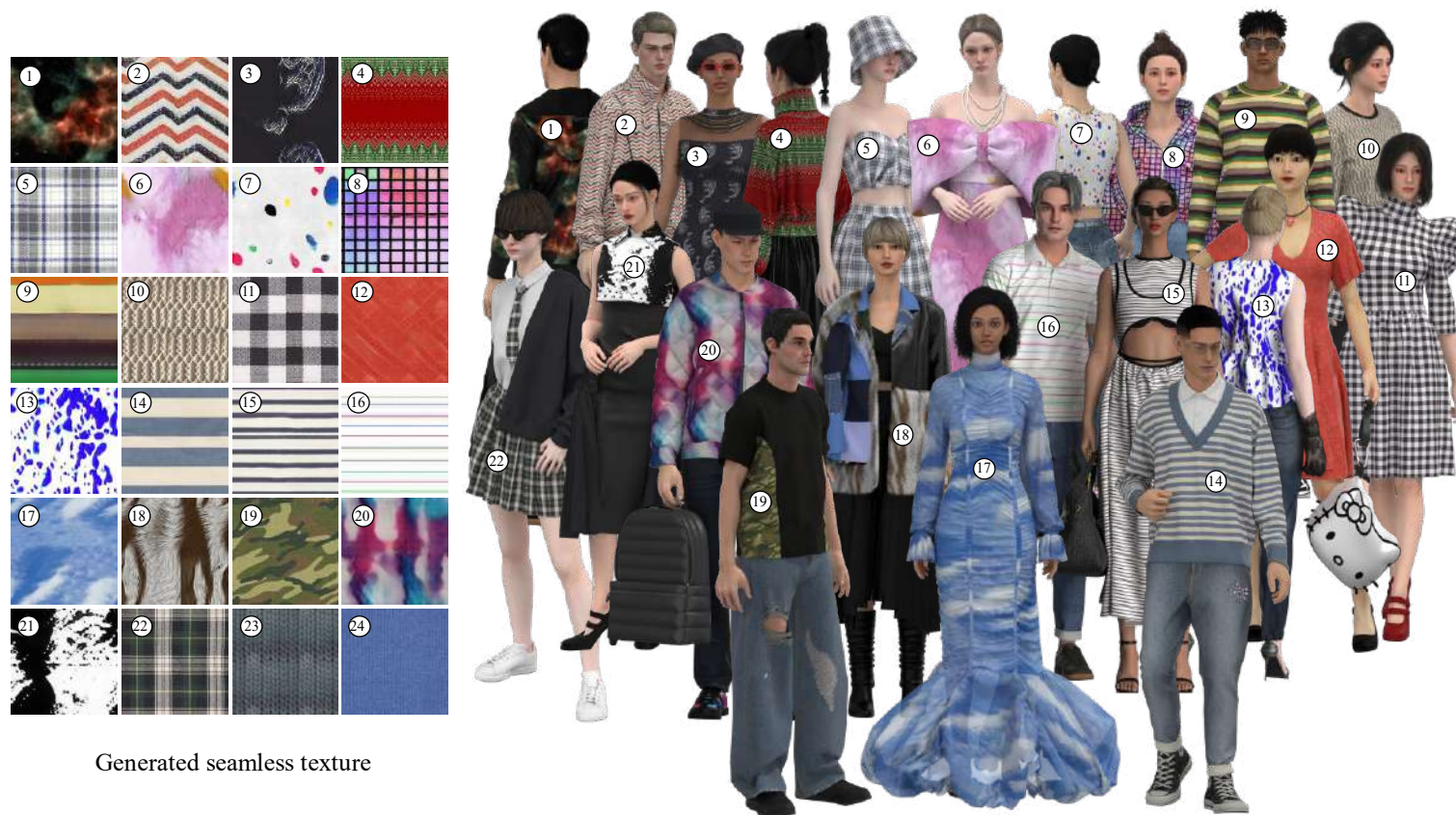


Figure 5-1 A cloth digitization pipeline, where high-quality seamless textures, generated by TiPGAN, can be tiled onto 3D garments. Here, 3D meshes with rich texture designs, such as seamless patterns of ④, ⑨, ⑭, ⑬, ①⑥, and materials, e.g., seamless normal map of knitted rib ⑬ or knitted cotton ⑭ are rendered, allowing textures to be rendered on different materials while maintaining consistent appearance.

Generating seamless texture, however, is not trivial, because the edges of the texture need to be carefully designed to create a continuous and uninterrupted pattern structure when repeated in any direction. Over the years, different strategies have been explored, with many methods developed. In earlier times, such textures were usually handled by skilled artists manually, through image transformation (Petrou & Petrou, 2010), or with the help of certain algorithms, like Image Quilting (Efros & Freeman, 2003; Efros & Leung, 1999) or Patch Match algorithms (Barnes et al., 2011; Barnes et al., 2009; Barnes et al., 2010). These methods, involved specially designed matching algorithms, often require manual intervention and have limited efficiency. The rapid development of deep learning models in the past decade has significantly advanced the work in texture synthesis. For example, Gatys et al. (2015) proposed an exemplar-based approach to texture synthesis using convolutional neural networks (CNNs). Instead of relying on manually designed filters as in traditional approach, their method created variations of a reference texture by employing Gram matrices from the feature map extracted by a pre-trained CNN encoder. To enhance the perceptual similarity between the synthesized and reference textures, Guo et al. (2016) integrated spectral information with CNN features. Sendik and Cohen-Or (2017) improved texture synthesis by leveraging spatial correlations. To model the variety of textures, subsequent methods (Ulyanov et al., 2017; Xie et al., 2016) adopted a generative approach that synthesis an image from random noises. These methods, however, primarily concentrate on producing input-like outputs by smoothing the transitions between texture blocks. The

generated textures of these methods have limited texture tileability, pattern continuity, and semantic consistency, as illustrated in Figure 5-2.

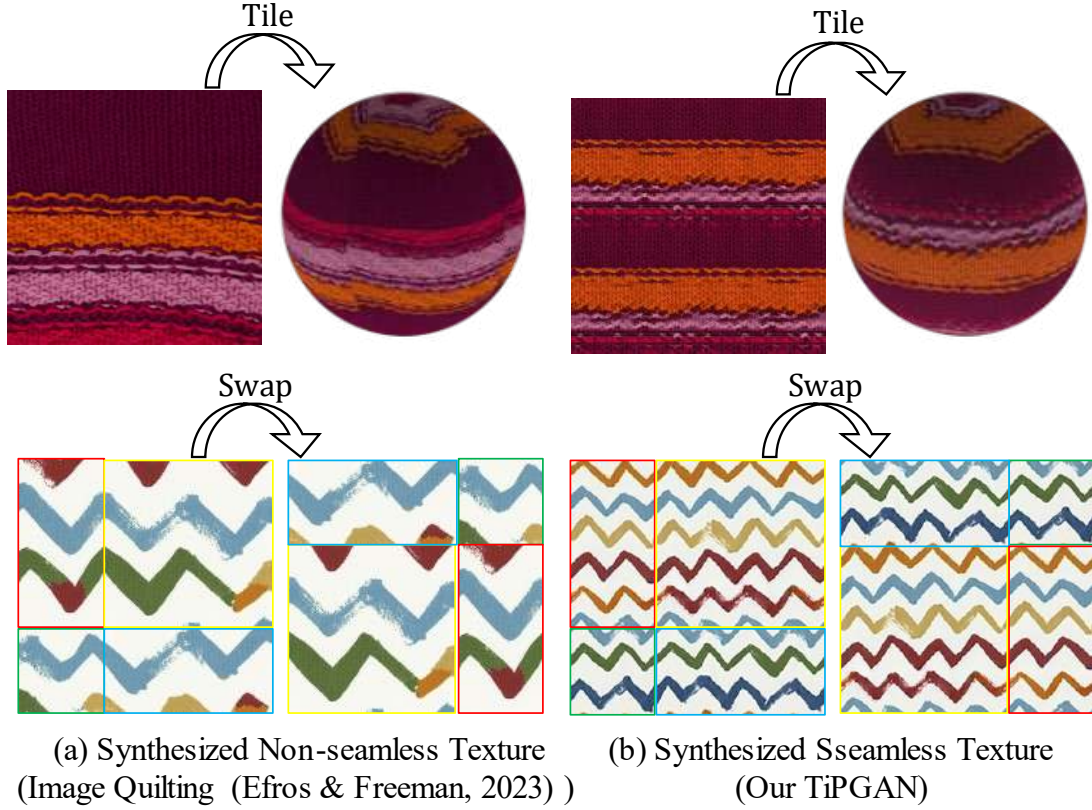


Figure 5-2 Texture continuity: (a) non-seamless vs. (b) seamless after tiling or swapping.

To synthesize seamless texture, other studies make use of periodic priors. For example, Bergmann et al. (2017) replaced initially randomized noise with sine waves to ensure the texture’s continuity. Paz et al. (2024) enhanced the continuity by incorporating Harmonics and Fourier features. Lin et al. (2023) integrated a multi-scale texture broadcasting module into the StyleGAN-V2 framework (Karras et al., 2020) to improve

the visual quality of the synthesized textures. Although periodic modeling enhances texture continuity and internal consistency of the outputs, these methods fail to ensure edge continuity, resulting in textures with noticeable artifacts (i.e., discontinuities) at the tiled boundaries. Other studies focus on latent space manipulation. Fröhstück et al. (2019) generated a large set of latent features and synthesized images by selecting and combining those most similar features. Li et al. (2020b) identified the optimal tiling positions within the feature space by processing image borders. Moritz et al. (2017) transformed photos into tileable textures through texture stationarization, optimizing for uniformity and repeatability. Rodriguez-Pardo and Garces (2022) employed feature repetition to achieve seamless transitions. These latent manipulations are effective in dealing with regular, near-regular, stochastic, near-stochastic patterns, but the feature continuity constraints of irregular, recurring patterns in the wild are insufficient.

The motivation to enhance texture tileability originates from the inherent nature of seamless textures, namely the homogeneous continuity in the spatial structure across all directions. This fundamental character enables a texture to be tiled infinitely without visible seams. Additionally, seamless textures can be divided into arbitrary 2×2 patches that preserve their tileability even after being swapped diagonally. This feature is illustrated in the second row of Figure 5-2. This capability inspired the development of TiPGAN, a Generative Adversarial Network (GAN) model designed to generate **tileable patterns**. The proposed TiPGAN exploits the above-mentioned intrinsic to

generate seamless textures. This is achieved through two key modules: Patch Swapping Module (PSM) and Patch Tiling Module (PTM). Initially, a patch randomly cropped from a reference texture is encoded into the latent space. The PSM then diagonally swaps patches that have been randomly split. To ensure seamlessness, a residual block merges the swapped features, eliminating any visible seams. Following this, the PTM duplicates the feature map both horizontally and vertically. Although this process might disrupt the texture's inherent pattern, the reference texture acts as a ground truth, guiding the continuity of the texture's inner pattern. This approach proves effective as the reference image inherently displays continuity within its inner region, ensuring the generated texture's inner pattern remains consistent. Furthermore, if the inner pattern remains continuous after the processes such as swapping and tiling, the resulting textures will be seamless.

To evaluate the effectiveness of TiPGAN, it is comprehensively evaluated and compared with state-of-the-art (SOTA) methods using image quality metrics, including structural similarity (SSIM) (WANGZ & Sheikh, 2004), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) and Single Image Fréchet Inception Distance (SI-FID) (Shaham et al., 2019). Although these metrics can evaluate perceptual consistency, they are not designed to assess tileability. To address this gap, a new metric is introduced, Relative Total Variation (RTV), in this study. This is inspired by the fact that the traditional total variation (tv) score (Rudin et al., 1992)

measures the smoothness within an image, with a higher value of the tv score indicating poorer smoothness. Poor continuity within an image’s pattern is reflected in a higher TV score. The RTV metric builds on an observation that any seamless textures maintain continuity upon tiling. By comparing the tv scores of a texture before and after tiling and analyzing the differences, we can measure tileability. A lower RTV score indicates better tileability. In summary, the key contributions of this study are as follows:

- This chapter proposed a novel way of exploiting intrinsic priors to improve seamless texture synthesis. Notably, the proposed TiPGAN is robust in dealing with special textures, such as irregular, unflatten patterns.
- The proposed TiPGAN enhances the cloth digitization pipeline by allowing easy and flexible mapping to ensure accurate visual representation of the fabrics.
- An intuitive metric is introduced, Relative Total Variation (RTV), to quantify texture pattern tileability.
- Extensive qualitative and quantitative experiments demonstrate that the synthesized seamless textures maximize the visual similarity with the reference texture.

5.2 Related Work

5.2.1 *Texture Synthesis*

Significant advances in texture synthesis research have given rise to a large number of synthesis techniques. Among traditional ones, exemplar-based texture synthesis (parametric and non-parametric) methods strives to synthesize textures that closely resemble the texture of a provided input sample. Parametric methods utilize statistical analysis of the reference texture to generate a new texture by imposing various statistical constraints (Galerne et al., 2010; Liu et al., 2023b; Rabin et al., 2012). The key limitation of such approach is that they cannot handle highly complex or abstract textures that are lack of clear statistical regularities (Akl et al., 2018). Non-parametric methods have achieved impressive results in generating high-quality textures with different characteristics, including some classic algorithms such as Image Quilting (Efros & Freeman, 2023; Efros & Leung, 1999), Graph-cuts (Kwatra et al., 2003), Genetic algorithms (Dong et al., 2007) and Patch Match (Barnes et al., 2011; Barnes et al., 2009; Barnes et al., 2010). These synthesis methods encourage the preservation of fine local details, emphasizing the ‘stitching’ of discontinuities (Efros & Freeman, 2023; Kwatra et al., 2003) between patches and fostering overall similarity (Kopf et al., 2007). However, they either require additional manual intervention or the synthesized textures lack global coherence, since synthesizing textures in a patch-by-patch manner may exhibit obvious seams over larger areas. Among different deep learning-based methods, the texture synthesis methods leveraging pre-trained networks or feedforward generative networks are the most groundbreaking. For example, Gatys et al. (2015) used a pre-trained VGG-19 network (Simonyan & Zisserman, 2014) to extract texture features. Ulyanov et al. (2016) attained comparable performance by employing a fast

feedforward generative network. Li et al. (2017) advanced a feedforward generative network to efficiently synthesize multiple textures within a single network and enable meaningful interpolation between them. There are also some studies on synthesizing textures based on implicit neural representation (INR) methods (Chen & Zhang, 2019; Park et al., 2019; Sitzmann et al., 2020). Henzler et al. (2020) and Portenier et al. (2020) employed raw pixel coordinates directly as input to predict the RGB values of the texture. Chen et al. (2022) expanded texture synthesis into the continuous domain through the implicit representation of texture. Recently, the diffusion model (DM) (Ho et al., 2020; Sohl-Dickstein et al., 2015) has achieved good results in image generation tasks (Dhariwal & Nichol, 2021). For example, Chen *et al.* Chen et al. (2023b) proposed text-driven texture synthesis, which is actually an image inpainting architecture based on the diffusion model. Most existing DM-based image synthesis techniques cannot be directly applied to texture synthesis because their common underlying concept (of manipulating attention maps) is not applicable to the texture domain (Guerrero-Viu et al., 2024). Controllable texture synthesis remains a challenging task. We therefore look to GANs (Goodfellow et al., 2014) for a solution, mainly because GAN’s architecture allows meaningful latent space manipulation, proven effective for different texture synthesis tasks, such as non-stationary texture (Zhou et al., 2018), tileable texture (Zhou et al., 2022) and multitexture synthesis (Xie et al., 2023a). The popularity of generative models has surged, largely due to the widespread applications of GANs and the various GAN variants. For example, MGAN (Li & Wand, 2016) used a pre-trained VGG network to capture the feature statistics of Markov blocks, improving efficiency while

maintaining texture visual credibility. SGAN (Jetchev et al., 2016) fused multiple source images in complex textures by expanding the noise distribution space. Zhou et al. (2018) generated extended textures iteratively by randomly cropping the reference texture. Yu et al. (2019a) performed linear interpolation in the latent space to fuse two different textures. TileGAN Frühstück et al. (2019) searches for the nearest latent tiles using neighborhood similarity matching and then feeds these pre-stored latent tiles into a trained generator. This generator combines the various tiles produced by the GAN to synthesize images. SinGAN (Shaham et al., 2019) was a fully convolutional architecture based on GANs. This trained model is capable of capturing the internal distribution of patches within an image and generating high-quality, diverse samples that retain the same visual content as the input. Some subsequent work, such as ConSinGAN (Hinz et al., 2021) and PetsGAN (Zhang et al., 2022b), aimed to accelerate the learning efficiency of SinGAN (Shaham et al., 2019) to quickly achieve high-quality texture optimization. LocoGAN (Struski et al., 2022) was another fully convolutional GAN trained on fixed-size sub-images (rather than on full images), enabling the model to generate fully periodic or nearly periodic ‘infinitely long’ textures. LocoGAN utilized local learning and incorporated a positioning channel to generate specific part of the image. These studies confirm that GANs are capable of and have known advantages in the field of texture synthesis.

5.2.2 Seamless Texture Synthesis

In terms of seamless textures, early methods endeavor to minimize visible seams by optimizing patch selection and texture blending (Cohen et al., 2003; Efros & Freeman, 2023), or by wrapping around the texture seams (Birsak et al., 2013), or by the use of Fourier analysis (Holten et al., 2006). Because these methods are mostly formulated as optimization problems matching between inputs and outputs, they are slow and have limited capability of complex texture synthesis. Non-parametric texture synthesis methods (e.g. (Moritz et al., 2017)) focus on generating tileable textures by maximizing texture smoothness through extrinsic measures. Deliot and Heitz (2019) generated new tileable texture mixtures via histogram-preserving patch-based synthesis. Parameter substitution methods are employed. For example, Rodriguez-Pardo et al. (2019) utilized deep features from a pre-trained CNN (Krizhevsky et al., 2012) to identify repetitive patterns in images and then created tileable images by seamlessly blending their boundaries. The creation of seamless textures can also be accomplished through particular image parameterization and neural network architectures, such as neural cellular automata (Niklasson et al., 2021). Some other work, such as ControlMat (Vecchio et al., 2023), synthesized material textures using diffusion models. They modified the sampling process to integrate multi-scale information, implemented rolling diffusion to ensure seamless tiling and patch diffusion to produce high-resolution textures. However, optimizing these models typically requires significant training time and computational resources. GANs-based seamless texture synthesis methods attempt to further optimize the quality of generated results. PSGAN (Bergmann et al., 2017) was the first attempt to generate periodic textures by

incorporating periodic patterns with random phase terms into the generator. (Jetchev et al., 2017) fine-tuned the latent noise space based on PSGAN (Bergmann et al., 2017) to generate textures with consistent content. User-Controllable GANs (Alanov et al., 2019) incorporates convolutional encoder to (Bergmann et al., 2017) in order to produce textures according to user preferences. Subsequently, the flexible feature decoupling performance of the StyleGAN family (Karras et al., 2021; Karras et al., 2019; Karras et al., 2020) has drawn a great deal of attention. For example, TileGen (Zhou et al., 2022), based on StyleGAN-V2, combined with a conditional architecture similar to CollageGAN (Li et al., 2021c), that only output tileable texture maps with material maps. Lin et al. (2023) integrated a multi-scale texture broadcast module to StyleGAN-V2 framework, enhancing seamless texture synthesis by injecting appropriate perceptual biases into the synthesis procedure. Frühstück et al. (2019) edited latent code at lower resolution levels to address seamless texture scalability issues. Rodriguez-Pardo and Garces (2022) achieved smooth transition areas by randomly cropping texture and repeating features. Paz et al. (2024) explored sinusoidal neural networks to represent periodic tileable textures, adding a regularization term based on Poisson’s equation to enforce this constraint. These methods have the ability to generate seamless textures by explicitly incorporating periodic patterns into generators or latent spatial modifications. However, for complex cases, such as irregular, non-flattened patterns, the output of their models is unsatisfactory. We, therefore, enhance the latent space manipulation by introducing a latent swap module and a latent tiling module to ensure compositing of seamless textures.

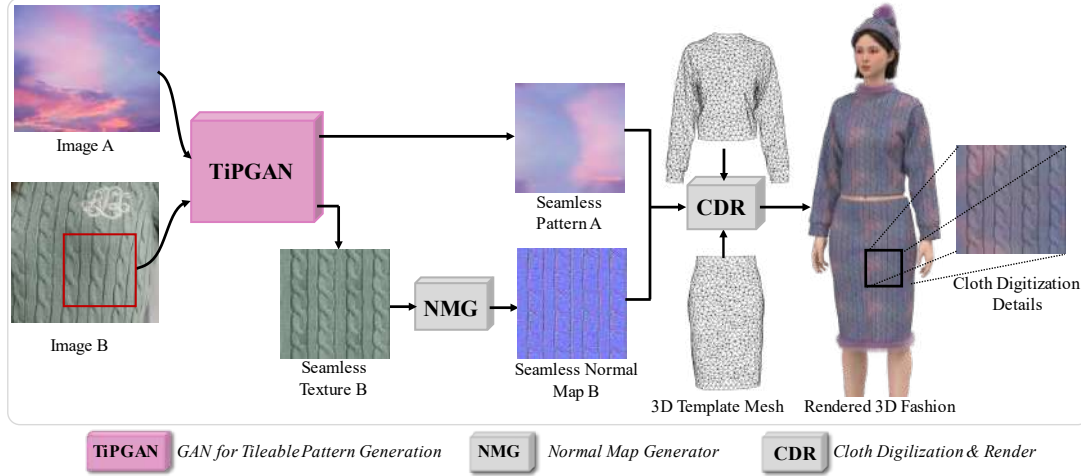


Figure 5-3 Realistic 3D seamless texture generation for cloth digitization application.

5.3 Method

This chapter proposes TiPGAN a GAN-based network to generate high quality seamless textures that retain the visual characters of the source textures (see Figure 5-4). Moreover, a cloth digitization pipeline, as shown in Figure 5-3, is also proposed, and this pipeline incorporates TiPGAN as a key step to effectively replicate the visual aspects of color and material textures (Deng et al., 2024) in 3D scenes. In the following, details of the TiPGAN architecture are first provided in Section 5.3.1, followed by an explanation of the incorporated seamless texture intrinsics and the details of the TiPGAN model in Section 5.3.2.. Next, the new metric RTV, designed to evaluate seamless quality, is introduced in Section 5.3.3. Lastly, the cloth digitization pipeline designed for texture 3D visualization is described in Section 5.3.4.

5.3.1 TiPGAN Architecture

Figure 5-4 shows the architecture of TiPGAN, which the generator consists of four main components: (i) the texture style encoder, (ii) patch swapping module, (iii) patch tiling module and (iv) the tileable decoder.

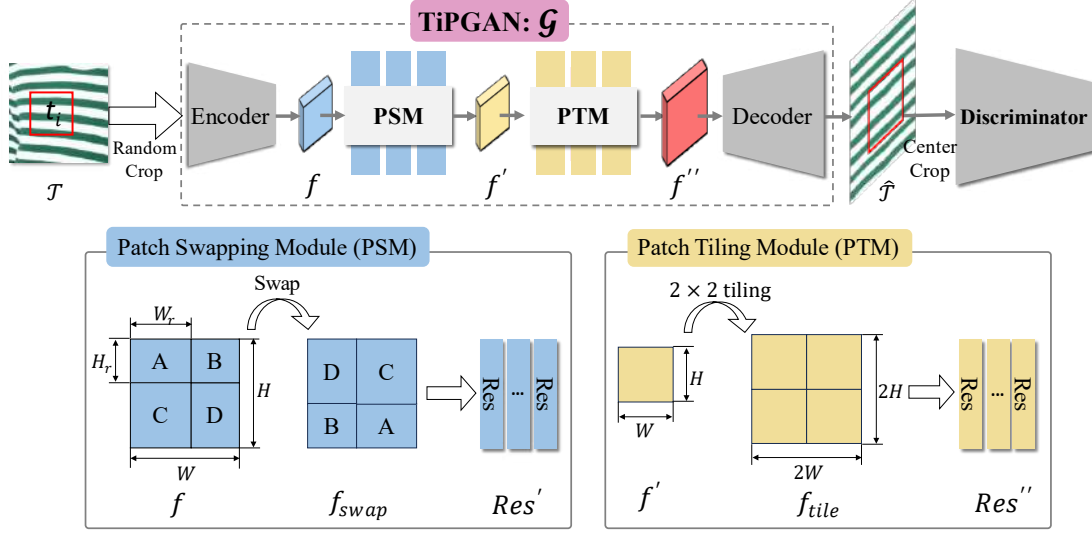


Figure 5-4 Architecture of TiPGAN, including the Patch Swapping Module and Patch Tiling Module.

Without a large-scale supervision data, a self-supervised learning approach (Shaham et al., 2019; Zhou et al., 2018) is employed for training TiPGAN. A PatchGAN Discriminator (Isola et al., 2017) is adopted to classify whether the generated texture is real, and TiPGAN is trained following typical adversarial training schemes (Goodfellow et al., 2014).

Given a target texture $\mathcal{T}$, the TiPGAN $\mathcal{G}$ is trained using randomly cropped patches $\mathbf{t}_i$ from $\mathcal{T}$ to reconstruct $\hat{\mathcal{T}}$, as described by the following Eq.:

$$\hat{\mathcal{T}} = \mathcal{G}(\mathbf{t}_i) \quad (5-1)$$

By reconstructing a texture from its small patches, model $\mathcal{G}$ learns the internal characters of $\mathcal{T}$. Specifically, $\mathcal{G}$ employs a classical convolutional neural network architecture that downsamples an image twice and upsamples it threefold. The design of the encoder-decoder structure for TiPGAN is inspired by the work of (Zhou et al., 2018). Through this structure, the model learns to expand a given texture into a non-stationary texture. Nevertheless, a non-stationary texture is not inherently tileable. Previous work (Frühstück et al., 2019; Karras et al., 2017) demonstrated that spatially concatenating various latent spaces within the generator can generate textures that integrate the visual details of the patches while ensuring smooth transitions between them. Building upon this idea, the seamless tiling properties of generated textures are proposed to be enhanced by incorporating intrinsic priors into two specially designed modules.

5.3.2 *Seamless Intrinsic Priors Guided Synthesis*

Patch Swapping Module (PSM). Seamless textures have an intrinsic property that enables the interchange of interior patches without compromising their seamless integrity. This is attributed to the lack of discernible boundaries or discontinuities between different parts of the texture, allowing for individual areas to be freely tiled while maintaining overall visual coherence. Based on this observation, a random patch swapping module is proposed to ensure swapping consistency by imposing a constraint.

For simplicity, when focusing solely on the dimensions of width and height, a feature map is denoted as $f \in \mathcal{R}^{W \times H}$. We partition this map by choosing a random center (W_r, H_r) :

$$\begin{bmatrix} f_a & f_b \\ f_c & f_d \end{bmatrix} = f \quad (5-2)$$

where the spatial dimensions for each segment of the feature map are specified as follows: $f_a \in \mathcal{R}^{W_r \times H_r}$, $f_b \in \mathcal{R}^{(W-W_r) \times H_r}$, $f_c \in \mathcal{R}^{W_r \times (H-H_r)}$, and $f_d \in \mathcal{R}^{(W-W_r) \times (H-H_r)}$.

The swapped feature map f_{swap} is produced by:

$$f_{\text{swap}} = \begin{bmatrix} f_d & f_c \\ f_b & f_a \end{bmatrix} \quad (5-3)$$

To ensure seamless integration at the splitting margin, residual blocks, Res' , are employed to fully merge the swapped feature map:

$$f' = \text{Res}'(f_{\text{swap}}) \quad (5-4)$$

Patch Tiling Module (PTM). When applying a seamless texture to a larger area, such as a cloth UV map, the most common tiling method involves repeatedly using a 2×2 tiling. If a texture is not tileable, artifacts will appear at the crossing regions. To improve the texture's tileability, 2×2 tiling in the feature space is introduced:

$$\begin{bmatrix} f_a & f_b \\ f_c & f_d \end{bmatrix} = f \quad (5-5)$$

where four copies of the same texture are seamlessly joined (see Figure 5-4). Initially, the latent field f' may exhibit significant discontinuities at the edges where the four

replicas meet. However, as the texture progresses through the network, these artifacts are progressively converted into seamless boundaries by the subsequent residual blocks and the tileable decoder. The additional residual blocks Res'' are expressed as follows:

$$f'' = \text{Res}''(f_{\text{tile}}) \quad (5-6)$$

Loss Function. The generator is modeled as a fully convolutional neural network. The encoder utilizes three standard convolutional blocks, each consisting of a convolutional layer, an instance normalization layer, and a ReLU activation layer. The spatial size is reduced by a factor of 8 compared to the input size. The PSM includes a patch swapping operation and one residual block. The PTM incorporates a tiling operation followed by six residual blocks. The decoder uses convolutional blocks similar to those in the encoder but with four blocks. This setup allows a 2×2 texture $\hat{\mathcal{T}}$ to be generated from $\mathbf{t}_i$ in a single forward pass through the network. Four types of losses are incorporated in the training, and the loss function for optimizing generation is presented as follows:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{L_1} + \lambda_2 \mathcal{L}_{\text{adv}} + \lambda_3 \mathcal{L}_{\text{style}} + \lambda_4 \mathcal{L}_{\text{hist}} \quad (5-7)$$

where λ_s denote the respective weights for the L1 loss $\mathcal{L}_{L_1}$, adversarial loss $\mathcal{L}_{\text{adv}}$ (Goodfellow et al., 2014), style loss $\mathcal{L}_{\text{style}}$ (Gatys et al., 2015), and histogram loss $\mathcal{L}_{\text{hist}}$ (Afifi et al., 2021). We follow the loss combination outlined in SeamlessGAN (Rodriguez-Pardo & Garces, 2022). The adversarial loss $\mathcal{L}_{\text{adv}}$ ensures significant semantic consistency. As proven by Non-Stationary GAN (Zhou et al., 2018) and SeamlessGAN (Rodriguez-Pardo & Garces, 2022), it stands out as the key component

in the overall loss function. The L1 loss serves as a regularization technique to eliminate artifacts. The style loss (Gatys et al., 2015) contributes additional details to fully represent the texture. The histogram loss derived from the reference texture histogram enhances the realism of texture colors. The exact definition of each term in this paper is given as follows:

$$\mathcal{L}_{L_1} = \mathbb{E}_{\mathcal{T}, \hat{\mathcal{T}}} [\|\mathcal{T} - \hat{\mathcal{T}}\|_1] \quad (5-8) \text{ a}$$

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{\mathcal{T}} [\log \mathcal{D}(\mathcal{T})] + \mathbb{E}_{\hat{\mathcal{T}}} [\log (1 - \mathcal{D}(\hat{\mathcal{T}}))] \quad (5-9) \text{ b}$$

$$\mathcal{L}_{\text{style}} = \sum_l \|G_l(F(\mathcal{T})) - G_l(F(\hat{\mathcal{T}}))\|_F^2 \quad (5-10) \text{ c}$$

$$\mathcal{L}_{\text{hist}} = \frac{1}{\sqrt{2}} \cdot \frac{1}{N} \sum_i \left\| \sqrt{H(\mathcal{T}_i)} - \sqrt{H(\hat{\mathcal{T}}_i)} \right\|_2^2 \quad (5-11) \text{ d}$$

The symbols in the equations are defined as follows: $\mathcal{T}$ and $\hat{\mathcal{T}}$ denote the target and synthesized texture, $\mathcal{D}$ is a discriminator, F is a pre-trained encoder (e.g., VGG-16), G_l denotes the Gram matrix of features at layer l extracted by F , $H(\cdot)$ represents the histogram features computed from an input image in the RGBuv color space, i is the index of each pixel, and N represents the total number of pixels in the image.

Our discriminator utilizes the PatchGAN discriminator architecture (Isola et al., 2017; Zhou et al., 2018; Zhu et al., 2017a), which is also a fully convolutional network, with five downsampling convolutional blocks. It outputs patchwise discrimination with a patch size that is 1/32 of that of the input spatial size. The architecture of TiPGAN is designed based on the Non-Stationary GAN framework (Zhou et al., 2018).

5.3.3 Seamless Quality Metric – Relative Total Variation

In this study, we define the seamless quality as the ability of a texture to be tiled in a 2×2 manner without any artifacts appearing in the intersection regions. A texture lacking seamless quality exhibits a discontinuous pattern, resulting in visually unsmooth effects. To quantify this continuity, we introduce the Relative Total Variation (RTV) metric as a means to assess the smoothness of the tiled texture.

We first introduced the Total Variation quality (TV) (Rudin et al., 1992) which is utilized in the calculation of the proposed RTV, and it is defined as follows:

$$TV(I) = \sum_{x,y} \sqrt{(\Delta_x I(x,y))^2 + (\Delta_y I(x,y))^2} \quad (5-12)$$

where $I(x,y)$ refers to a pixel at position (x,y) of an image I . Δ_x and Δ_y denotes the horizontal and vertical gradients of I , and are calculated as follows:

$$\Delta_x I(x,y) = I(x+1,y) - I(x,y) \quad (5-13) a$$

$$\Delta_y I(x,y) = I(x,y+1) - I(x,y) \quad (5-14) b$$

The traditional Total Variation (TV) method involves computing the difference between each pixel. This pixel-by-pixel gradient calculation across the entire image can be highly sensitive to local variations, especially in textures containing repetitive patterns. To address this, we propose calculating total variation on a patch-wise basis rather than a pixel-wise basis. Specifically, we split the image into patches of size (p, p) , and we modify the Δx and Δy in Eq. 5-10 into the following:

$$\Delta_x I(x, y) = I(x + p, y) - I(x, y) \quad (5-15) \ a$$

$$\Delta_y I(x, y) = I(x, y + p) - I(x, y) \quad (5-16) \ b$$

The total variation is then calculated within these patches. In our experiments, we set $p = 16$. We conducted an ablation study on Sec. 5.4.4.2 to investigate how the choice of p influences the results.

Although the patch total variation improves its robustness to local texture variations, it does not address how these textures perform when they are tiled. To overcome this limitation, the patch total variation is proposed to be calculated before and after a texture undergoes 2×2 tiling. The underlying idea is that if a texture is truly seamless, its patch total variation would remain unchanged by 2×2 tiling. For an image I , the transformed image $\hat{I}$ can be obtained by center cropping the tiled texture:

$$\hat{I} = \mathcal{C}(\text{tile}(I)) \quad (5-17)$$

where $C(\cdot)$ denotes the center crop for aligning the size before tiling. The tiling process $\text{tile}(\cdot)$ is defined in eq. 5-12. RTV is computed from I to $\hat{I}$ and vice versa:

$$RTV_{I \rightarrow \hat{I}} = \left| 1 - \frac{TV(I)}{TV(\hat{I}) + \epsilon} \right|, RTV_{\hat{I} \rightarrow I} = \left| 1 - \frac{TV(\hat{I})}{TV(I) + \epsilon} \right| \quad (5-18)$$

where ϵ is a small constant for avoiding zero on the denominator. We set $\epsilon = 1e - 5$ in the experiments. The term $TV_{I \rightarrow \hat{I}}$ quantifies the difference moving from I to $\hat{I}$. Given the asymmetry of this process, to eliminate any bias, we adjust the RTV calculation as follows:

$$RTV(I) = 0.5 * RTV_{I \rightarrow \hat{I}} + 0.5 * RTV_{\hat{I} \rightarrow I} \quad (5-19)$$

5.3.4 Cloth Digitization Pipeline

This study proposes a novel cloth digitization pipeline empowered by seamless texture synthesis, as illustrated in Figure 5-3. Central to this pipeline is TiPGAN, a model that generates seamless textures while retaining the visual characters of the source textures, including colors and patterns. This pipeline not only generates high quality textures, which can be tiled without visible seams, but also the corresponding albedo and normal maps. The generated textures and maps are applied to 3D garment meshes by Physically Based Rendering (PBR), effectively replicating the visual aspects of color and material textures (Deng et al., 2024) and reflecting the quality of reference textures.

5.4 Experiments

5.4.1 Datasets and Implementation Details

Datasets and Data Processing. To examine the effectiveness of the proposed *Model iii*, both qualitative and quantitative evaluations are conducted on two texture datasets. The DTD dataset (Cimpoi et al., 2014) comprises 5,640 images across 47 varied texture classes. The PSU NRTDB near-regular texture dataset (Liu et al., 2015) covers textures ranging from completely regular through nearly regular to irregular. We collected a distinct set of images from these two datasets, characterized by diversity in regularity and content, specifically related to cloth patterns and materials, to showcase more fashionable and artistic cloth digitization pipeline.

Implementation Details. When training the model, the corresponding weights in Eq. (5-7) are set to $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 10$, and $\lambda_4 = 1e-4$ for $\mathcal{L}_1$ loss $\mathcal{L}_{L_1}$, adversarial loss $\mathcal{L}_{adv}$, style loss $\mathcal{L}_{style}$, and histogram loss $\mathcal{L}_{hist}$, respectively. These weights were chosen following prior works (Afifi et al., 2021; Rodriguez-Pardo & Garces, 2022) to ensure fair comparison with baseline methods. While a comprehensive exploration of these weights is beyond the scope of this study, preliminary observations indicate that the current settings provide a reasonable balance between different loss components and support the effectiveness of the proposed approach.

For optimization, we used the Adam optimizer with a learning rate of $2e-4$, $\beta_1 = 0.5$, and $\beta_2 = 0.999$ chosen for its proven effectiveness in similar tasks. Each experiment is

run for 50,000 iterations, with learning rate adjustments made at the 30,000 and 40,000 iteration marks by a factor of 0.2. All experiments are conducted on a single RTX 3090 GPU.

Regarding the selection of the hyper-parameter p in the Relative Total Variation (RTV) metric, it was initially set to 16, following empirical settings. Subsequent ablation studies revealed that a smaller p (e.g., 4) yields better results in most scenarios, as discussed in Section 5.4.4.2.

Comparative Studies. The TiPGAN is compared with multiple state-of-the-art (SOTA) seamless texture synthesis methods, including the following:

- Image Quilting (Efros & Freeman, 2023) a traditional patch-based method for image synthesis
- Histogram Blending (Deliot & Heitz, 2019) a classical Patch Match method
- PSGAN (Bergmann et al., 2017) a GAN-based model utilizing periodic priors
- Self-organizing (Niklasson et al., 2021) a GAN-based method adopting an NCA-based approach, and
- SeamlessGAN (Rodriguez-Pardo & Garces, 2022) a GAN-based model that uses adversarial expansion

For Image Quilting (Rudin et al., 1992) and Histogram Blending (Deliot & Heitz, 2019), their project codes are leveraged from the two papers to implement tileable texture synthesis and modified the output size and resolution of both methods to match our own. For PSGAN (Bergmann et al., 2017), that version of PSGAN that trained on a single image is utilized to approximate our training configuration. The network parameters are set exactly as described in (Bergmann et al., 2017). For Self-organizing (Niklasson et al., 2021), their model is trained using PyTorch for a total of 5000 iterations. The standard training settings are followed, starting with a learning rate of 0.002 and then reducing it by half in 2000 and 4000 iterations. For SeamlessGAN (Rodriguez-Pardo & Garces, 2022), a simplified version of the generator that only outputs a single seamless albedo map is used to ensure a fair comparison. In terms of quantitative metrics, SSIM (WANGZ & Sheikh, 2004), SI-FID (Shaham et al., 2019), LPIPS (Zhang et al., 2018) and Relative Total Variation (RTV), the new metric introduced in Section 5.3.3, are employed.

5.4.2 Quantitative Evaluation

Table 5-1 shows the quantitative evaluation results. TiPGAN performs the best in comparison to SSIM metric. This indicates that the PSM and PTM latent space operation modules allow seamless boundaries between tiling feature maps, outperforming other methods in terms of perceptual and texture structure similarity. Moreover, the proposed TiPGAN has the lowest SI-FID score compared to the previous

deep learning-based methods, including PSGAN (Bergmann et al., 2017), Self-Organizing (Niklasson et al., 2021) and SeamlessGAN (Rodriguez-Pardo & Garces, 2022), which means that the quality and fidelity for images generated by TiPGAN is the closest to that of the ground truths, with well learned global continuity, style, and semantic content inherited from source references. In terms of LPIPS metric, the proposed TiPGAN, SeamlessGAN (Rodriguez-Pardo & Garces, 2022), and Image Quilting (Rudin et al., 1992) rank the top three methods, showing that the patch-based method preserves the structure of the reference texture better than previous neural parameter models in terms of perceptual and structural similarity. More importantly, the proposed TiPGAN achieves the smallest RTV score, demonstrating that the synthesized textures of the proposed *Model iii* have the highest texture tileability.

Table 5-1 Quantitative comparison.

	SSIM↑	SI-FID↓	LPIPS↓	RTV↓
Image Quilting (Efros & Freeman, 2023)	0.1414	<u>1.1670</u>	0.4937	1.8268
Histogram Blending (Deliot & Heitz, 2019)	0.1257	12.8638	0.5460	1.0473
PSGAN (Bergmann et al., 2017)	0.1209	1.8550	0.5435	1.1363
Self-Organizing (Niklasson et al., 2021)	<u>0.1466</u>	1.6789	0.5466	<u>0.5815</u>
SeamlessGAN (Rodriguez-Pardo & Garces, 2022)	0.1462	12.8726	<u>0.4795</u>	1.6901
TiPGAN (Ours)	0.1496	0.7793	0.4704	0.5587

We bold the **best** results and underscore the second-best results.

5.4.3 Qualitative Evaluation

Figure 5-5 shows a gallery of seamless textures produced by the proposed TiPGAN, along with qualitative comparisons with other SOTA methods listed in Section 5.4.1.

Overall, the proposed method is capable of generating seamless textures across a variety of fabric categories, including highly irregular banded patterns (see Figure 5-5 (a, b, c), uneven knitted (d), and unflatten dotted patterns (e). Its performance remains robust even with natural textures, such as hive (f), floral (i), and peacock feather patterns (h).

The other methods struggle either to preserve perceptual consistency or to ensure tileability. For instance, Image Quilting algorithm (Efros & Freeman, 2023) for

generating seamless textures depended heavily on the pattern structure of the reference texture. Image Quilting (Efros & Freeman, 2023) can only handle tileable or near-tileable patterns. For textures that are irregular, however, this algorithm fails to produce seamless textures that meet satisfactory standards. Histogram Blending (Deliot & Heitz, 2019) is only compatible with tileable stochastic and near-stochastic textures, particularly those textures that appear as random noise when observed from a distance, as well as natural-looking materials that lack precise geometric shapes. The textures it produces often exhibit ‘repetitive texture stitching fusion artifacts’ in the details. PSGAN (Bergmann et al., 2017) performs well in keeping periodicity patterns, yet it struggles to accurately capture the finer details present in natural textures (e.g., Figure 5-5 (g, i)). The texture generated by Self-Organizing (Niklasson et al., 2021) can be tiled seamlessly on a global scale, and it excels in the reconstruction of patterns such as grid and polka dot. Nonetheless, this approach disrupts the inherent geometric structure of the texture pattern, which limits the application of the generated seamless texture in task such as real-world textures digitization for 3D presentation or visualization. SeamlessGAN (Rodriguez-Pardo & Garces, 2022) excels at retaining the semantic characters of the reference texture but falls short in achieving satisfactory tileability performance for textures with irregular patterns or unflatten patterns. The qualitative comparison results in Figure 5-5 are consistent with that of quantitative comparison in Table 5-1. This demonstrates that the designed PSM and PTM enable the network to learn global continuity of patterns and seamless boundaries between tiling textures in the latent space.

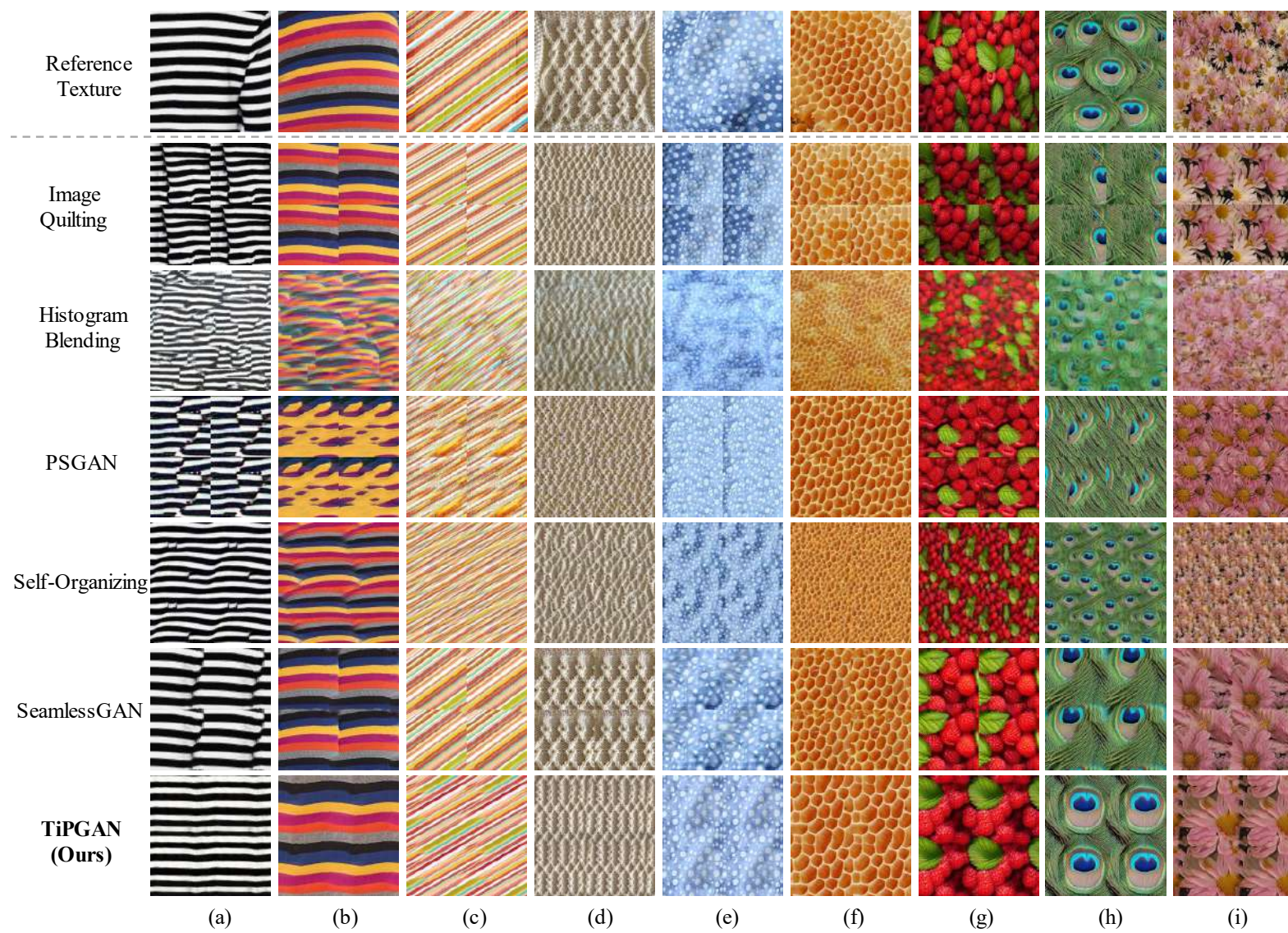


Figure 5-5 Qualitative comparison with Image Quilting (Efros & Freeman, 2023), Histogram Blending (Deliot & Heitz, 2019), PSGAN (Bergmann et al., 2017), Self-Organizing (Niklasson et al., 2021), and SeamlessGAN (Rodriguez-Pardo & Garces, 2022). To more intuitively visualize texture tileability, all the results are processed by 2×2 tiling. The proposed TiPGAN excels at preserving the continuity of the overall pattern of textures while providing seamless and semantically consistent boundaries.

5.4.4 Ablation Study

5.4.4.1 Ablation on the PSM & PTM Modules

To examine the effectiveness of TiPGAN in synthesizing seamless textures, some textures that are initially warped and non-flattened are carefully selected. Figure 5-6 depicts the qualitative results of the ablation study. The baseline column denotes the network structure in Figure 5-4 without the patch swapping module (PSM) and patch tiling module (PTM) removed. It is evident that generated textures of the baseline model are simply stitched patches together, with noticeable misalignments and displacements within the textures. The ‘+PSM’ column represents the results of the baseline model using only PSM. PSM cannot work by itself as it increases variation to the textures rather than provides guidance for tileability. The ‘+PTM’ column denotes the results of baseline model with only the PTM. As shown, the overall tileability is improved compared to the baseline. However, the pattern is still misaligned at the edges of the tiles, especially along the vertical direction (see cases of diagonal stripes, polka

dots and plaid patterns in Figure 5-6). The last column ‘+ PSM & PTM’ shows the results of the proposed TiPGAN. It is noteworthy that TiPGAN successfully generates relatively smooth, tileable textures, while preserving the perceptual quality in alignment with the reference texture. Table 5-2 presents the quantitative ablation results. It shows that PTM positively contributes to seamless quality, as evidenced by the reduction in the RTV score. However, it may affect the visual fidelity in order to make the pattern tileable, as shown by the slight increase in SI-FID score. The incorporation of both PSM and PTM can effectively enhance tileability and preserve visual quality, as reflected by the improvements in the SI-FID, LPIPS, and RTV metrics. There is an acceptable decrease in the SSIM score, only 0.005 lower than that achieved by PTM alone.

Table 5-2 Quantitative ablation result.

	SSIM↑	SI-FID↓	LPIPS↓	RTV↓
Baseline	0.2634	6.6274	0.3839	4.8549
+PSM	0.2596	6.5293	0.3926	3.8072
+PTM	0.2730	6.6287	0.4060	4.5393
+PSM&PTM	0.2680	1.9574	0.3743	0.7259

*We bold the **best** results and underscore the second-best results.*



Figure 5-6 Qualitative ablation results on patch swapping module and patch tiling module.
All outputs are tiled 2×2 to visualize texture tileability.

5.4.4.2 Ablation on the Choice of Patch Size p

The patch size p plays a crucial role in our RTV metric by determining the granularity of local texture analysis. Initially, $p = 16$ was selected empirically based on reasonable

preliminary performance. To better understand its impact, we conducted an ablation study with $p \in \{1, 2, 4, 8, 16, 32, 64\}$ across various transformation types and degrees, as shown in Tab. 3. For this study, we collected 5,599 royalty free seamless textures from (3DJungle, 2024) and applied various types and degrees of transformation to artificially degrade their tileability. These transformations were designed to simulate real-world challenges and assess the robustness of the RTV metric under different conditions.

Our results reveal that RTV is robust to the choice of p for $p \leq 16$, as it consistently increases with transformation degree, aligning with the intended design of the metric. However, for larger patch sizes ($p = 32$ and $p = 64$), this correlation weakens, indicating reduced sensitivity to fine-grained transformations. Smaller patch sizes, particularly $p = 4$, yield the smallest standard deviations, making them optimal for capturing local variations. Despite this, the choice of $p = 16$ does not affect the key correlation between transformation degree and RTV, ensuring the validity of our conclusions. For computational efficiency and consistency with the initial experiments, we retained $p = 16$ but recommend $p = 4$ for future work to further reduce variation and improve robustness.

Table 5-3 Quantitative ablation result.

Transformation Type	Transformation Degree	Patch Size p							
		1	2	4	8	16	32	64	
No Transformation		0.3063 ± 0.68	0.2655 ± 0.53	0.3826 ± 0.87	0.5104 ± 1.01	0.7502 ± 1.30	1.2546 ± 2.19	2.1084 ± 3.40	
	Deformation	Medium	1.2970 ± 0.93	0.7989 ± 0.68	0.6767 ± 0.62	0.6875 ± 0.70	0.8325 ± 1.03	1.2451 ± 1.68	2.0094 ± 2.63
	High	1.3280 ± 1.33	0.9310 ± 1.00	0.7752 ± 0.88	0.7789 ± 0.85	0.9079 ± 1.02	1.2437 ± 1.58	1.9123 ± 2.28	
Scale	Medium	1.3826 ± 2.45	0.9325 ± 1.73	0.8578 ± 1.73	1.0289 ± 1.80	1.1910 ± 1.99	1.7152 ± 2.45	2.8874 ± 3.65	
	High	2.7897 ± 5.29	1.7182 ± 3.66	1.3225 ± 3.06	1.3929 ± 3.12	1.8193 ± 3.76	2.2408 ± 3.70	5.0124 ± 6.14	
Rotation	Medium	1.4498 ± 1.84	0.8917 ± 1.25	0.7724 ± 1.09	0.8111 ± 1.14	0.9644 ± 1.39	1.3716 ± 1.86	2.4087 ± 3.02	
	High	1.6813 ± 2.64	1.0337 ± 1.91	0.8515 ± 1.68	0.8555 ± 1.59	0.9755 ± 1.66	1.3887 ± 2.02	2.3380 ± 2.92	

5.5 Discussion

5.5.1 *Effectiveness of the proposed RTV metric*

To evaluate the tileability of texture, this section proposed the relative total variation (RTV) metric based on the observation that seamless textures maintain continuity when tiled. If the texture is tileable, its continuity performance will not be affected after splicing itself, and the value of total variation will not change significantly. The value of RTV reflects the relative changes in global smoothness and continuity before and after texture splicing. The smaller the RTV value, the better the texture's tileability, and vice versa. To prove the effectiveness and robustness of the RTV metric, a wide array of seamless textures is specially selected and manipulated into non-seamless textures. The manipulations adopted to break seamless textures include deformation, scale and rotation. These disruptions are applied to seamless examples. In Figure 5-7, from top to bottom, the level of their tileability is gradually decreasing, as evident by the increasing RTV value below texture samples, quantifying the texture's tileability. Higher the level of disruptions, the lower the tileability property of the resulting textures, as reflected in higher RTV scores. For scaling disruptions, RTV is more sensitive to pattern amplification. When the texture is scaled up, its capacity to capture regular and complete patterns decreases. Consequently, this leads to a more pronounced visual disruption in the texture's coherence at the junctions. In rotation disruption, the textures

in the middle column are most sensitive to rotation disturbance, and their RTV changes are also more obvious, because texture tileability takes into account the smoothness and the continuity of both *petal-like* patterns and *line-like* patterns in this case.

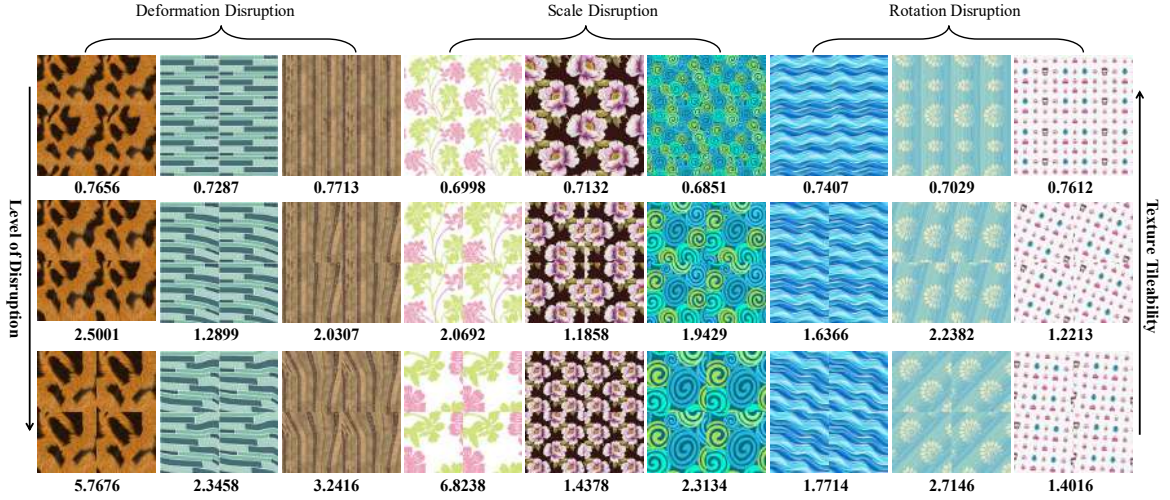


Figure 5-7 Effectiveness of the proposed RTV metric in evaluating seamless quality.

5.5.2 Capability in Cloth Digitization

TiPGAN generates outputs that can seamlessly tile, effectively maintaining the reference texture's pattern, color, style, geometry. The output seamless textures can be easily applied to 3D meshes using the cloth digitization pipeline. With reference to Figure 5-3, the seamless textures generated by TiPGAN are divided into two main types: one is pattern A with rich style features, and the other is fine-grained material B (e.g., knitted texture). For input B, the normal map generator (NMG) (Shi et al., 2022) is used to produce the corresponding high-quality and tileable normal map. The tileability feature of TiPGAN output ensures that both pattern A and normal map B can be seamlessly merged, enabling the creation of novel textures that boast rich styles and

true-to-life materials, all of which can be realistically rendered using PBR techniques. In the implementation, the subsequent Cloth Digitization & Rendering (CDR) process shown in Figure 5-3 was completed using CLO3D (CLO3D, 2024), a commercial software, and Blender (Blender, 2024), an open-source 3D computer graphics software. Figure 5-1 show the astonishing results of various seamless textile products being generated and presented on 3D models. The tileable textures generated by the TiPGAN can extend beyond cloth digitization and be easily applied to different 3D objects. See Figure 5-9 for other applications.

To showcase the capabilities of the proposed cloth digitization pipeline, comparative experiments are conducted with Pix2Surf (Mir et al., 2020). The idea of (Mir et al., 2020)'s work was to use a non-rigid 3D to 2D registration method to generate data pairs of clothing images and 3D garment, and then based on these data pairs, a neural network was employed to learn how to map the pixels of the image onto the 3D garment's surface. Their garment images are used to generate a seamless texture with TiPGAN, which is then directly applied to the 3D garment's surface. We straightforwardly map the synthesized textures onto UV maps by tiling the small texture repeatedly in all directions, ensuring continuity within each panel. While this approach naturally results in seams at the boundaries where panels are stitched together, these discontinuities are unavoidable and align with standard clothing manufacturing processes. Our motivation

focuses on achieving seamless tiling within individual panels, as the boundary seams are not our concern and are acceptable in practical garment production.



Figure 5-8 Comparison of 3D cloth digitization performance. The proposed Model iv' results outperform others with high-resolution textures and flexible pattern tiling sizes. Zoom in for a better view.

Figure 5-8 shows the comparison visualization results. In (a_1, a_2, a_3) of Figure 5-8, it is evident that Pix2Surf model (Mir et al., 2020) is not able to maintain the details of high-quality textures, resulting in a blurred appearance of the generated textures on the T-shirt and shorts. The cloth digitization pipeline (*Model iv*) can successfully capture the distinct features of the reference texture, including style, color, and geometric structure.

It allows easy modification of the texture’s tiling and orientation, as demonstrated in (b_1, b_2, b_3) , changing angle for tiling adjustments, as shown in (c_1, c_2, c_3) of Figure 5-8. Additionally, in the 3D garment renderings, we have applied normal maps with patterns such as polka dots and plaids to simulate the shading effects of various fabrics or materials, as shown in (d_1, d_2, d_3) . For a closer view of their detailed effects, please refer to (e_1, e_2, e_3) of Figure 5-8. A video is also provided in the supplementary material, showcasing a 360° rotation of resulting 3D garments, to compare the two methods comprehensively. The results shown in Figure 5-8 and the accompanied video have the same camera perspective and angle, lighting conditions, image size, and resolution during the 3D garment rendering process

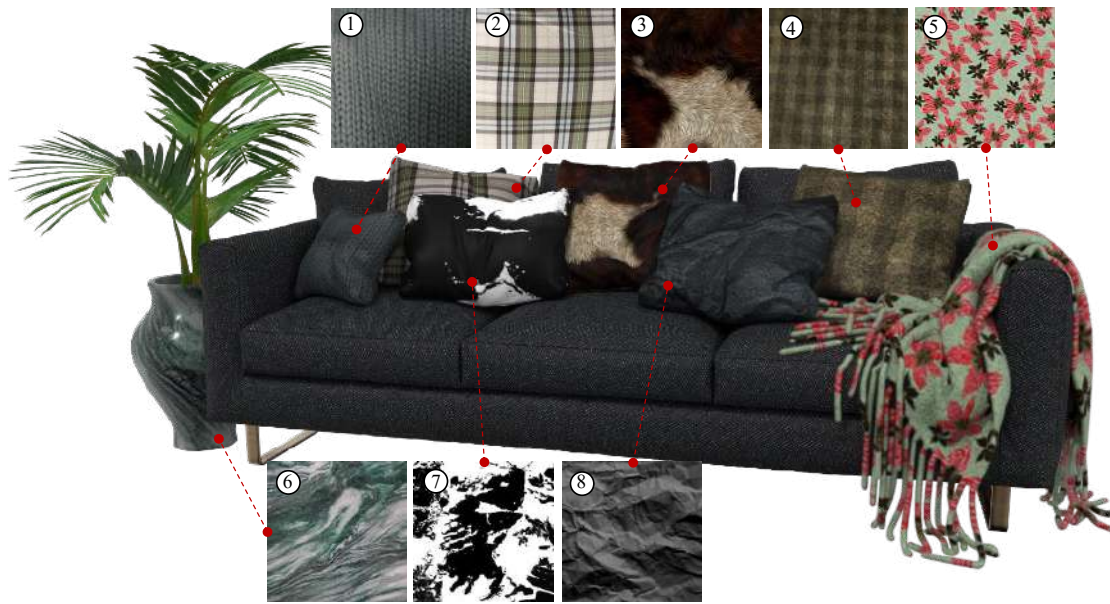


Figure 5-9 Other applications of the proposed tileable texture: the tileable textures are generated by TiPGAN, utilizing the non-flatten and non-seamless reference textures ① to ⑧.

5.5.3 Limitation of TiPGAN

TiPGAN is capable of ensuring tileability across a wide range of textures. However, if the reference texture is not ideally flattened, the synthesized textures will contain certain unwanted warping patterns. For instance, Figure 5-11 (a) shows that the synthesized bricks are laid in an angled pattern due to the perspective transformation of the reference texture. For non-periodic patterns or non-stationary textures, TiPGAN may fail because it is difficult to obtain regular tiling patterns (e.g., Figure 5-10 (b) keeps the gathered wrinkles). Moreover, the inherent design of the TiPGAN model has prevented it from capturing multiple elements within a single image. For example, Figure 5-10 (c) blends the textures from the trousers and the top. Additionally, during the random sampling process, the texture sampling frequency for the trousers is lower than that for the top, leading to more artifacts in the respective region of the generated texture.

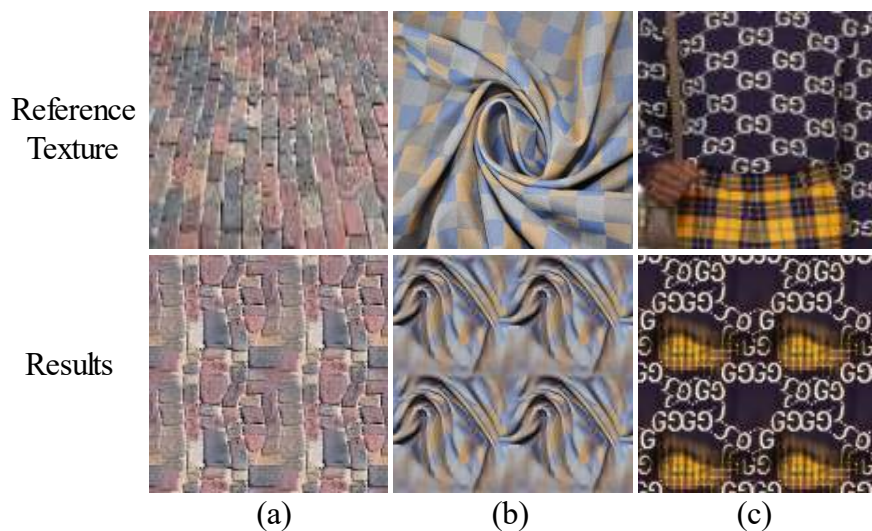


Figure 5-10 Failure examples synthesized by TiPGAN.

5.6 Augmented Reality Mobile Application

The proposed 3D visual try-on method has been incorporated and utilized in a mobile app designed for trying on fashion using Augmented Reality (AR) technology. The AR application called 'easy wear' helps users measure their own body measurements, try on their favorite virtual fashion products, and generate their own customized avatars. If the user launches the try-on room from the easy wear store application, all the user's information can be transferred to the try-on room application. 'Easy Wear' has two integrated parts, the fitting room APP and the store APP. The former is used to collect information about the user's body measurements and help users to perform body-aware virtual try-on. The latter store APP presents the textured 3D garments database, where users can select their favourite 3D garments to try on. Furthermore, textured 3D garments can be combined with Avatars bound to skeletons to realize virtual catwalks. Figure 5-11 shows the easy wear app's all features page.

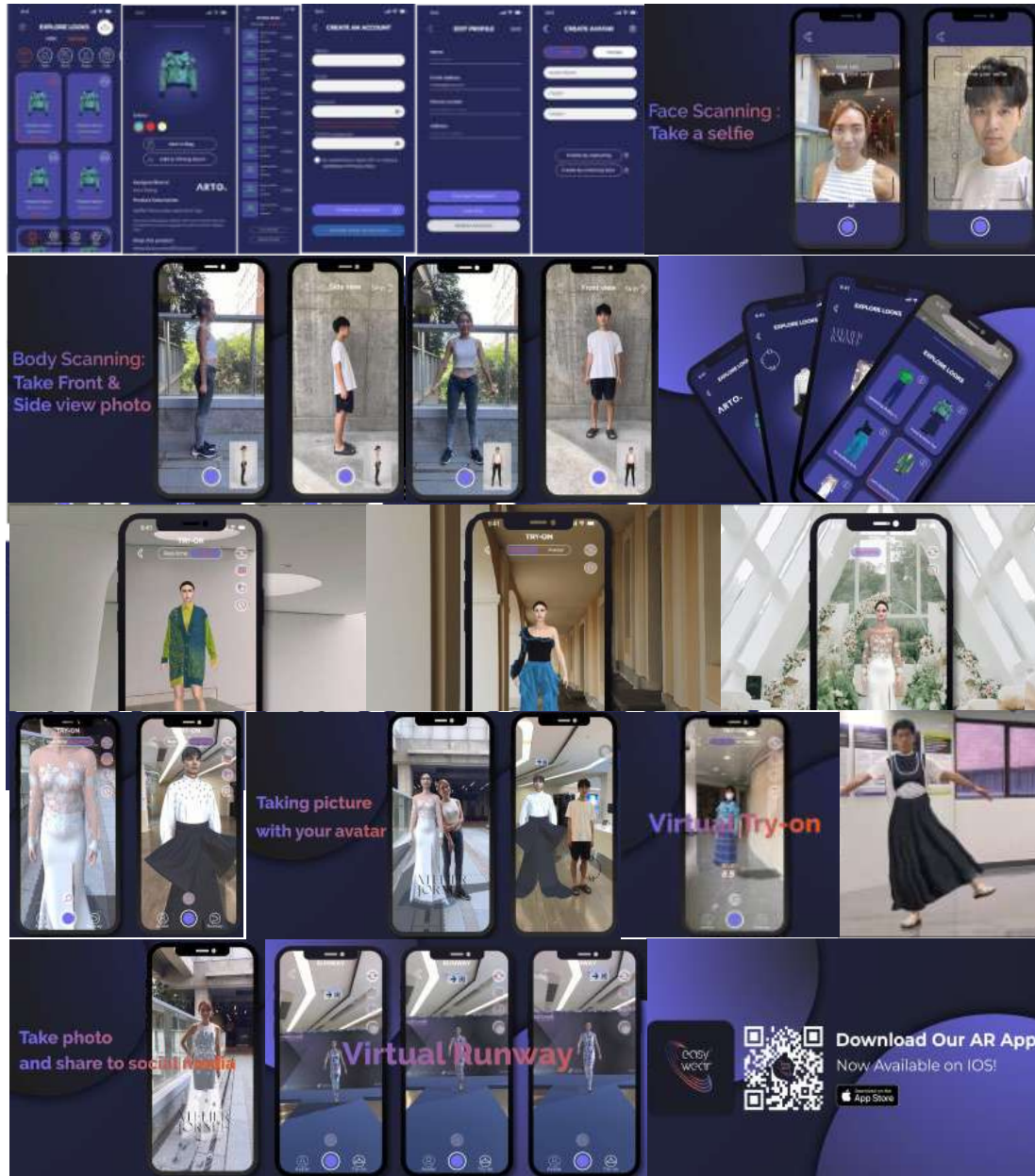


Figure 5-11 Screenshot of demo video for of Easy Wear APP.

5.7 Chapter Summary

In this chapter, we report the proposed tileable texture generation model (*Model iii*), a novel texture synthesis framework (named TiPGAN). This framework combines recent advances in deep texture synthesis, adversarial neural networks, and latent space

manipulation. TiPGAN is able to synthesize high-quality tileable textures while maintaining perceptual consistency with reference textures. Compared with previous SOTA methods, TiPGAN has been verified by extensive qualitative and quantitative experiments that it is capable of generating textures with finer pattern details and with semantic properties better preserved. Moreover, a new metric, called Relative Total Variation (RTV), has been introduced, which can be used for the evaluation of texture synthesis methods, measuring quantitatively the seamless quality of the synthesized textures. RTV computes a human-friendly score that effectively quantifies the intrinsic tileability of any texture. Extensive experiments have also been conducted to demonstrate that RTV can evaluate various methods for tileable texture synthesis, outperforming previous existing metrics that lead to irrelevant evaluations (i.e., a good tileable texture may have a high SI-FID (Shaham et al., 2019) score, or a poor tileable texture may have a low LPIPS (Zhang et al., 2018)). Finally, the potential applications of TiPGAN are discussed, and a cloth digitization pipeline for the fashion industry is proposed. TiPGAN is capable of accurately capturing various aspects of the reference texture, such as style, color, and geometric structure, synthesizing high-quality tileable textures. Consequently, the cloth digitization pipeline (*Model iv*) offers several advantages over certain end-to-end methods. For example, it allows easy adjustments of the number and direction of the repeats, as well as the angle of tiles for mapping the texture onto a 3D object’s surface.

In addition, this chapter also showcase an AR mobile APP developed using the proposed technology, allowing users to visual realistic try-on effects in AR scenes. The use of AR technology is reshaping the marketing landscape of fashion brands, injecting innovation and appeal into their strategies, thereby attracting consumer interest and increasing product conversion rates. The proposed tileable texture generation model (*Model iii*) highlights the great utility of high-quality tileable textures in mapping textures to 3D garment surfaces. The proposed cloth digitization pipeline (*Model iv*) can easily and flexibly map the high-quality tileable textures generated by *Model iii* to 3D garment, generating not only textures but also tileable albedo and normal maps at the same time, achieving more realistic 3D visual try-on through 3D garment rendering. These 3D visual try-on effects can be integrated into AR scenes to provide customers with an immersive online shopping experience.

Chapter 6. Conclusions and Recommendations for Future Work

6.1 Conclusions

This thesis proposes novel 2D virtual try-on and 3D visual try-on solutions. It reports three developments including fashion items localization, human body shape-garment relationship and high-quality tileable texture reconstruction.

The first study reports a fashion landmark localization method based on pose-aware segmentation, which can automatically locate functional key points on a garment even without Ground Truth. Unlike deep learning-based models, the proposed fashion landmark localization model does not rely on massive amounts of data but rather locates the key points of a garment through the relationship between the garment and the human body. This modeling method of defining explicit patterns enables controlled body-aware garment morphing. The proposed fashion landmark localization method is evaluated through experimental comparison with other state-of-the-art methods. It can greatly improve the efficiency of locating key points while ensuring accuracy.

The second development is the BSLT-VTON method for fashion landmark transformation based on human size. It encompasses a two-step development strategy covering (i) a fashion logo localization model (reported in Chapter 3). and (ii) a body-aware garment morphing model. *Model ii* aims to guide the source garment morphing

onto the target body while responding to the target body's body shape, height/posture. The BSLT-VTON method proposed in Chapter 4 models the wearable looseness and wearing position between the target person and the source garment in a 2D visual form to achieve a realistic and practical fitting effect. Extensive experiments have shown that BSLT-VTON is satisfactory in terms of garment deformation and outperforms some state-of-the-art VTON methods in terms of garment texture quality preservation.

The third study reported in Chapter 5 covers the proposal of a high-quality texture reconstruction (TiPGAN) method and a new cloth digitization pipeline for 3D VTON. The proposed TiPGAN model has been verified through a large number of qualitative and quantitative experiments and can generate textures with finer pattern details and better preservation of semantic properties. In addition, a new metric, relative total variation (RTV), is reported to evaluate TiPGAN model, quantitatively measuring the seamless quality of the synthesized texture. It fills the research gap for objective quantitative evaluation regarding texture tileability. The proposed cloth digitization pipeline can easily and flexibly map the high-quality tileable texture generated by TiPGAN onto 3D clothing, generating not only textures but also tileable albedo and normal maps at the same time, achieving more realistic 3D visual try-on through 3D clothing rendering. Finally, all the results of visualizing 3D VTON have been applied to the AR mobile application process flow. In addition, the 3D VTON with skeleton avatar completed a virtual runway. These all demonstrate the applicability of the

developed method in the fashion domain.

6.2 Recommendations for Future Work

In this study, a total of two solutions covering four models/developments have been reported. Each of these models/methods have been extensively evaluated and verified through experiment, from which a number of promising avenues for future research and development have emerged:

1. Chapter 3 reports a method for locating specific points on the human body that are critical to the fit and shape of a garment, such as shoulder points, waistlines, and necklines. The proposed method used neural networks to parse garment regions and extract texture and pose estimation to detect body joint points. This is done explicitly by combining garment and human body relationships. This approach deserves to be explored further, combining deep learning methods to address the accuracy of fashion landmark localization in the presence of complex pose and self-obscuration to enhance the robustness of the current algorithms.
2. Chapter 4 describes a new 2D VTON method. This method uses a fashion landmark localization model to locate key functional points and then deforms and scales the garments for different body shapes through a body-aware garment deformation model. Looking ahead, a possible future direction is to pursue real-time performance optimization. By integrating the fashion landmark localization and body-aware clothing deformation model in Chapter 3 into the 2D image

synthesis task. An end-to-end 2D virtual try-on task may not only ensure effective try-on results but also help improve efficiency.

3. Chapter 5 reports the development of TiPGAN to generate high-quality tileable texture based on latent space manipulation. As future work for this study, more texture manipulations in latent space, such as de-artifacting (Dekel et al., 2015), de-distortion (Li et al., 2019b), and automatic texture rotation and scaling (Chatillon et al., 2023), are planned to be explored, in order to provide a generalized deep synthesis for different texture synthesis tasks. In addition, the combination of a seamless intrinsic prior with a latent diffusion model (Rombach et al., 2022) for seamless texture synthesis with text control will be explored.
4. The proposed 2D VTON and 3D VTON solutions are mainly developed to serve fashion applications. In fact, virtual try-on technology can have many applications beyond the fashion industry. For example, the healthcare sector utilizes virtual try-on technology to enhance patient care, improve the fit of medical devices (orthopedic braces and prostheses) and facilitate teleconsultations (virtual vein viewers). The requirements of these different application domains are different from those of the fashion domain. It would be interesting to explore the specific requirements of the different domains and develop more generalized solutions in the future.
5. In future research, the value of virtual try-on system in practical commercial applications should be evaluated comprehensively. While evaluation metrics like

SSIM and FID evaluate the technical proficiency of the virtual try-on system's image generation in the research, they do not entirely capture its practical commercial significance. The commercial success of a virtual try-on system hinges not only on image quality but also on user experience, market performance, and business benefits. Thus, a holistic evaluation necessitates a fusion of technical benchmarks and commercial metrics. Technical indicators scrutinize image generation accuracy, realism, and consistency, while business metrics like user satisfaction, conversion rates, return rates, user retention, and average order value gauge the systems influence on user behavior and market performance. By employing A/B testing, user studies, data analytics, and ROI analysis, a scientific foundation is laid for refining technology and adjusting business strategies. Future evaluation criteria will integrate technical and commercial indicators to comprehensively assess the technical prowess and commercial viability of virtual trial systems, aiming to optimize market potential and enhance user experiences.

In conclusion, virtual try-on technology has great potential to revolutionize various industries. By delving into these future directions, researchers can contribute to the continuous improvement of technologies that provide enhanced realism, efficiency and adaptability in a wide range of applications.

References

- 3DJungle. (2024). *3djungle.net*. <https://3djungle.net/>
- Abels, A., & Kruusmaa, M. (2013). Construction of a female shape-changing robotic mannequin. *Journal of Automation and Control Engineering*, 1(2), 132-134.
- Afifi, M., Brubaker, M. A., & Brown, M. S. (2021). Histogan: Controlling colors of gan-generated and real images via color histograms. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,
- Akl, A., Yaacoub, C., Donias, M., Da Costa, J.-P., & Germain, C. (2018). A survey of exemplar-based texture synthesis methods. *Computer Vision and Image Understanding*, 172, 12-24.
- Alanov, A., Kochurov, M., Volkhonskiy, D., Yashkov, D., Burnaev, E., & Vetrov, D. (2019). User-controllable multi-texture synthesis with generative adversarial networks. *arXiv preprint arXiv:1904.04751*.
- Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein Generative Adversarial Networks. International Conference on Machine Learning,
- Azuma, R. T. (1997). A survey of augmented reality. *Presence: teleoperators & virtual environments*, 6(4), 355-385.
- Barnes, C., Goldman, D. B., Shechtman, E., & Finkelstein, A. (2011). The patchmatch randomized matching algorithm for image manipulation. *Communications of the ACM*, 54(11), 103-110.
- Barnes, C., Shechtman, E., Finkelstein, A., & Goldman, D. B. (2009). PatchMatch: A

- randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3), 24.
- Barnes, C., Shechtman, E., Goldman, D. B., & Finkelstein, A. (2010). The generalized patchmatch correspondence algorithm. *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III* 11,
- Bartle, A., Sheffer, A., Kim, V. G., Kaufman, D. M., Vining, N., & Berthouzoz, F. (2016). Physics-driven pattern adjustment for direct 3D garment editing. *ACM Trans. Graph.*, 35(4), 50:51-50:11.
- Batool, R., & Mou, J. (2023). A systematic literature review and analysis of try-on technology: Virtual fitting rooms. *Data and Information Management*, 100060.
- Belagiannis, V., Rupprecht, C., Carneiro, G., & Navab, N. (2015). Robust optimization for deep regression. *Proceedings of the IEEE international conference on computer vision*,
- Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE transactions on pattern analysis and machine intelligence*, 24(4), 509-522.
- Bergmann, U., Jetchev, N., & Vollgraf, R. (2017). Learning texture manifolds with the periodic spatial GAN. *arXiv preprint arXiv:1705.06566*.
- Berman, B. (2019). Flatlined: Combatting the death of retail stores. *Business Horizons*, 62(1), 75-82.

- Bhatnagar, B. L., Tiwari, G., Theobalt, C., & Pons-Moll, G. (2019). Multi-garment net: Learning to dress 3d people from images. *Proceedings of the IEEE/CVF international conference on computer vision*,
- Bialkova, S., & Barr, C. (2022). Virtual try-on: How to enhance consumer experience? 2022 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW),
- Billinghurst, M., Clark, A., & Lee, G. (2015). A survey of augmented reality. *Foundations and Trends® in Human–Computer Interaction*, 8(2-3), 73-272.
- Bińkowski, M., Sutherland, D. J., Arbel, M., & Gretton, A. (2018). Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.
- Birsak, M., Musialski, P., Arikan, M., & Wimmer, M. (2013). Seamless texturing of archaeological data. 2013 Digital Heritage International Congress (DigitalHeritage),
- Blanco, F. R., & Oliveira, M. M. (2008). Instant mesh deformation. *Proceedings of the 2008 symposium on Interactive 3D graphics and games*,
- Blender. (2024). <https://www.blender.org/>
- Bookstein, F. L. (1989). Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE transactions on pattern analysis and machine intelligence*, 11(6), 567-585.
- Bossard, L., Dantone, M., Leistner, C., Wengert, C., Quack, T., & Van Gool, L. (2013). Apparel classification with style. *Computer Vision–ACCV 2012: 11th Asian*

- Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012,
Revised Selected Papers, Part IV 11,
- Browzwear. (2024). <https://browzwear.com/>
- Busto, P. P., Eisenacher, C., Lefebvre, S., & Stamminger, M. (2010). Instant Texture
Synthesis by Numbers. VMV,
- Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2d pose
estimation using part affinity fields. Proceedings of the IEEE conference on
computer vision and pattern recognition,
- Carmigniani, J., Furht, B., Anisetti, M., Ceravolo, P., Damiani, E., & Ivkovic, M. (2011).
Augmented reality technologies, systems and applications. *Multimedia tools
and applications*, 51, 341-377.
- Casciani, D., Chkanikova, O., & Pal, R. (2022). Exploring the nature of digital
transformation in the fashion industry: opportunities for supply chains, business
models, and sustainability-oriented innovations. *Sustainability: Science,
Practice and Policy*, 18(1), 773-795.
- Catmull, E. E. (1974). *A subdivision algorithm for computer display of curved surfaces*.
The University of Utah.
- Chatillon, P., Gousseau, Y., & Lefebvre, S. (2023). A geometrically aware auto-encoder
for multi-texture synthesis. International Conference on Scale Space and
Variational Methods in Computer Vision,
- Chaudhuri, B., Sarafianos, N., Shapiro, L., & Tung, T. (2021). Semi-supervised

- synthesis of high-resolution editable textures for 3d humans. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,
- Chen, C.-Y., Chen, Y.-C., Shuai, H.-H., & Cheng, W.-H. (2023a). Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network. Proceedings of the IEEE/CVF International Conference on Computer Vision,
- Chen, D. Z., Siddiqui, Y., Lee, H.-Y., Tulyakov, S., & Nießner, M. (2023b). Text2tex: Text-driven texture synthesis via diffusion models. Proceedings of the IEEE/CVF International Conference on Computer Vision,
- Chen, H., Liu, J., Chen, W., Liu, S., & Zhao, Y. (2022). Exemplar-based pattern synthesis with implicit periodic field network. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,
- Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834-848.
- Chen, L.-C., Yang, Y., Wang, J., Xu, W., & Yuille, A. L. (2016). Attention to scale: Scale-aware semantic image segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition,
- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV),

- Chen, M., Qin, Y., Qi, L., & Sun, Y. (2019a). Improving fashion landmark detection by dual attention feature enhancement. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*,
- Chen, X., & Yuille, A. L. (2014). Articulated pose estimation by a graphical model with image dependent pairwise relations. *Advances in neural information processing systems*, 27.
- Chen, Z., Fu, Y., Wang, Y.-X., Ma, L., Liu, W., & Hebert, M. (2019b). Image deformation meta-networks for one-shot learning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,
- Chen, Z., & Zhang, H. (2019). Learning implicit fields for generative shape modeling. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,
- Chidambaram, V., Rana, N. P., & Parayitam, S. (2024). Antecedents of consumers online apparel purchase intention through Virtual Try On technology: A moderated moderated-mediation model. *Journal of Consumer Behaviour*, 23(1), 107-125.
- Cho, H., & Schwarz, N. (2010). I like those glasses on you, but not in the mirror: Fluency, preference, and virtual mirrors. *Journal of Consumer Psychology*, 20(4), 471-475.
- Choi, S., Park, S., Lee, M., & Choo, J. (2021). Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. *Proceedings of the IEEE/CVF*

- Conference on Computer Vision and Pattern Recognition,
- Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., & Vedaldi, A. (2014). Describing textures in the wild. *Proceedings of the IEEE conference on computer vision and pattern recognition*,
- CLO3D. (2024). <https://clo3d.com>
- Cloud, J. M., & Perilloux, C. (2015). “Drawing” conclusions about perceptions of ideal male and female body shapes. *Evolutionary Psychological Science*, 1, 163-171.
- Cohen, M. F., Shade, J., Hiller, S., & Deussen, O. (2003). Wang tiles for image and texture generation. *ACM Transactions on Graphics (ToG)*, 22(3), 287-294.
- Cordier, F., Seo, H., & Magnenat-Thalmann, N. (2003). Made-to-measure technologies for an online clothing store. *IEEE computer graphics and applications*, 23(1), 38-48.
- Craig, A. B. (2013). Understanding augmented reality: Concepts and applications.
- Dekel, T., Michaeli, T., Irani, M., & Freeman, W. T. (2015). Revealing and modifying non-local variations in a single image. *ACM Transactions on Graphics (ToG)*, 34(6), 1-11.
- Deliot, T., & Heitz, E. (2019). Procedural stochastic textures by tiling and blending. *GPU Zen*, 2(2), 4-6.
- Deng, W., Ke, W., Deng, Z., & Wang, X. (2024). Virtual design of woven fabrics based on parametric modeling and physically based rendering. *Computer-Aided Design*, 173, 103717.

- Detlefsen, N. S., Borovec, J., Schock, J., Jha, A. H., Koker, T., Di Liello, L., Stancl, D., Quan, C., Grechkin, M., & Falcon, W. (2022). Torchmetrics-measuring reproducibility in pytorch. *Journal of Open Source Software*, 7(70), 4101.
- Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34, 8780-8794.
- Dong, H., Liang, X., Shen, X., Wang, B., Lai, H., Zhu, J., Hu, Z., & Yin, J. (2019). Towards multi-pose guided virtual try-on network. Proceedings of the IEEE/CVF International Conference on Computer Vision,
- Dong, W., Zhou, N., & Paul, J.-C. (2007). Optimized tile-based texture synthesis. Proceedings of Graphics Interface 2007,
- Dou, S., Jiang, X., Tu, Y., Gao, J., Qu, Z., Zhao, Q., & Zhao, C. (2024). DROP: Decouple Re-Identification and Human Parsing with Task-specific Features for Occluded Person Re-identification. *arXiv preprint arXiv:2401.18032*.
- Douglas, D. H., & Peucker, T. K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization*, 10(2), 112-122.
- Dubey, S. R. (2021). A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5), 2687-2704.
- Efros, A. A., & Freeman, W. T. (2023). Image quilting for texture synthesis and transfer.

- In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2* (pp. 571-576).
- Efros, A. A., & Leung, T. K. (1999). Texture synthesis by non-parametric sampling. Proceedings of the seventh IEEE international conference on computer vision,
- Emarketer. (2024). *Worldwide ecommerce sales*.
<https://www.emarketer.com/insights/worldwide-ecommerce-sales-to-top-7-trillion/>
- Erra, U., & Colonnese, V. (2015). Experiences in the development of an augmented reality dressing room. International Conference on Augmented and Virtual Reality,
- Fan, X., Zheng, K., Lin, Y., & Wang, S. (2015). Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. Proceedings of the IEEE conference on computer vision and pattern recognition,
- Fang, N., Qiu, L., Zhang, S., Wang, Z., Hu, K., & Li, H. (2022). Toward multi-category garments virtual try-on method by coarse to fine TPS deformation. *Neural Computing and Applications*, 34(15), 12947-12965.
- Faragallah, O. S., El-Hoseny, H., El-Shafai, W., Abd El-Rahman, W., El-Sayed, H. S., El-Rabaie, E.-S. M., Abd El-Samie, F. E., & Geweid, G. G. N. (2020). A comprehensive survey analysis for present solutions of medical image fusion and future directions. *IEEE Access*, 9, 11358-11371.
- Fincato, M., Cornia, M., Landi, F., Cesari, F., & Cucchiara, R. (2022). Transform, warp, and dress: a new transformation-guided model for virtual try-on. *ACM*

Transactions on Multimedia Computing, Communications, and Applications (TOMM), 18(2), 1-24.

Fincato, M., Landi, F., Cornia, M., Cesari, F., & Cucchiara, R. (2021). VITON-GT: an image-based virtual try-on model with geometric transformations. 2020 25th International Conference on Pattern Recognition (ICPR),

Frühstück, A., Alhashim, I., & Wonka, P. (2019). Tilegan: synthesis of large-scale non-homogeneous textures. *ACM Transactions on Graphics (ToG)*, 38(4), 1-11.

Fu, J., Wang, J., Li, Z., Xu, M., & Lu, H. (2013). Efficient Clothing Retrieval with Semantic-Preserving Visual Phrases. In (pp. 420-431). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37444-9_33

Galerne, B., Gousseau, Y., & Morel, J.-M. (2010). Random phase textures: Theory and synthesis. *IEEE transactions on image processing*, 20(1), 257-267.

Gao, D., Chen, X., Zhang, X., Wang, Q., Sun, K., Zhang, B., Bo, L., & Huang, Q. (2024). Cloth2Tex: A Customized Cloth Texture Generation Pipeline for 3D Virtual Try-On. 2024 International Conference on 3D Vision (3DV),

Gatys, L., Ecker, A. S., & Bethge, M. (2015). Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28.

Ge, C., Song, Y., Ge, Y., Yang, H., Liu, W., & Luo, P. (2021a). Disentangled cycle consistency for highly-realistic virtual try-on. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., & Luo, P. (2021b). Parser-free virtual try-

- on via distilling appearance flows. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,
- Ge, Y., Zhang, R., Wang, X., Tang, X., & Luo, P. (2019). Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,
- Gong, K., Gao, Y., Liang, X., Shen, X., Wang, M., & Lin, L. (2019). Graphonomy: Universal human parsing via graph transfer learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,
- Gong, K., Liang, X., Zhang, D., Shen, X., & Lin, L. (2017). Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. Proceedings of the IEEE conference on computer vision and pattern recognition,
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Gou, J., Sun, S., Zhang, J., Si, J., Qian, C., & Zhang, L. (2023). Taming the power of diffusion models for high-quality virtual try-on with appearance flow. Proceedings of the 31st ACM International Conference on Multimedia,
- Guerrero-Viu, J., Hasan, M., Roullier, A., Harikumar, M., Hu, Y., Guerrero, P., Gutierrez, D., Masia, B., & Deschaintre, V. (2024). Texsliders: Diffusion-based texture editing in clip space. ACM SIGGRAPH 2024 Conference Papers,

- Guler, R. A., & Kokkinos, I. (2019). Holopose: Holistic 3d human reconstruction in-the-wild. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). Improved Training of Wasserstein GANs. *arXiv preprint arXiv:1704.00028*.
- Guo, Z., Zhang, D., Zhang, S., Liu, X., & Li, J. (2016). The design of a form-changing female fitting robot. *Journal of Advanced Mechanical Design, Systems, and Manufacturing*, 10(8), JAMDSM0097-JAMDSM0097.
- Han, X., Hu, X., Huang, W., & Scott, M. R. (2019a). Clothflow: A flow-based model for clothed person generation. Proceedings of the IEEE/CVF international conference on computer vision,
- Han, X., Wu, Z., Huang, W., Scott, M. R., & Davis, L. S. (2019b). Compatible and diverse fashion image inpainting. *arXiv preprint arXiv:1902.01096*.
- Han, X., Wu, Z., Wu, Z., Yu, R., & Davis, L. S. (2018). Viton: An image-based virtual try-on network. Proceedings of the IEEE conference on computer vision and pattern recognition,
- He, H., Zhang, J., Zhuang, B., Cai, J., & Tao, D. (2023). End-to-end one-shot human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity Mappings in Deep Residual Networks. In (pp. 630-645). Springer International Publishing.
https://doi.org/10.1007/978-3-319-46493-0_38

- He, S., Song, Y.-Z., & Xiang, T. (2022). Style-based global appearance flow for virtual try-on. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,
- Henzler, P., Mitra, N. J., & Ritschel, T. (2020). Learning a neural 3d texture space from 2d exemplars. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- Hilken, T., Chylinski, M., Keeling, D. I., Heller, J., de Ruyter, K., & Mahr, D. (2022). How to strategically choose or combine augmented and virtual reality for improved online experiential retailing. *Psychology & Marketing*, 39(3), 495-507.
- Hinz, T., Fisher, M., Wang, O., & Wermter, S. (2021). Improved techniques for training single-image gans. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*,
- Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 6840-6851.
- Hollensen, S., Kotler, P., & Opresnik, M. O. (2022). Metaverse—the new marketing universe. *Journal of Business Strategy*, 44(3), 119-125.
- Holten, D., Wijk, J. J. V., & Martens, J.-B. (2006). A perceptually based spectral model

- for isotropic textures. *ACM Transactions on Applied Perception (TAP)*, 3(4), 376-398.
- Hsieh, C.-W., Chen, C.-Y., Chou, C.-L., Shuai, H.-H., & Cheng, W.-H. (2019). Fit-me: Image-based virtual try-on with arbitrary poses. 2019 IEEE international conference on image processing (ICIP),
- Hu, P., Li, D., Wu, G., Komura, T., Zhang, D., & Zhong, Y. (2018a). Personalized 3D mannequin reconstruction based on 3D scanning. *International Journal of Clothing Science and Technology*.
- Hu, Y., Xiao, L., Wang, Y., & Jin, Y. (2018b). Fashion pose machine for fashion landmark detection. 2018 International Conference on Image and Video Processing, and Artificial Intelligence,
- Hua, W., Liu, Y., Zhang, Z., Li, M., & Yu, X. (2024). Exploring users' adoption intention of virtual try-on apps: how users' individual characteristics affect post-use feelings. *Asia Pacific Journal of Marketing and Logistics*.
- Huang, J., Feris, R. S., Chen, Q., & Yan, S. (2015). Cross-domain image retrieval with a dual attribute-aware ranking network. Proceedings of the IEEE international conference on computer vision,
- Hwangbo, H., Kim, E. H., Lee, S.-H., & Jang, Y. J. (2020). Effects of 3D virtual “try-on” on online sales and customers' purchasing experiences. *IEEE Access*, 8, 189479-189489.
- Igarashi, T., & Igarashi, Y. (2009). Implementing as-rigid-as-possible shape

- manipulation and surface flattening. *journal of graphics, gpu, and game tools*, 14(1), 17-30.
- Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*,
- Javornik, A. (2016). *“Reality is in the air”: concept of perceived augmentation and exploration of its impact on consumer experience* Università della Svizzera italiana].
- Jetchev, N., Bergmann, U., & Seward, C. (2017). GANosaic: Mosaic creation with generative texture manifolds. *arXiv preprint arXiv:1712.00269*.
- Jetchev, N., Bergmann, U., & Vollgraf, R. (2016). Texture synthesis with spatial generative adversarial networks. *arXiv preprint arXiv:1611.08207*.
- Kalantidis, Y., Kennedy, L., & Li, L.-J. (2013). Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*,
- Kanazawa, A., Black, M. J., Jacobs, D. W., & Malik, J. (2018). End-to-end recovery of human shape and pose. *Proceedings of the IEEE conference on computer vision and pattern recognition*,
- Kang, J.-Y. M. (2014). Augmented reality and motion capture apparel e-shopping values and usage intention. *International Journal of Clothing Science and*

Technology.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation. International Conference on Learning Representations,

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34, 852-863.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Kenney, C. S., Zuliani, M., & Manjunath, B. (2005). An axiomatic approach to corner detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05),

Kho, Y., & Garland, M. (2005). Sketching mesh deformations. Proceedings of the 2005 symposium on Interactive 3D graphics and games,

Kiapour, M. H., Yamaguchi, K., Berg, A. C., & Berg, T. L. (2014). Hipster wars:

- Discovering elements of fashion styles. Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13,
- Kim, H. J., Lee, D. H., Niaz, A., Kim, C. Y., Memon, A. A., & Choi, K. N. (2021). Multiple-clothing detection and fashion landmark estimation using a single-stage detector. *IEEE Access*, 9, 11694-11704.
- Kim, J.-H., Kim, M., Park, M., & Yoo, J. (2023). Immersive interactive technologies and virtual shopping experiences: Differences in consumer perceptions between augmented reality (AR) and virtual reality (VR). *Telematics and Informatics*, 77, 101936.
- Kim, J., Gu, G., Park, M., Park, S., & Choo, J. (2024). Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,
- Kolisnik, B., Hogan, I., & Zulkernine, F. (2021). Condition-CNN: A hierarchical multi-label fashion image classification model. *Expert Systems with Applications*, 182, 115195.
- Kopf, J., Fu, C.-W., Cohen-Or, D., Deussen, O., Lischinski, D., & Wong, T.-T. (2007). Solid texture synthesis from 2d exemplars. In *ACM SIGGRAPH 2007 papers* (pp. 2-es).
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing*

systems, 25.

- Kuang, Z., Gao, Y., Li, G., Luo, P., Chen, Y., Lin, L., & Zhang, W. (2019). Fashion retrieval via graph reasoning networks on a similarity pyramid. Proceedings of the IEEE/CVF International Conference on Computer Vision,
- Kwatra, V., Schödl, A., Essa, I., Turk, G., & Bobick, A. (2003). Graphcut textures: Image and video synthesis using graph cuts. *ACM Transactions on Graphics (ToG)*, 22(3), 277-286.
- Lavoye, V., Sipilä, J., Mero, J., & Tarkiainen, A. (2023). The emperor's new clothes: self-explorative engagement in virtual try-on service experiences positively impacts brand outcomes. *Journal of Services Marketing*, 37(10), 1-21.
- Lazova, V., Insaftudinov, E., & Pons-Moll, G. (2019). 360-degree textures of people in clothing from a single image. 2019 International Conference on 3D Vision (3DV),
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436-444.
- Lee, Rokkyu, Kang, M., Cho, M., & Park, G. (2019a). LA-VITON: A network for looking-attractive virtual try-on. Proceedings of the IEEE/CVF international conference on computer vision workshops,
- Lee, J., Hwang, S., Lee, J., & Kang, S. (2020). Comparative performance characterization of mobile AR frameworks in the context of AR-based grocery shopping applications. *Applied Sciences*, 10(4), 1547.

- Lee, S., Oh, S., Jung, C., & Kim, C. (2019b). A global-local embedding module for fashion landmark detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*,
- Levin, D. (1998). The approximation power of moving least-squares. *Mathematics of computation*, 67(224), 1517-1531.
- Li, C., & Wand, M. (2016). Combining markov random fields and convolutional neural networks for image synthesis. *Proceedings of the IEEE conference on computer vision and pattern recognition*,
- Li, J., Wang, S., & Paquette, E. (2024a). Texture-driven adaptive mesh refinement with application to 3D relief. *Computer-Aided Design*, 167, 103640.
- Li, J., Weng, J., Xu, H., Zhou, C., Zhang, D., & Lu, G. (2019a). Design of robotic mannequin formed by flexible belt net. *Computer-Aided Design*, 110, 1-10.
- Li, K., Wang, S., Zhang, X., Xu, Y., Xu, W., & Tu, Z. (2021a). Pose recognition with cascade transformers. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,
- Li, P., Xu, Y., Wei, Y., & Yang, Y. (2021b). Self-Correction for Human Parsing. *IEEE transactions on pattern analysis and machine intelligence*, 1-1.
<https://doi.org/10.1109/tpami.2020.3048039>
- Li, R., Guillard, B., & Fua, P. (2024b). Isp: Multi-layered garment draping with implicit sewing patterns. *Advances in neural information processing systems*, 36.
- Li, S., Liu, Z.-Q., & Chan, A. B. (2014). Heterogeneous multi-task learning for human

- pose estimation with deep convolutional neural network. Proceedings of the IEEE conference on computer vision and pattern recognition workshops,
- Li, T., Liang, Z., Zhao, S., Gong, J., & Shen, J. (2020a). Self-learning with rectification strategy for human parsing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,
- Li, X., Zhang, B., Sander, P. V., & Liao, J. (2019b). Blind geometric distortion correction on images through deep learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,
- Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., & Yang, M.-H. (2017). Diversified texture synthesis with feed-forward networks. Proceedings of the IEEE conference on computer vision and pattern recognition,
- Li, Y., Li, Y., Lu, J., Shechtman, E., Lee, Y. J., & Singh, K. K. (2021c). Collaging class-specific gans for semantic image synthesis. Proceedings of the IEEE/CVF International Conference on Computer Vision,
- Li, Y., Tang, S., Ye, Y., & Ma, J. (2019c). Spatial-aware non-local attention for fashion landmark detection. 2019 IEEE international conference on multimedia and expo (ICME),
- Li, Z., Shafiei, M., Ramamoorthi, R., Sunkavalli, K., & Chandraker, M. (2020b). Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

- Liang, X., Gong, K., Shen, X., & Lin, L. (2018). Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4), 871-885.
- Liang, X., Liu, S., Shen, X., Yang, J., Liu, L., Dong, J., Lin, L., & Yan, S. (2015a). Deep human parsing with active template regression. *IEEE transactions on pattern analysis and machine intelligence*, 37(12), 2402-2414.
- Liang, X., Xu, C., Shen, X., Yang, J., Liu, S., Tang, J., Lin, L., & Yan, S. (2015b). Human parsing with contextualized convolutional neural network. Proceedings of the IEEE international conference on computer vision,
- Lin, J., Sharma, G., & Pappas, T. N. (2023). Toward universal texture synthesis by combining texton broadcasting with noise injection in StyleGAN-2. *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, 3, 100092.
- Lin, Y.-L., & Wang, M.-J. J. (2016). The development of a clothing fit evaluation system under virtual environment. *Multimedia Tools and Applications*, 75(13), 7575-7587.
- Liu, J., & Lu, H. (2018). Deep fashion analysis with feature map upsampling and landmark-driven attention. Proceedings of the European Conference on Computer Vision (ECCV) Workshops,
- Liu, L., Xu, X., Lin, Z., Liang, J., & Yan, S. (2023a). Towards garment sewing pattern reconstruction from a single image. *ACM Transactions on Graphics (TOG)*, 42(6), 1-15.

- Liu, S., Feng, J., Domokos, C., Xu, H., Huang, J., Hu, Z., & Yan, S. (2013). Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia*, 16(1), 253-265.
- Liu, S., Ng, T.-T., Sunkavalli, K., Do, M. N., Shechtman, E., & Carr, N. (2015). PatchMatch-based automatic lattice detection for near-regular textures. Proceedings of the IEEE international conference on computer vision,
- Liu, S., Song, Z., Wang, M., Xu, C., Lu, H., & Yan, S. (2012). Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. Proceedings of the 20th ACM international conference on Multimedia,
- Liu, X., Li, J., & Lu, G. (2023b). Generating High-Fidelity Texture in RGB-D Reconstruction using Patches Density Regularization. *Computer-Aided Design*, 160, 103516.
- Liu, X., Lu, L., Sharf, A., Yan, X., Lischinski, D., & Tu, C. (2020). Fabricable dihedral Escher tessellations. *Computer-Aided Design*, 127, 102853.
- Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016a). Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. Proceedings of the IEEE conference on computer vision and pattern recognition,
- Liu, Z., Yan, S., Luo, P., Wang, X., & Tang, X. (2016b). Fashion landmark detection in the wild. European Conference on Computer Vision,
- Liu, Z., Yan, S., Luo, P., Wang, X., & Tang, X. (2016c). Fashion landmark detection in the wild. Computer Vision—ECCV 2016: 14th European Conference,

- Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14,
- Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (ToG)*, 34(6), 1-16.
- Loureiro, S. M. C. (2023). Artificial intelligence and extended reality in luxury fashion retail: Analysis and reflection. In *Augmented reality and artificial intelligence: The fusion of advanced technologies* (pp. 323-348). Springer.
- Luo, Y., Zheng, Z., Zheng, L., Guan, T., Yu, J., & Yang, Y. (2018). Macro-micro adversarial network for human parsing. Proceedings of the European conference on computer vision (ECCV),
- Luvizon, D. C., Tabia, H., & Picard, D. (2019). Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, 85, 15-22.
- Ma, J., Zhao, J., & Tian, J. (2013). Nonrigid image deformation using moving regularized least squares. *IEEE Signal Processing Letters*, 20(10), 988-991.
- Ma, Y., Zheng, J., & Xie, J. (2014). Foldover-free mesh warping for constrained texture mapping. *IEEE Transactions on Visualization and Computer Graphics*, 21(3), 375-388.
- Meng, Y., Mok, P. Y., & Jin, X. (2010). Interactive virtual try-on clothing design systems. *Computer-Aided Design*, 42(4), 310-321.
- Micchelli, C. A., & Pontil, M. (2005). On learning vector-valued functions. *Neural*

computation, 17(1), 177-204.

Minar, M. R., Tuan, T. T., Ahn, H., Rosin, P., & Lai, Y.-K. (2020). Cp-vton+: Clothing

shape and texture preserving image-based virtual try-on. CVPR Workshops,

Mir, A., Alldieck, T., & Pons-Moll, G. (2020). Learning to transfer texture from clothing

images to 3d humans. Proceedings of the IEEE/CVF Conference on Computer

Vision and Pattern Recognition,

Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., & Cucchiara, R. (2023).

Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on.

Proceedings of the 31st ACM International Conference on Multimedia,

Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., & Cucchiara, R. (2022).

Dress code: High-resolution multi-category virtual try-on. Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition,

Moritz, J., James, S., Haines, T. S., Ritschel, T., & Weyrich, T. (2017). Texture

stationarization: Turning photos into tileable textures. Computer graphics forum,

Nakamura, T., & Goto, R. (2018). Outfit generation and style extraction via

bidirectional lstm and autoencoder. *arXiv preprint arXiv:1807.03133*.

National-Retail-Federation. (2021). *National Retail Federation, 2021*.

<https://nrf.com/tag/nrf-2021>

Newell, A., Huang, Z., & Deng, J. (2017). Associative embedding: End-to-end learning

for joint detection and grouping. *Advances in neural information processing*

systems, 30.

- Niklasson, E., Mordvintsev, A., Randazzo, E., & Levin, M. (2021). Self-organising textures. *Distill*, 6(2), e00027. 00003.
- Optitex. (2024). <https://optitex.com/>
- Pan, Z., Wang, P., Xin, J. H., Zhang, M., & Shen, H.-L. (2005). An illumination distribution preserved colour substitution algorithm based on dichromatic reflection model. *Displays*, 26(3), 121-127.
- Pantano, E. (2015). *Successful technological integration for competitive advantage in retail settings*. IGI Global.
- Pantano, E., & Gandini, A. (2017). Exploring the forms of sociality mediated by innovative technologies in retail settings. *Computers in Human Behavior*, 77, 367-373.
- Papagiannidis, S., See-To, E., & Bourlakis, M. (2014). Virtual test-driving: The impact of simulated products on purchase intention. *Journal of Retailing and Consumer Services*, 21(5), 877-887.
- Park, J. J., Florence, P., Straub, J., Newcombe, R., & Lovegrove, S. (2019). Deepsdf: Learning continuous signed distance functions for shape representation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,
- Parmar, G., Zhang, R., & Zhu, J.-Y. (2022). On aliased resizing and surprising subtleties in gan evaluation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

- Paz, H., Novello, T., & Velho, L. (2024). Implicit Neural Representation of Tileable Material Textures. *arXiv preprint arXiv:2402.02208*.
- Peng, S. (2013). Development of an interactive robotic mannequin for fashion industry.
- Pérez, P., Gangnet, M., & Blake, A. (2003). Poisson image editing. In *ACM SIGGRAPH 2003 Papers* (pp. 313-318).
- Petrou, M. M., & Petrou, C. (2010). *Image processing: the fundamentals*. John Wiley & Sons.
- Pfister, T., Charles, J., & Zisserman, A. (2015). Flowing convnets for human pose estimation in videos. Proceedings of the IEEE international conference on computer vision,
- Portenier, T., Arjomand Bigdeli, S., & Goksel, O. (2020). Gramgan: Deep 3d texture synthesis from 2d exemplars. *Advances in neural information processing systems*, 33, 6994-7004.
- Pumarola, A., Sanchez-Riera, J., Choi, G., Sanfeliu, A., & Moreno-Noguer, F. (2019). 3dpeople: Modeling the geometry of dressed humans. Proceedings of the IEEE/CVF international conference on computer vision,
- Rabin, J., Peyré, G., Delon, J., & Bernot, M. (2012). Wasserstein barycenter and its application to texture mixing. *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*,
- Raffee, A. H., & Sollami, M. (2021). Garmentgan: Photo-realistic adversarial fashion

- transfer. 2020 25th International Conference on Pattern Recognition (ICPR),
- Raj, A., Sangkloy, P., Chang, H., Hays, J., Ceylan, D., & Lu, J. (2018). Swapnet: Image based garment transfer. European Conference on Computer Vision,
- Rauschnabel, P. A., Felix, R., & Hinsch, C. (2019). Augmented reality marketing: How mobile AR-apps can improve brands through inspiration. *Journal of Retailing and Consumer Services*, 49, 43-53.
- Rauschnabel, P. A., Felix, R., Hinsch, C., Shahab, H., & Alt, F. (2022). What is XR? Towards a framework for augmented and virtual reality. *Computers in human behavior*, 133, 107289.
- Rigby, C. (2013). Fashion retailers opt for virtual fitting rooms. *Internet Retailing*.
- Rincon, L. (2023). Virtually try on clothes with a new ai shopping feature. *Google, Jun*.
- Roberts-Islam, B. (2020). How digital fashion could replace fast fashion, and the startup paving the way. *Forbes*, 21(2020).
- Rodriguez-Pardo, C., & Garces, E. (2022). SeamlessGAN: Self-Supervised Synthesis of Tileable Texture Maps. *arXiv preprint arXiv:2201.05120*.
- Rodriguez-Pardo, C., Suja, S., Pascual, D., Lopez-Moreno, J., & Garces, E. (2019). Automatic extraction and synthesis of regular repeatable patterns. *Computers & Graphics*, 83, 33-41.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

- Ruan, T., Liu, T., Huang, Z., Wei, Y., Wei, S., & Zhao, Y. (2019). Devil in the details: Towards accurate single and multiple human parsing. Proceedings of the AAAI conference on artificial intelligence,
- Rudin, L. I., Osher, S., & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4), 259-268.
- Santos, L. R., Montagna, G., & Neto, M. J. P. (2020). The virtualization of the fashion product. Advances in Industrial Design: Proceedings of the AHFE 2020 Virtual Conferences on Design for Inclusion, Affective and Pleasurable Design, Interdisciplinary Practice in Industrial Design, Kansei Engineering, and Human Factors for Apparel and Textile Engineering, July 16–20, 2020, USA,
- Särmäkari, N. (2021). “Digital fashion” on its way from niche to the new norm. In *«Новая норма» Гардеробные и телесные практики в эпоху пандемии* (pp. 117-134). Novoe literaturnoe obozrenie.
- Schaefer, S., McPhail, T., & Warren, J. (2006). Image deformation using moving least squares. In *ACM SIGGRAPH 2006 Papers* (pp. 533-540).
- Sederberg, T. W., & Parry, S. R. (1986). Free-form deformation of solid geometric models. Proceedings of the 13th annual conference on Computer graphics and interactive techniques,
- Sendik, O., & Cohen-Or, D. (2017). Deep correlations for texture synthesis. *ACM Transactions on Graphics (ToG)*, 36(5), 1-15.
- Shaham, T. R., Dekel, T., & Michaeli, T. (2019). Singan: Learning a generative model

- from a single natural image. Proceedings of the IEEE/CVF international conference on computer vision,
- Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. Proceedings of the 1968 23rd ACM national conference,
- Shi, Z., Xu, Y., Shen, Y., Zhao, D., Chen, Q., & Yeung, D.-Y. (2022). Improving 3d-aware image synthesis with a geometry-aware discriminator. *Advances in neural information processing systems*, 35, 7921-7932.
- Shilkrot, R., Cohen-Or, D., Shamir, A., & Liu, L. (2012). Garment personalization via identity transfer. *IEEE computer graphics and applications*, 33(4), 62-72.
- Shin, E., & Baytar, F. (2014). Apparel fit and size concerns and intentions to use virtual try-on: Impacts of body satisfaction and images of models' bodies. *Clothing and Textiles Research Journal*, 32(1), 20-33.
- Simo-Serra, E., Fidler, S., Moreno-Noguer, F., & Urtasun, R. (2015). Neuroaesthetics in fashion: Modeling the perception of beauty. CVPR,
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Sitzmann, V., Martel, J., Bergman, A., Lindell, D., & Wetzstein, G. (2020). Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33, 7462-7473.
- Smythe, D. B. (1990). A two-pass mesh warping algorithm for object transformation and image interpolation. *Rapport technique*, 1030, 31.

- Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *International conference on machine learning*,
- Song, D., Li, T., Mao, Z., & Liu, A.-A. (2020). SP-VITON: shape-preserving image-based virtual try-on network. *Multimedia Tools and Applications*, 79, 33757-33769.
- Statistics, O.-s. (2024). *online shopping statistics*. <https://optinmonster.com/online-shopping-statistics/>
- Struski, Ł., Knop, S., Spurek, P., Daniec, W., & Tabor, J. (2022). Locogan—locally convolutional gan. *Computer Vision and Image Understanding*, 221, 103462.
- Sun, G.-L., Wu, X., Chen, H.-H., & Peng, Q. (2015). *Clothing Style Recognition using Fashion Attribute Detection* Proceedings of the 8th International Conference on Mobile Multimedia Communications,
- Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,
- Sun, X., Shang, J., Liang, S., & Wei, Y. (2017). Compositional human pose regression. *Proceedings of the IEEE international conference on computer vision*,
- Toshev, A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*,

- Ulyanov, D., Lebedev, V., Vedaldi, A., & Lempitsky, V. (2016). Texture networks: Feed-forward synthesis of textures and stylized images. *arXiv preprint arXiv:1603.03417*.
- Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2017). Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. Proceedings of the IEEE conference on computer vision and pattern recognition,
- Umetani, N., Kaufman, D. M., Igarashi, T., & Grinspun, E. (2011). Sensitive couture for interactive garment modeling and editing. *ACM Trans. Graph.*, 30(4), 90.
- Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., & Schmid, C. (2018). Bodynet: Volumetric inference of 3d human body shapes. Proceedings of the European conference on computer vision (ECCV),
- Vecchio, G., Martin, R., Roullier, A., Kaiser, A., Rouffet, R., Deschaintre, V., & Boubekur, T. (2023). Controlmat: a controlled generative approach to material capture. *arXiv preprint arXiv:2309.01700*.
- Wang, Zheng, H., Liang, X., Chen, Y., Lin, L., & Yang, M. (2018a). Toward characteristic-preserving image-based virtual try-on network. Proceedings of the European conference on computer vision (ECCV),
- Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., & Yang, M. (2018b). Toward characteristic-preserving image-based virtual try-on network. Proceedings of the European conference on computer vision (ECCV),

- Wang, J., Zhong, Y., Li, Y., Zhang, C., & Wei, Y. (2019a). Re-identification supervised texture generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,
- Wang, P., Zhang, M., Xin, J. H., Pan, Z., & Shen, H. (2003). An image-based texture mapping technique for apparel products exhibition and interior design. *Displays*, 24(4-5), 179-186.
- Wang, T. Y., Ceylan, D., Popović, J., & Mitra, N. J. (2018c). Learning a shared shape space for multimodal garment design. *ACM Transactions on Graphics (ToG)*, 37(6), 1-13.
- Wang, W., Xu, Y., Shen, J., & Zhu, S.-C. (2018d). Attentive fashion grammar network for fashion landmark detection and clothing category classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,
- Wang, W., Zhang, Z., Qi, S., Shen, J., Pang, Y., & Shao, L. (2019b). Learning compositional neural information fusion for human parsing. Proceedings of the IEEE/CVF International Conference on Computer Vision,
- Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), 600-612.
- Wang, Z., Gu, Y., Zhang, Y., Zhou, J., & Gu, X. (2017). Clothing retrieval with visual attention model.

- WANGZ, B. A., & Sheikh, H. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600-612.
- Wei, S.-E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*,
- Weng, Y., Xu, W., Wu, Y., Zhou, K., & Guo, B. (2006). 2D shape deformation using nonlinear least squares optimization. *The visual computer*, 22, 653-660.
- Wu, Z., Lin, G., Tao, Q., & Cai, J. (2019). M2e-try on net: Fashion from model to everyone. *Proceedings of the 27th ACM international conference on multimedia*,
- Xie, J., Lu, Y., Zhu, S.-C., & Wu, Y. (2016). A theory of generative convnet. *International conference on machine learning*,
- Xie, X., Zhu, Z., Wu, Z., Xu, Z., & Zhu, Y. (2023a). Learning in a single domain for non-stationary multi-texture synthesis. *arXiv preprint arXiv:2305.06200*.
- Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., & Liang, X. (2023b). Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,
- Xu, H., Li, J., Weng, J., Deng, H., Zhou, C., Zhang, D., & Lu, G. (2018). The design of a deformable robotic mannequin. *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*,

- Xu, Y., Gu, T., Chen, W., & Chen, C. (2024). Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv preprint arXiv:2403.01779*.
- Yamaguchi, K., Berg, T. L., & Ortiz, L. E. (2014). Chic or social: Visual popularity analysis in online fashion networks. Proceedings of the 22nd ACM international conference on Multimedia,
- Yamaguchi, K., Hadi Kiapour, M., & Berg, T. L. (2013). Paper doll parsing: Retrieving similar styles to parse clothing items. Proceedings of the IEEE international conference on computer vision,
- Yamaguchi, K., Kiapour, M. H., Ortiz, L. E., & Berg, T. L. (2012). Parsing clothing in fashion photographs. 2012 IEEE Conference on Computer vision and pattern recognition,
- Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., & Luo, P. (2020). Towards photo-realistic virtual try-on by adaptively generating-preserving image content. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,
- Yang, W., & Feng, J. (2009a). 2D shape manipulation via topology-aware rigid grid. *Computer Animation and Virtual Worlds*, 20(2-3), 175-184.
- Yang, W., & Feng, J. (2009b). 2D shape morphing via automatic feature matching and hierarchical interpolation. *Computers & Graphics*, 33(3), 414-423.
- Yim, M. Y.-C., & Park, S.-Y. (2019). “I am not satisfied with my body, so I like augmented reality (AR)”: Consumer responses to AR-based product

- presentations. *Journal of Business Research*, 100, 581-589.
- Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., & Wang, J. (2021). Lite-hrnet: A lightweight high-resolution network. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,
- Yu, N., Barnes, C., Shechtman, E., Amirghodsi, S., & Lukac, M. (2019a). Texture mixer: A network for controllable synthesis and interpolation of texture. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,
- Yu, R., Wang, X., & Xie, X. (2019b). Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. Proceedings of the IEEE/CVF international conference on computer vision,
- Yu, W., Liang, X., Gong, K., Jiang, C., Xiao, N., & Lin, L. (2019c). Layout-graph reasoning for fashion landmark detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,
- Yuan, Y., Chen, X., & Wang, J. (2020). Object-contextual representations for semantic segmentation. European conference on computer vision,
- Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the IEEE conference on computer vision and pattern recognition,
- Zhang, X., Chen, Y., Tang, M., Wang, J., Zhu, X., & Lei, Z. (2022a). Human parsing with part-aware relation modeling. *IEEE Transactions on Multimedia*, 25, 2601-2612.

- Zhang, X., Chen, Y., Zhu, B., Wang, J., & Tang, M. (2020a). Part-aware context network for human parsing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,
- Zhang, Z., Liu, Y., Han, C., Shi, H., Guo, T., & Zhou, B. (2022b). Petsgan: Rethinking priors for single image generation. Proceedings of the AAAI Conference on Artificial Intelligence,
- Zhang, Z., Su, C., Zheng, L., & Xie, X. (2020b). Correlating edge, pose with parsing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,
- Zhao, J., Li, J., Cheng, Y., Sim, T., Yan, S., & Feng, J. (2018). Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. Proceedings of the 26th ACM international conference on Multimedia,
- Zhao, J., Ma, J., Tian, J., Ma, J., & Zhang, D. (2011). A robust method for vector field learning with application to mismatch removing. CVPR 2011,
- Zheng, Z., Yu, T., Wei, Y., Dai, Q., & Liu, Y. (2019). Deepphuman: 3d human reconstruction from a single image. Proceedings of the IEEE/CVF International Conference on Computer Vision,
- Zhou, T., Wang, W., Liu, S., Yang, Y., & Van Gool, L. (2021). Differentiable multi-granularity human representation learning for instance-aware human semantic parsing. Proceedings of the IEEE/CVF conference on computer vision and

pattern recognition,

Zhou, X., Hasan, M., Deschaintre, V., Guerrero, P., Sunkavalli, K., & Kalantari, N. K.

(2022). TileGen: Tileable, Controllable Material Generation and Capture.

SIGGRAPH Asia 2022 Conference Papers,

Zhou, Y., Zhu, Z., Bai, X., Lischinski, D., Cohen-Or, D., & Huang, H. (2018). Non-

stationary texture synthesis by adversarial expansion. *arXiv preprint*

arXiv:1805.04487.

Zhu, H., Cao, Y., Jin, H., Chen, W., Du, D., Wang, Z., Cui, S., & Han, X. (2020). Deep

fashion3d: A dataset and benchmark for 3d garment reconstruction from single

images. Computer Vision—ECCV 2020: 16th European Conference, Glasgow,

UK, August 23–28, 2020, Proceedings, Part I 16,

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017a). Unpaired image-to-image

translation using cycle-consistent adversarial networks. Proceedings of the

IEEE international conference on computer vision,

Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., &

Kemelmacher-Shlizerman, I. (2023). Tryondiffusion: A tale of two unets.

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition,

Zhu, S., Urtasun, R., Fidler, S., Lin, D., & Change Loy, C. (2017b). Be your own prada:

Fashion synthesis with structural coherence. Proceedings of the IEEE

international conference on computer vision,

Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., & Bai, X. (2019). Progressive pose attention transfer for person image generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,