International Journal of Science and Engineering (issued by NUTN) Vol.11 No.2(2021):41-57

Semantic-Guided and Detail-Preserving Image-Based 2D Virtual Try-On Networks

Hsin-Hui Wang[1], Man-Ling Guo[2], Chow-Sing Lin[3*]

[1,2,3]Department of Computer Science & Information Engineering, National University of Tainan, Tainan, 70005, Taiwan

[1]E-mail : whh871002@gmail.com
[2]Email : guomanling@gmail.com
[3*]Email : mikelin@mail.nutn.edu.tw

Abstract

2D virtual try-on has become a hot topic in recent years. It can change what a person image wearing by inputting a desired clothes image. In this study, we propose a visual try-on network, namely Semantic-guided and Detail-preserving Image-based 2D Virtual Try-On Networks(SD-VTON), which improve the architecture of a novel network, ACGPN. Considering the three major modules of ACGPN, including the Semantic Generation Module(SGM), the Clothes Warping Module (CWM), and the Content Fusion Module(CFM), we find out that SGM is the main reason causing the problems mentioned above. Consequently, we substitute UNet++ for the original neural network to improve human parsing, and hoping it makes the same or even better result with fewer epochs. In the test result of 200 epochs, there are 86.1% SSIM scores higher than ACGPN's. Moreover, the result of 140 epochs is extremely close to that of 200 epochs, which means that we can save more time on training. In terms of practicality, SD-VTON add a Clothes Tailoring Module at the head of the overall architecture, generating semantic segmentation with JPPNet, and develop a virtual try-on system with higher quality.

, -

Keywords: Neural Network, Generative Adversarial Network(GAN), Semantic Segmentation, Body Parsing Virtual Try on

---

* Corresponding author: mikelin@mail.nutn.edu.tw
DOI：10.3966/222344892021101102004

基於圖像之語意導向與細節保留的 2D 虛擬試穿網路

王新慧，郭曼玲，林朝興*

國立臺南大學資訊工程系

摘要

2D 虛擬試穿是近年來熱門的研究項目，只要輸入欲穿上的衣服圖片，便能更換指定人物圖像的服裝。本研究提出了基於圖像之語意導向與細節保留的 2D 虛擬試穿網路(SD-VTON)，將基於較新穎的 ACGPN 網路架構去做改善，透過 ACGPN 三大架構：語意生成、衣服變形、內容合成，我們發現語意生成是造成錯誤的最大原因，因此將神經網路修改為 UNet++來改善人物解析，欲減少 epoch 來產生同等或更佳的生成結果，修改神經網路訓練深度或進行其他參數的調整，在 200 epochs 的測試結果中，有 86.1% 的 SSIM 分析值優於 ACGPN，且於 140 epochs 就有與其相近的試穿結果，這意味著我們將可以省去不少訓練的時間。在實用性方面，SD-VTON 在整個架構最初加入剪取衣服的步驟，利用 LIP-JPPNet 生成語意模板，開發一個品質更好的虛擬試穿系統。

關鍵詞：類神經網路、生成對抗網路、語意分割、人物解析、虛擬試穿

1. 簡介

「衣裝」是每個人天天需要思考的問題，隨著時尚理念的提升，我們也更注重衣服對於自身的適合度。當不方便出門而選擇網購衣服時；或是到服飾店但店家規定白色衣服不能試穿時；又或者想找件上衣來搭配自己擁有的單品時，總是因為缺乏實際視覺效果的問題而猶豫是否要購買。如果能檢視自己穿上衣服後的圖片，是不是就能解決這個煩惱呢？同樣是條紋上衣，也可能因為線條大小、顏色、方向影響衣服對於個人的適合度；同樣是藍色襯衫，每個人穿起來都會有不同的韻味。

本研究欲改良現有的虛擬試穿網路，讓使用者可以透過輸入任意姿勢的人物照片及欲穿上的衣服，得到試穿後的圖片結果。圖像的虛擬試穿模型，已有許多研究關注，其利用圖像生成研究，促使產生影像更接近真實，但除了有常見的像素到像素損失(pixel-to-pixel losses)、感知損失(perceptual loss)[4]外，對抗損失(adversarial loss)[5]在某種程度上雖然可以降低模糊問題，卻仍容易遺漏關鍵細節。此外，延伸出的虛擬試穿網路VITON[6]、CP-V TON[7]，這些方法只能處理輸入和輸出大致對齊的情形，而當處理姿勢複雜的人像、案例時，保留細節的能力會變得很差，也無法保留完整的人像，或是在目標上衣扭曲、覆蓋的過程中，易產生與預期結果無關的偽影，當人物原本衣物與目標上衣差別越大時，越容易造成實驗結果出現非預期的突出或模糊色塊，實驗結果尚存有很大的挑戰性；而較新發表的論文 ACGPN，對虛擬試穿結果提出了許多優化設計，但仍有衣服試穿區塊框取錯誤、變形不佳、圖案模糊等問題，是值得重新審視的。

我們認為，一個好的虛擬試穿網路必須包含以下重點：應保留目標服裝的紋理、圖案及刺繡等細節、目標衣物應完全適合該人的姿勢及身體部位、應保留不打算更換的衣物，例如褲子、裙子、在保持人像姿勢及身形的情況下，清楚地填補身體部位。因此，現有的虛擬試穿需要更縝密、更精確的深度學習網路來提升，以達到符合現實情形的預期結果。此外，由於現有 2D 虛擬試穿網路中，大多目標上衣的輸入只能是單純一張衣服的圖片，但往往在生活情境中，幾乎都是看見模特兒試穿而產生購買慾望，若能將模特兒試穿圖片作為輸入，轉移其穿著於另一人像上，將能更貼近實際情形。

本研究提出基於圖像之語意導向與細節保留的 2D 虛擬試穿網路(SD-VTON)，其主要目的為以下兩點：1.以 ACGPN 為主要架構，SD-VTON 針對問題所佔比例最高的人物解析瑕疵去做優化，修改 SGM 階段的G_1 網路架構，降低G_1 錯誤率，生成更貼近實際情形的試穿結果。2.採用 JPPNet 作為人物解析之方法，經訓練後得到模特兒身上正確的目標上衣區塊，再利用此生成目標上衣遮罩，利用遮罩將其剪取，即作為目標上衣，如此一來，目標上衣的輸入形式變得更加彈性，可以是一張圖片或模特兒試穿圖片。

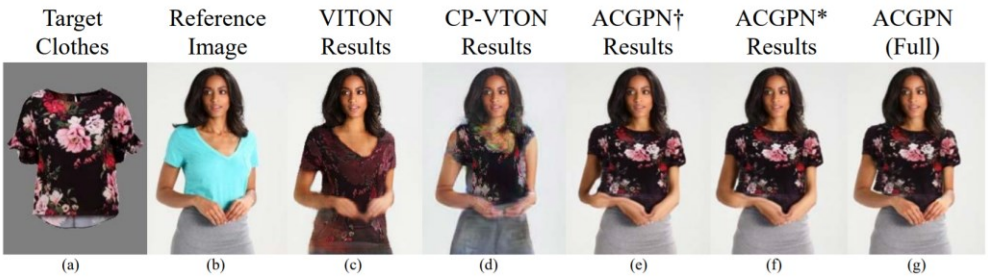

圖 1 虛擬試穿網路合成結果比較[1]

2. 相關技術

2.1 圖像合成 (Image Synthesis)

生成對抗網絡(GANs)由兩個神經網路組成，分別是鑑別網路(Discriminating Network)與生成網路(Generative Network)，生成網路試圖產生與真實無法區分的樣本，而鑑別網路要判別真實樣本與生成樣本，透過互相對抗產生出以假亂真的輸出圖像。條件生成對抗網絡(cGAN)可以將任意的輸入圖像轉換成真實圖像。GAN 對於高分辨率圖像生成一直存在著許多問題，其中對抗損失(adversarial loss)雖稱能使生成圖像看起來自然，但卻不穩定，因圖像到圖像的轉換中，大多情形是輸入和輸出彼此大致對齊，且有相同的基礎結構。

而在先前研究中，處理較大的空間變形時，就會遇到一些問題。大多數以未對齊圖像為條件的圖像到圖像轉換，都採用從粗略到精緻的方式來提高最終結果的質量，使用姿勢點(pose points)在 GAN 中進行變形，VITON 將目標上衣和預測上衣遮罩使用形狀上下文(shape context)匹配來估計 TPS[8]轉換，並生成變形的服裝圖像，由於形狀的匹配過於耗時，CP-VTON 設計了卷積神經網路(CNN)來估計輸入之目標上衣和輸出之合成圖像間衣服的 TPS 轉換，而不需要任何對應點。ACGPN 則是預先框出要合成的區塊，再進行後續訓練。

2.2 人物解析與生成 (Human Parsing and Person Image Generation)

近年來人物解析受到很大的關注，將人物身體各個部位或所穿著之衣物加以識別，亦稱作人體解析、服裝解析，所有組成人體的像素均被標記，分別分割出人體臉部、頭髮、上衣、褲子等區域。最著名的 SSL[9]在不受限制的環境與人像外觀條件下，引入 LIP (Look into Person)，擴展性、多樣性、難度上都有顯著的進步，利用數據集帶有廣泛視角(人物被遮擋且背景複雜)的 19 個語意部位標籤(semantic part labels)，分析最終結果的成功與失敗。為了實現更精確的人物解析，Liang 等人提出了 JPPNet，該網路結構與先前方法不同的是，他們加入了對人體結構的考慮，將人體姿勢估測(pose estimate)與人物解析有效地結合，而姿勢估測所標記的是關節點位置，能夠提供高層結構訊息，在預測姿勢訊息的同時進行人物解析，並將兩個任務的預測結果多次融合迭代，達到相互促進的作用，進而形成更準確的語意分割圖。像這樣的技術能延伸應用到將一個人身上的衣物剪取下來，如 LGVTON[10]。

虛擬試穿的目標是將服裝逼真地合成於新的人像上，因此 Lassner 等人提出[11]生成逼真服裝於人像身上，該模型在人像處理方面生成人物解析圖(parsing maps)，但合成圖像尚不清楚如何控制生成的服裝；而 Zhao 等人提出[12]解決了此問題，可以給定任意視角生成多視角的服裝圖像；PG2[13]以任意目標姿勢作為條件，將人物生成該姿勢的合成圖像，使服裝視角及人物姿勢得以控制生成圖像的結果；另外，FashionGAN[14]不只能改變人物穿著，還能透過文字描述產生新的服裝。

2.3 虛擬試穿 (Virtual Try-on)

最早的虛擬試穿系統大多是基於 3D 圖像模型來實作，如：利用有深度感測的相機進行人體形狀之 3D 測量，或是使用 SCAPE 實作 3D 人形建模等方式。但相較於 3D 圖像模型，基於 2D 圖像的生成模型具有更高的運算效率，因此本研究的目標是直接利用

2D 圖像合成出逼真的試穿圖像。第一篇相關的論文為 VITON，接著，CP-VTON 基於該方法提出了對齊網路(alignment network)與單次生成框架(single pass generative framework)，以保留服裝之特徵細節，皆為由簡到精的架構。ACGPN 與以往不同的是，第一步就先將人像將要試穿衣服的區塊預測出來，接著再做後續的生成遮罩、變形、合成，如此一來，可以直接降低人像原本衣服造成結果有誤的可能性，並非由簡到精的訓練，而是事先預測，再進行後續網路訓練，其為近期最有彈性且優越的虛擬試穿網路，因此我們將做為參考研究之。

VITON 可說是 2D 虛擬試穿技術的始祖，也是適合初步踏入該領域研讀的入門論文，他先透過 Person Representation 對輸入的人像參考圖分解為三個部分，分別是 Pose heatmap[15]、Body shape、和 Reserved regions。Pose heatmap 用來表示人體姿勢關鍵點；Body shape 能夠粗略地涵蓋人體形狀；Reserved regions 則是保留了人像的頭部 RGB 圖像，以便維持人物身分。這些特徵架構可以進一步限制圖像合成過程，使其盡可能與真實試穿時衣服該有的紋理、變形、圖案、露出人體部位相同。接著採取兩階段的合成訓練，第一階段為 Encoder-Decoder Generator Stage，此階段利用 Encoder-Decoder Generator 生成粗糙地合成圖像與預測的衣服遮罩，再到第二階段的 Refinement stage，將預測的衣服遮罩與目標上衣遮罩用來計算 TPS 變換參數，並將此參數套用到目標上衣圖像上，生成變形的上衣，最後將變形的上衣與第一階段產出的粗糙合成圖像融合，形成最終試穿結果。此研究跨出了虛擬試穿技術的一大步，但也存在以下幾個主要問題：1.目標衣服之質地、圖案及刺繡等細節未完整保留。2.目標衣服未正確變形成適合該人像姿勢之形狀。3.未保留不應更換的人像下著。4.在保持人像姿勢及身形的情況下，未清楚呈現身體部位。除了這些具體能辨別的標準未達成外，甚至有些圖像合成後是完全模糊而無法分辨原始輸入的資料。

CP-VTON 基於 VITON 的網路架構，在模組的部分作了調整，主要在第一階段改為 Geometric Matching Module (GMM)，而第二階段改為 Try-On Module (TOM)。GMM 先對輸入的目標上衣和人像參考圖提取高級特徵，接著通過 Correlation-layer 將兩個高級特徵組合為單個張量，並用來預測空間變換參數以進行 TPS 變換，產生變形的上衣。而 TOM 將人像參考圖與變形的上衣輸入到 U-Net 中，經過 encoder-decoder 得到粗糙的合成圖像，並預測了一個合成遮罩，接著利用遮罩將 TOM 的輸入融合在一起，得到最終結果。該研究結果相較於 VITON 保留了更多的衣服細節，但仍然有以下主要問題：1.目標衣服交界處模糊。2.目標衣服未正確變形成適合該人像姿勢之形狀。3.未保留人像下著，且每張人像下半身皆有特定一件褲子的偽影。4.在保持人像姿勢及身形的情況下，未清楚呈現身體部位。

以上 VITON 及 CP-VTON 皆有幾乎相似的幾個大問題，且這兩個方法都無法處理姿勢較複雜的人像，如：人物肢體遮蔽了衣服部分，或有手臂交叉情形，生成結果幾乎都會有缺陷，像是手指變得模糊不清、手臂糊成一團，細節還原能力很差，有時更會有無法預期的錯誤發生。

ACGPN 主要在 SGM、CWM、CFM 三大部分加入新的設計：首先，SGM 先預測出穿上衣服的區塊，形成遮罩，這使得往後進行訓練時，不會受到原始人像穿著的影響，而造成試穿錯誤；CWM 主要採用 STN[16]和 TPS 進行衣服變形轉換，為了防止衣服過度扭曲變形，加入了二階差分約束去控制變形範圍，使衣服的圖案、紋理在變形後依舊保持於正確位置上；CFM 利用前面步驟產生出來的遮罩，去還原填滿在試穿後會暴露或遮蔽的人體部位。這三大部分改善了 VITON 及 CP-VTON 部分問題，卻依舊有不足之處或造成另外新生的錯誤發生：1.目標衣服之袖子或領子等未符合人像姿勢變形。2.少量的合成圖像下身未完整保留。3.在保持人像姿勢及身形的情況下，未清楚呈現身體

部位。以上為 ACGPN 常見的問題，但明顯錯誤範圍縮小了許多，不再有整張合成圖像模糊的情形發生，衣服之圖案、紋理也幾乎能成功保留。

上述三篇主要論文中，欲試穿上衣的格式皆為單純衣服圖像，而另有其他論文是以輸入一張模特兒圖像，並以該模特兒所穿上衣作為目標上衣，也就是說，要將兩個人像所穿衣物互換，這種技術實作上更加困難，第一，從模特兒身上取得衣服時，必須擺脫各種姿勢與身形，不失其服裝特徵完整地取下；第二，在轉移人像中，要呈現真實自然的試穿情形，並保留衣服原有的版型、圖案、紋理。關於此技術的相關研究，以 SwapNet[17]、LGVTON 兩種方式最為具體而有參考價值：SwapNet 可於圖片中任意姿勢的人像間做服裝轉移，此網路提出一種利用弱監督學習(weakly supervised learning)的方式來訓練。該研究先將欲試穿上衣剪取，這啟發我們思考以另一種形式取得輸入的目標上衣，但於重現該網路過程中，發現其生成結果並不理想，大多數圖像變得模糊不清，甚至完全無法辨認出原始圖像的人臉、目標衣服，因此，該網路架構設計上不列入本研究之主要參考，但利用人物解析來剪取目標上衣的作法仍有參考價值。LGVTON 是以一種新穎的自我監督(self-supervised)學習方法，自我監督的結合解決了模型對虛擬試穿場景中缺少配對訓練數據的問題。LGVTON 根據人的形狀和姿勢，使用兩種類型的標誌(landmark)來變形模型區塊，一是人類標誌，為人物之姿勢關鍵點；另一是時尚標誌，為服裝之結構關鍵點，接著利用標誌對應去變形為符合人像姿勢的試穿合成。為了避免場景雜質干擾造成變形效果不佳，LGVTON 提出了遮罩生成器模組，該模組試圖預測人體上服裝之真實分割遮罩，從而使圖像合成器解決變形問題。

圖 2 LGVTON 人物標誌與時尚標誌，及生成結果[10]

3. SD-VTON

自適內容生成與保留網路(Adaptive Content Generating and Preserving Network，ACGPN)為本研究主要參考的架構。為了增加實用性及提升試穿效果，SD-VTON 設計讓欲試穿衣服的輸入型態不再只能是一張完整且單純的衣服圖片，而可以是模特兒的照片，如同於把模特兒身上的衣服轉移到另一人像身上。那麼，於 ACGPN 的 SGM、CWM、CFM 三大步驟之前，必須先處理模特兒身上的衣服，因此加入了 CT 此步驟作為優化目標之一。

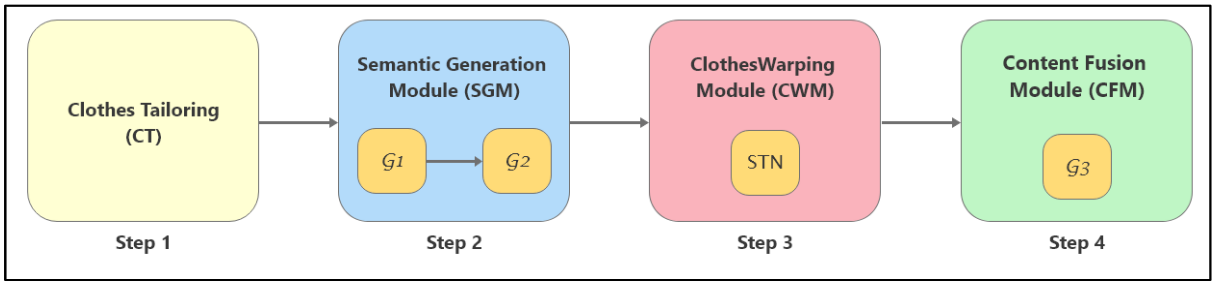

圖 3 SD-VTON 之主要流程圖

3.1 剪取上衣 (Clothes Tailoring, CT)

如圖 4，(a.)人物所穿之上衣為欲試穿之上衣，SD-VTON 利用 LIP_JPPNet 進行人物解析， 得到上衣的語意分割圖(b.)，接著將該部分形成上衣遮罩(c.)，最後利用(c.)剪取(a.)上模特兒身上的衣服，即可取得目標上衣(d.)，並作為下個階段的輸入：T_c。

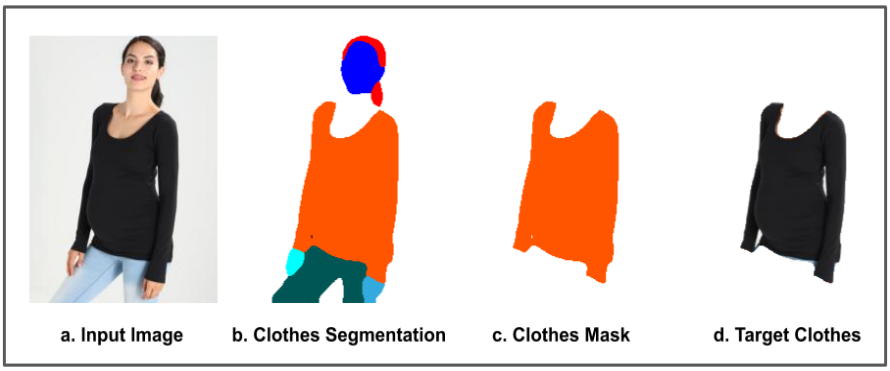

圖 4 剪取上衣示意圖

將人像參考圖 I 的手臂與身軀作為一區塊，當作輸入遮罩M^F，並輸入人物姿勢關鍵點M_p，及要穿上的目標衣服T_C。

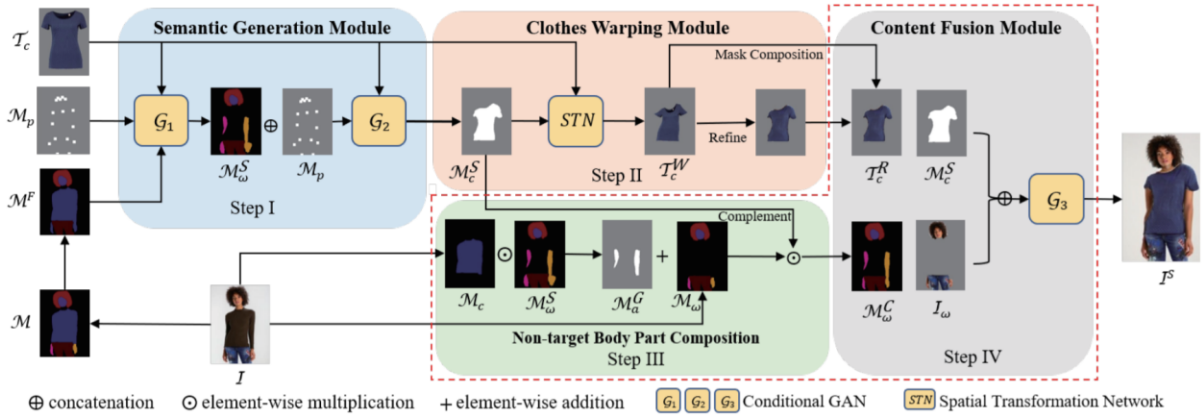

圖 5 ACGPN 流程架構[1]

3.2 語意生成模組 (Semantic Generation Module, SGM)

與一般的虛擬試穿網路不同，SGM 分為兩階段，利用身體部位和目標衣服的語意分割(semantic segmentation)，預先生成應暴露的身體部位和衣服變形區域的遮罩，使得在後續網路中，人像參考圖的原始服裝為完全未知，而不會干擾試穿網路，這就是自適性的提升身體部位與目標衣服區塊的精確度，有助於優化後續步驟結果。

如圖 5，SGM 的第一階段為訓練人物解析(body parsing) G_1，利用輸入的遮罩M^F、人物姿勢關鍵點M_p、目標衣服T_C，生成人物語意模板 M_w^S，預測會暴露出來的部位及目標衣服要穿上的區塊；第二階段，G_2 依據M_w^S、M_p、T_C，生成目標衣服的合成遮罩M_C^S。

兩階段都採用條件式生成對抗網路(cGAN)，利用 U-Net 作為生成器、pix2pixHD[18]中的判別器(discriminator)，來區分生成遮罩和真實遮罩。其中 cGAN 的損失(loss)公式為：

$$\begin{aligned}\mathcal{L}_1 =& \mathbb{E}_{x,y}\left[\log\left(\mathcal{D}\left(x,y\right)\right)\right] \\ &+ \mathbb{E}_{x,z}\left[\log\left(1-\mathcal{D}\left(x,\mathcal{G}\left(x,z\right)\right)\right)\right]\end{aligned} \tag{1}$$

x 表示輸入，y 表示真實遮罩，z 噪聲(noise)是從標準正態分佈採樣的輸入的附加通道。而試穿遮罩生成模組每個階段的總體目標函數表示為：

$$\mathcal{L}_m = \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2 \tag{2}$$

L_2 是 pixel-wise cross entropy loss[19]，是常用於語意分割的損失函數，可提升合成遮罩的質量，並獲得更精確的語意分析結果：λ_1 和 λ_2 是損失項的權衡(trade-off)參數，數值設為 1。

3.3 衣服變形模組 (Clothes Warping Module, CWM)

衣服變形的目的是讓衣服符合人像姿勢在視覺上自然穿上的情形，並保留人像與衣服該有的細節特徵。該模組根據生成的語義佈局對衣服進行變形操作，如果只訓練空間轉換網絡(Spatial Transformer Networks，STN)和運用薄板樣條(Thin-Plate Spline，TPS)，不能得到精確的轉換，尤其是在處理困難的情況下，如：複雜的紋理和豐富的色彩，會導致沒有對齊或模糊的結果。

圖 6 STN 有無約束條件之變形結果比較[1]

因此，為了解決這些問題，本研究引入二階差分約束(second-order difference constraint)，使得變形過程更加穩定，尤其是對於材質複雜的衣服。如圖 6 所示，與有約束條件的相比，沒有約束條件的目標衣服轉換顯示出明顯的變形錯誤和不合理的紋理。

以 SGM 生成的目標衣服合成遮罩M_C^S以及目標衣服T_C 作為輸入，訓練 STN 的學習參數進行變換，得到兩者間的對應，來生成變形的衣服T_C^w，引入以下約束L_3：

$$\begin{aligned}\mathcal{L}_3 =& \sum_{p\in\mathbf{P}} \lambda_r \left|\|pp_0\|_2 - \|pp_1\|_2\right| + \left|\|pp_2\|_2 - \|pp_3\|_2\right| \\ &+ \lambda_s \left(\left|S\left(p,p_0\right) - S\left(p,p_1\right)\right| + \left|S\left(p,p_2\right) - S\left(p,p_3\right)\right|\right)\end{aligned} \tag{3}$$

λ_r 和λ_s 是權衡超參數(hyper-parameters)，數值設為 0.1，利用最小化 max(L_3 -△, 0)來進行限制，而△是一個超參數。如圖 6 所示，p(x, y)代表限制點，上下左右各以p_0 (x_0，y_0)、p_1 (x_1，y_1)、 p_2 (x_2，y_2)、 p_3 (x_3，y_3)表示控制範圍，S(p, p_i)為斜率(i=0, 1, 2, 3)，L_3 作為 TPS 轉換的約束條件，最小化各軸的兩個相鄰區間的距離和斜率的距離來保持轉換的共線性、平行度和不變性。為了避免除數為零的錯誤：

$$\begin{aligned} &|S(p,p_i)-S(p,p_j)| \\ &=|(y_i-y)(x_j-x)-(y_j-y)(x_i-x)| \end{aligned} \tag{4}$$

而變形損失(warping loss) L_w，用來衡量變形的衣服T_c^w和真實人像參考圖I_C

$$\mathcal{L}_w = \mathcal{L}_3 + \mathcal{L}_4 \tag{5}$$

其中，$L_4 = \| T_c^w - I_c \|_1$。接著將變形好的衣服放入精緻化網路來產生更多細節部分，最後，利用學習矩陣α ($0 \le \alpha_{ij} \le 1$)生成精緻化後的衣服圖像T_c^R

$$\mathcal{T}_c^R = (1-\alpha)\odot \mathcal{T}_c^W + \alpha \odot \mathcal{T}_c^R \tag{6}$$

⊙為 element-wise 矩陣乘法，每個矩陣元素對應相乘。此部分是參考 CP-VTON，訓練的目標是使網路輸出T_c^w和T_c^R最小化差異，VGG 感知損失的定義為：

$$\mathcal{L}_{\mathrm{VGG}}(\mathcal{T}_c^R, \mathcal{T}_c^W) = \sum_{i=1}^{5} \lambda_i \left\| \phi_i(\mathcal{T}_c^R) - \phi_i(\mathcal{T}_c^W) \right\|_1 \tag{7}$$

T_c^R在精緻化後可以充分保留目標衣服的特徵，本研究認為 CWM 加入二階差分約束可以有效防止每個局部過度錯誤變形，同時又保有 TPS 的靈活性。

3.4 內容合成模組 (Content Fusion Module, CFM)

完成衣服變形後，自適性的合成出視覺上自然的試穿效果，更是一大挑戰，目標上衣必須清楚明確呈現出細節，人像的部分也須雕刻出細節。先前的技術都是採用由粗到細去刻劃出試穿效果，但是效果都不如預期，且無法重建更精緻的細節，因此，CFM 設計了兩個步驟，如圖 5 中的步驟 3 和 4，尤其是步驟 3，自適性的保留會暴露出來的人體部位，如：手臂、手指等；接著，步驟 4 利用此生成結果，G3 填補暴露的部位，同時與先前 CWM 生成的精緻化衣服變形圖像T_c^R合成。

CFM 整合了來自合成人像部位的遮罩、變形的衣服圖像、原始人像參考圖像的人像，以自適地確定合成圖像中不同人體部位的生成或保存。生成人體部位實作如下：

$$\mathcal{M}_a^G = \mathcal{M}_\omega^S \odot \mathcal{M}_c, \tag{8}$$

$$\mathcal{M}_\omega^C = (\mathcal{M}_a^G + \mathcal{M}_\omega) \odot (1 - \mathcal{M}_c^S), \tag{9}$$

$$\mathcal{I}_\omega = \mathcal{I}_{\omega'} \odot (1 - \mathcal{M}_c^S), \tag{10}$$

M_C 和M_w是人像參考圖 I 之身體和其他部位的遮罩；M_a^G是由M_C 和生成人物語意模板M_w^S組成；而合成的身體遮罩M_w^C是由M_w加上填補遮罩M_a^G，再與目標衣服合成遮罩M_c^S進

行合成。M_w^c保有與M_w^s相似的佈局，且利用M_a^c產生I_w恢復原始參考人像圖所要露出的人體部位。這就是 CFM 的精隨，可以將短袖衣服轉移到原本為長袖衣服的人像參考圖；反之，若$M_a^c=0$，多餘的身體部位將被陰影化。圖 7 為遭遮擋之人體部位經過填補後之效果。

圖 7 CFM 填補原本遭長袖遮擋之手臂效果呈現[1]

4. 研究結果

4.1 ACGPN 結果分析

ACGPN 此篇論文在探討虛擬試穿結果時，分別利用了定性分析與定量分析的方式與 VITON、CP-VTON、VTNFP[20]三個試穿網路做比較。但經由我們實作訓練結果時，重新審視，發現結果並不如論文中所描述：

(一)定量分析 (Quantitative Analysis)

採用 Structural SIMilarity(SSIM)來測量合成圖像和實際情形之間的相似性，和 Inception Score(IS)來評估合成圖像的視覺質量，這可以說是客觀、量化性指標方式來比較試穿效果，比對結果相似性由低到高依序為 VITON、CP-VTON、VTNFP、ACGPN，這更證明了 ACGPN 設計的優越性。

(二)定性分析 (Qualitative Analysis)

在視覺上，VITON 生成的圖像都有模糊、混色、偽影和雜質紋理等問題；CP- VTON 與 VITON 相比，雖然有得到更好的試穿效果，但處理難度較高的參考人像圖時，仍會造成下半身、身體部位模糊等嚴重問題；而 VTNFP 使用佈局表示法(segmentation representation)進一步保留身體部位的完整性，但保留部分的還原效果或細節依舊不佳，像是手指細節或衣服圖案扭曲，都無法呈現與現實相符的自然情形；ACGPN 的方法可以更有效地降低偽影、提高真實度，並且能更好的保留衣服紋理細節，達到與真實相符的試穿效果。

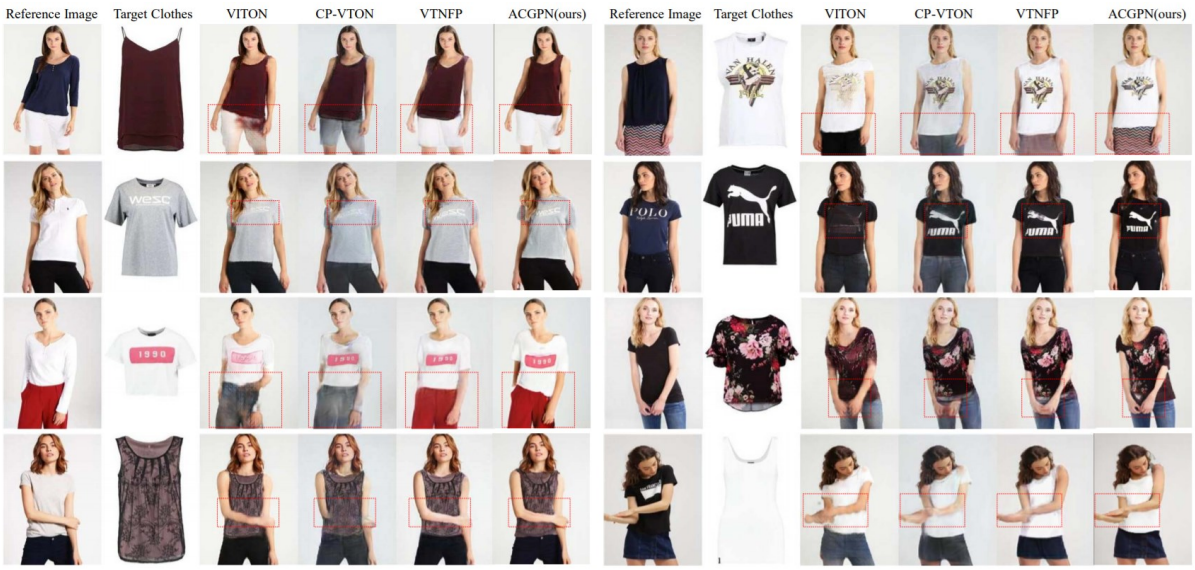

圖 8 視覺比較四種虛擬試穿方法[1]

(三)實際結果

由於只有 CP-VTON、ACGPN 原始碼完整可得，其中，又以 ACGPN 為主要研究架構，重現試穿網路後，我們將其結果分為以下 6 類加以分析：

1. G_1 階段錯誤：預測的人物語意模板(M_w^S)有問題
2. G_2 階段錯誤：M_w^S正確，預測的衣服遮罩有誤
3. Step2 錯誤：衣服變形、歪斜或角度有問題、圖案消失或模糊。
4. Step3 錯誤：填補錯誤(膚色不對或有雜質、非人物解析問題的手部怪異、領口受到參考人像原本的衣服影響)
5. 其他：目標上衣未正確讀取(輸入非衣服正面、不完整的衣服或未正確去背)、人物姿勢特殊(輸入人像為背面)
6. 正常：成功合成逼真的試穿圖像

利用視覺判定 ACGPN 與 CP-VTON 的結果，發現兩者試穿成功率不相上下，ACGPN 並不如論文中所描述的真實，但可以發現，CP-VTON 的錯誤都是嚴重偽影、身體部位模糊等，相對來說，ACGPN 穩定而錯誤範圍小。於是我們從中思考到底哪一步出了問題，檢視後發現，大多數在 SGM 步驟時生成的人物語意模板M_w^S就出現錯誤，造成後續試穿效果不佳，因而推測該部分是最需要改善的。

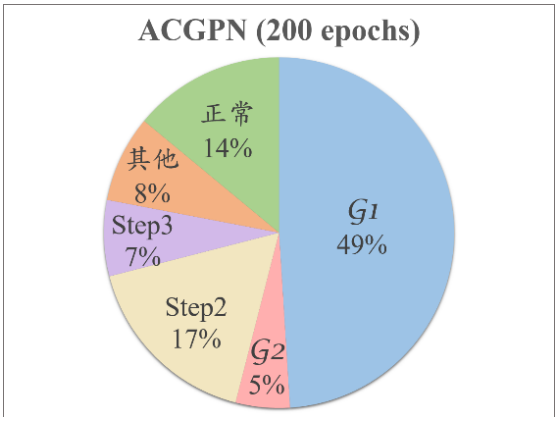

圖 9 ACGPN 結果分類情形

4.2 優化人物解析

經研究發現 ACGPN 大部分的問題都是出在 SGM 步驟生成的人物語意模板有誤，因此試著找尋人物解析的替代方案，現有的方法較常見的有 SSL 及 JPPNet，圖 10 是比較結果。JPPNet 較不注重細節，邊角會偏圓，但較穩定，錯誤不明顯；SSL 細節表現較好，但偶爾會有較不合理的錯誤，下身較常消失。

經過比較後發現 JPPNet 是較新、也是目前較好的方法，因此嘗試以 JPPNet 來實作 ACGPN 的 dataset，並將產出的結果與原先 ACGPN 所使用的 SSL 的人物語意模板做比較，由以上結果可見，雖然 JPPNet 的錯誤率是較高的，但他仍有參考價值，可以思考如何將 JPPNet 的穩定性套用在 SSL 的人物解析上，提高正確率，以提升整體試穿效果。推測改善此部分即能解決大多關於試穿區塊錯誤的問題，接著繼續設計下一步改良虛擬試穿網路的方法，檢視錯誤問題是否有共通點，推估在哪一階段開始出問題，進一步去測試如何修改架構能最佳化或加入新的研究方法，使該研究盡可能趨於完善、自然，能正確呈現視覺上的真實試穿效果。

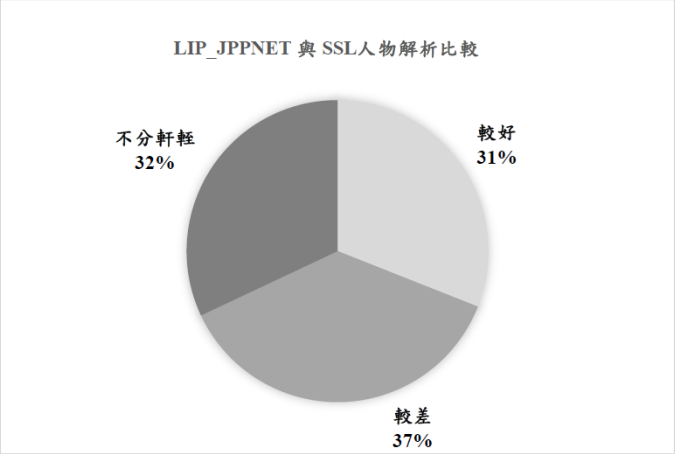

圖 10 JPPNet 與 SSL 人物解析比較結果

4.3 剪取衣服

目前較常見的虛擬試穿網路大多採用輸入一張衣服商品圖作為目標衣物的方式，且衣服必須為正面、不得有遮擋、不得扭曲變形，包含本研究的主要架構 ACGPN 亦是如此。然而人們在挑選衣服時，往往是看見模特兒穿上的效果而產生購買慾望，這使得我們嘗試將衣服從一張圖像的模特兒身上擷取衣服，此方法多樣性又直觀。虛擬試穿網路相關研究中，有些採取的是輸入一張人像照，並以該人像所穿的上衣取下作為目標衣物，這與我們的想法一致，但現有的研究中，SwapNet 剪取衣服的效果不佳而無法採用，生成結果幾乎模糊到無法辨識，因此，SD-VTON 研究出另一方法：將模特兒圖像利用 JPPNet 得到人物語意模板，並產生其上衣遮罩，再利用此遮罩剪取模特兒身上的衣服作為目標上衣。

圖 11 剪取衣服結果

由圖 11 可見，目前剪取衣服的方法受到諸多限制，像是衣服遭遮擋形成的缺漏很可能造成後續的合成效果不佳，另外，人物語意模板的品質也直接影響到剪取衣服的效果。我們將裁剪下來的衣服放到試穿網路中測試，發現結果如預期所估，當剪取的衣服及人物語意模板皆完整無缺時，生成的試穿圖像就會較逼真；反之，當其中一者有瑕疵時，也會造成合成結果不佳。

要如何完整得到人像身上的衣服又不失其真實性，目前尚有很大的限制，造成進步空間也有限，例如：衣服不能遭頭髮或其他身體部位遮擋，否則衣服不完整有缺角可能造成後續的合成效果不佳，對此必須在輸入模特兒人像圖片時就加以限制，但是這些限制也增加了不便性，勢必得再多加思考如何降低輸入資料的限制，開發出更具彈性的虛擬試穿系統。

圖 12 以從人像圖上剪取之上衣作為輸入的合成結果

4.4 改善試穿網路

現有的虛擬試穿網路大多採用 U-Net，但是這種模型有侷限性，到底神經網路要幾層，是否越多層越好，還是一個開放性問題，U-Net 容易花費很多時間，最終沒有得到預期目標的結果。為了克服這個限制，SD-VTON 將 U-Net 替換成 UNet++，這是今年新提出的一篇運用在醫學上結構分割之論文，當中 UNet++有效緩解未知網絡深度，不同

深度的 U-Net 集合使用深度監控器同時進行編碼器和共同學習，並重新設計跳過複雜的連接，於解碼器子網絡上聚合深度不同的特徵，從而產生高度靈活的特徵融合，且加快了訓練的速度。ACGPN 原先採用的是 U-Net(L^4)，因此，SD-VTON 改以 UNet ++進行調整，測試此模型是否對改善試穿網路有幫助。

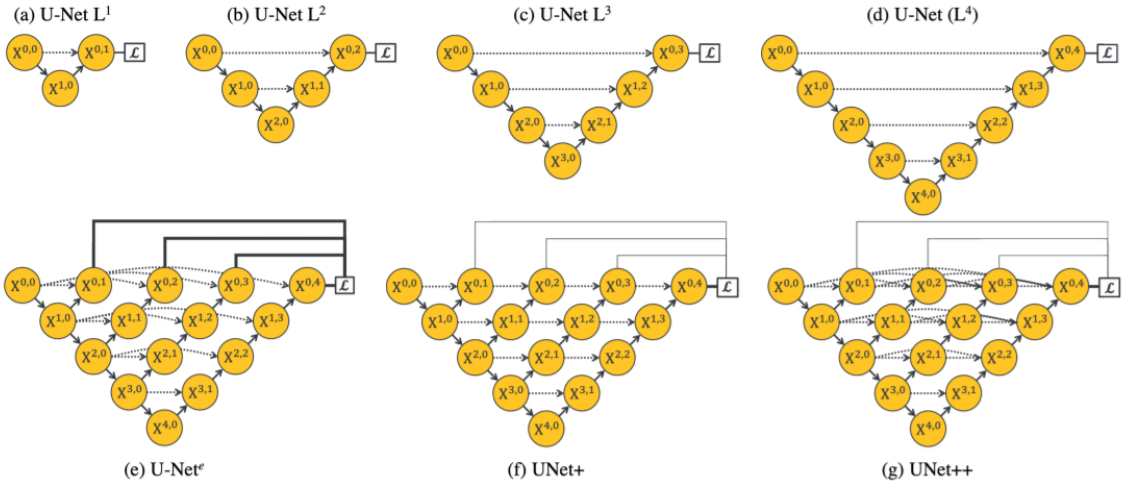

圖 13 從 U-Net 到 UNet++的演變[2]

(一) 定量分析 (Quantitative Analysis)

在資料集裡含有單純目標上衣及人像真實的試穿圖，將此兩項資料作為輸入，測試虛擬試穿結果，並以人像真實的試穿圖做為原圖，與生成結果進行結構相似度(SSIM)分析，SD-VTON 訓練 20 epochs 時，就與 ACGPN 的 200 epochs 有相近的試穿效果，而到 50 epochs 以後更是超越其值，最佳結果仍為 200 epochs，SSIM 值達到了 0.861，但事實上，訓練到 140 epochs 時就有與其相近且穩定的結果了；另外，將每張試穿合成圖分別作與原圖的結構相似度比較，SD-VTON 優於 ACGPN 記錄為 1，否則記錄為 0，最後取平均值即為 Quality Improvement(QI)，該值顯示，在整個訓練過程中，皆有優於 ACGPN 的結果，並且持續提升中，在 170 epochs 時達到了 0.807，因此，透過客觀的分析方法，推測將G_1 生成器改為 UNet++能有效解決部分預測試穿人物語意模板的問題。

表 1 ACGPN 與 SD-VTON 的 SSIM 結果比較

-	ACGPN	SD-VTON																		
Epochs	200	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
SSIM	0.845	0.843	0.846	0.845	0.841	0.847	0.849	0.845	0.849	0.853	0.851	0.853	0.857	0.851	0.860	0.857	0.860	0.860	0.860	0.858	0.861
QI	-	0.556	0.585	0.537	0.585	0.599	0.527	0.501	0.590	0.670	0.590	0.675	0.688	0.547	0.785	0.698	0.781	0.807	0.642	0.561	0.562

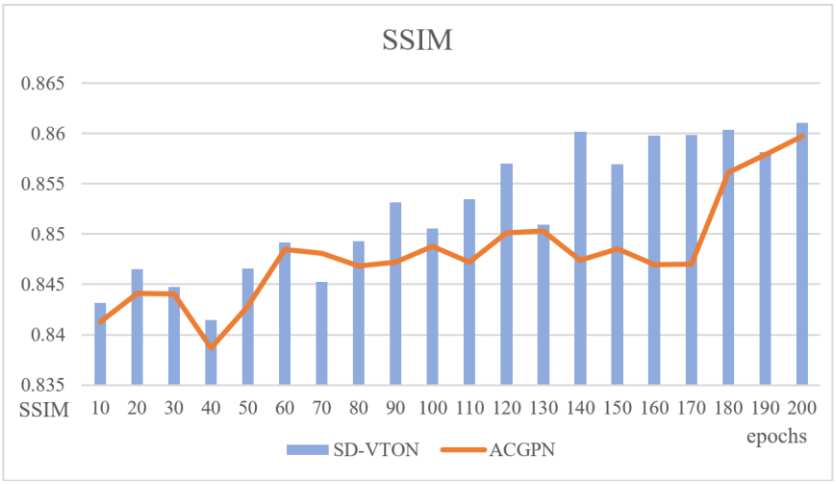

圖 14 ACGPN 與 SD-VTON 的 SSIM 結果比較

(二)定性分析 (Qualitative Analysis)

於定量分析結果達到超越 ACGPN 結構相似性之後，我們將其產出之試穿合成圖做更精細的分析，觀察合成結果分別在哪個步驟出錯。經過 SD-VTON 的改良後，發現於 50 epochs 時，G_1 錯誤比例下降 4%；而試穿成功的比例則增加了 4%，更加驗證了 UNet++ 對整體結果有所改善；而到了 200 epochs 時，G_1 錯誤比例下降了 7%；但試穿成功的比例卻維持與 ACGPN 200 epochs 相同的比例，由此可見即使 epochs 數提高可能讓G_1 的改善幅度增加，但同時也可能造成其他神經網路出現問題，進而影響到整體的結果。

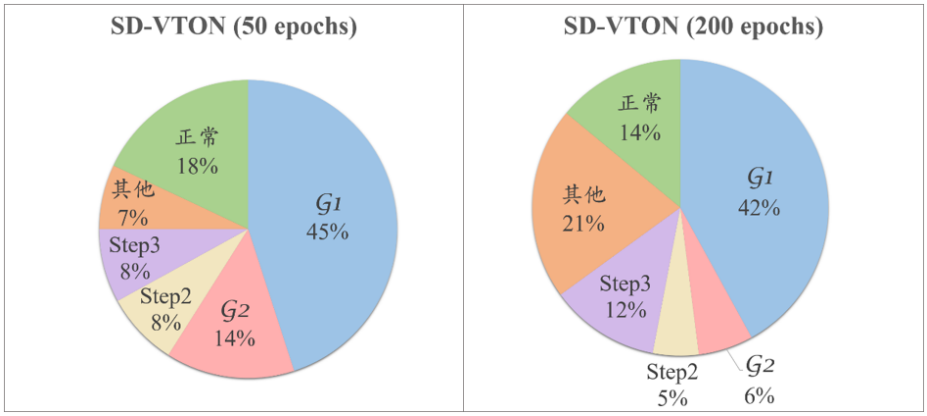

圖 15 SD-VTON 定性分析結果分類情形

5. 結論

本研究提出了一個新穎的 SD-VTON，意旨透過加強語意分割和優化 ACGPN 之試穿網路，同時保留衣服特徵及細節，又不失其人像原始的姿勢、身體部位、下著，生成逼真的試穿合成結果。SD-VTON 新增了一個階段到 ACGPN 模組裡，作為輸入資料的前處理，CT 可從模特兒試穿圖中剪取目標上衣作為輸入，增加試穿系統整體實用性，不再受到目標上衣型態的限制；接著，改善 ACGPN 三階段，分別用來達成不同部分的合成效果，SGM 預測目標上衣試穿區塊，CWM 約束衣服變形使過程不至於過度扭曲，CFM 將試穿後會暴露出來的人體部位正確填補，並產生最終合成結果。最後，分析現有虛擬試穿網路的重現結果，VITON、CP-VTON 由簡到精的合成方法，容易造成試穿效果模糊、遮蓋人體部位等難以預測的試穿錯誤，而 ACGPN 的架構有效對抗前者大部

分的問題，但重現結果大多在 SGM 的G_1發生錯誤，導致最終試穿區塊也跟著有誤，因此，SD-VTON 研究如何改善預測試穿的人物語意模板，進行試穿網路的優化，將 G_1 生成器改為 UNet++，並同樣對生成結果做定性分析與定量分析，結果顯示皆優於 ACGPN，200 epochs 達到最佳 SSIM 值 0.861，且於 140 epochs 時就有與其相近且穩定的試穿結果，QI 持續高於 0.5，證明其結果優越性。

雖然 SD-VTON 已改善輸入彈性及試穿效果，但仍有許多會影響試穿結果的資料限制：測試時人物肢體應盡量不遮蔽到軀幹、拍攝背景單純為佳，推測是由於訓練集過於單一所導致，因此，未來能以 3 個方法來改進試穿系統：1.於訓練時加入更多樣的資料，或許能適用於更多種不同條件的圖像。2.網路架構不斷推陳出新，使用新架構進行訓練。3.擴展下半身試穿，並結合穿搭推薦功能，使整個系統更能達成實用性。

6. 參考文獻

[1] H.Yang, R.Zhang, X.Guo, W.Liu, W.Zuo, andP.Luo, “Towards photo-realistic virtual try-on by adaptively generating↔preserving image content,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 7847–7856, 2020, doi: 10.1109/CVPR42600.2020.00787.

[2] Z.Zhou, M. M. R.Siddiquee, N.Tajbakhsh, andJ.Liang, “UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation,” *IEEE Trans. Med. Imaging*, 2020, doi: 10.1109/TMI.2019.2959609.

[3] X.Liang, K.Gong, X.Shen, andL.Lin, “Look into Person: Joint Body Parsing & Pose Estimation Network and a New Benchmark,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019, doi: 10.1109/TPAMI.2018.2820063.

[4] J.Johnson, A.Alahi, andL.Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” 2016, doi: 10.1007/978-3-319-46475-6_43.

[5] P.Isola, J. Y.Zhu, T.Zhou, andA. A.Efros, “Image-to-image translation with conditional adversarial networks,” 2017, doi: 10.1109/CVPR.2017.632.

[6] X.Han, Z.Wu, Z.Wu, R.Yu, andL. S.Davis, “VITON: An Image-Based Virtual Try-on Network,” 2018, doi: 10.1109/CVPR.2018.00787.

[7] B.Wang, H.Zheng, X.Liang, Y.Chen, L.Lin, andM.Yang, “Toward characteristic-preserving image-based virtual try-on network,” 2018, doi: 10.1007/978-3-030-01261-8_36.

[8] S.Belongie, J.Malik, andJ.Puzicha, “Shape matching and object recognition using shape contexts,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002, doi: 10.1109/34.993558.

[9] K.Gong, X.Liang, D.Zhang, X.Shen, andL.Lin, “Look into Person: Self-supervised Structure-sensitive Learning and a new benchmark for human parsing,” 2017, doi: 10.1109/CVPR.2017.715.

[10] D.Roy, S.Santra, andB.Chanda, “LGVTON: A Landmark Guided Approach to Virtual Try-On,” *arXiv*. 2020.

[11] C.Lassner, G.Pons-Moll, andP.V.Gehler, “A Generative Model of People in Clothing,” 2017, doi: 10.1109/ICCV.2017.98.

[12] B.Zhao, X.Wu, Z. Q.Cheng, H.Liu, Z.Jie, andJ.Feng, “Multi-view image generation from a single-view,” 2018, doi: 10.1145/3240508.3240536.

[13] L.Ma, X.Jia, Q.Sun, B.Schiele, T.Tuytelaars, andL.VanGool, “Pose guided person image generation,”

2017.

[14] S.Zhu, S.Fidler, R.Urtasun, D.Lin, andC. C.Loy, “Be Your Own Prada: Fashion Synthesis with Structural Coherence,” 2017, doi: 10.1109/ICCV.2017.186.

[15] Z.Cao, T.Simon, S. E.Wei, andY.Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” 2017, doi: 10.1109/CVPR.2017.143.

[16] M.Jaderberg, K.Simonyan, A.Zisserman, andK.Kavukcuoglu, “Spatial transformer networks,” 2015.

[17] A.Raj, P.Sangkloy, H.Chang, J.Hays, D.Ceylan, andJ.Lu, “SwapNet: Image based garment transfer,” 2018, doi: 10.1007/978-3-030-01258-8_41.

[18] T. C.Wang, M. Y.Liu, J. Y.Zhu, A.Tao, J.Kautz, andB.Catanzaro, “High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs,” 2018, doi: 10.1109/CVPR.2018.00917.

[19] I.Goodfellow, Y.Bengio, andA.Courville, “Deep Learning - An MIT Press book,” *MIT Press*, 2016.

[20] R.Yu, X.Wang, andX.Xie, “VTNFP: An image-based virtual try-on network with body and clothing feature preservation,” 2019, doi: 10.1109/ICCV.2019.01061.