基於可組裝串聯外觀流之虛擬服裝試穿

柯良穎　夏至賢　許良亦　陳麒安
國立宜蘭大學資訊工程學系

摘要

隨著深度學習(deep learning, DL)、電腦視覺(computer vision, CV)，以及物聯網(internet of things, IoT)的快速發展，用於線上服飾購物的虛擬試穿(virtual try-on, VTON)技術也獲得突破性的成果並且開始在市場上普及。然而，現今的 VTON 技術卻仍然難以對複雜的服飾進行變形，並且變形後的服裝普遍存在紋理失真的問題。因此，為解決上述問題，本研究提出一種基於可組裝的串聯外觀流網路(cascade appearance flow assemble network, CAFANet)來進行 VTON。該模型透過串聯全域-局部對齊模組(cascade global-local alignment module, CGLAM)來有效地避免在對齊服裝形狀和服裝紋理上導致的服裝紋理失真的問題。根據實驗結果顯示，本研究提出的模型架構在 VITON 資料庫上具有良好的服裝合成能力，並且在測試集上達到 25.34 的 FID 數值。

關鍵詞：深度學習、對抗式生成網路、虛擬試穿、外觀流。

1. 前言

隨著資訊與通訊技術(information and communications technology, ICT)和物聯網(internet of things, IoT)的蓬勃發展，線上服飾購物(online apparel shopping)已逐漸轉變成當代消費者的一種新興購物模式，並且市場迎來大幅度地增長[1]。然而，與傳統的服飾購物相比，線上服務購物難以在購買時進行服飾試穿和檢查服飾狀態，使得許多消費者在購買服飾上仍較傾向於傳統的服飾購物。因此，發展一種讓消費者能在線上進行服裝試穿的虛擬試穿(virtual try-on, VTON)技術，已成為推動線上服飾購物發展的核心，這使得電子商務(electronic commerce, E-Commerce)的業者能夠減少產品退貨產生的運輸成本。為實現上述提到的目標，近年來有許多文獻透過深度學習(deep learning, DL)技術結合電腦視覺(computer vision, CV)技術來實現 VTON 技術，並且透過大量的研究使得 VTON 技術獲得突破性的成果。其中，VTON 技術的研究問題主要能夠分成兩個子問題(sub-problems)，分別是將目標服飾變形(warping)至特定人物的姿態，以及將變形後的服飾與特定人物進行合成(synthesis)。

2. 文獻回顧

在過往的研究中已經有大量的研究論文提出方法來對妝容轉換的問題進行解決。因此，本研究將在下方的內容對過往的妝容轉換研究論文進行描述。Han *et al.*[2]提出一種基於影像的兩階段虛擬試穿網路(virtual try-on network, VITON)來實現 VTON。首先，該方法會先編碼器-解碼器架構(encoder-decoder architecture)來生成出變形服裝的粗略(coarse)遮罩。接著，藉由形狀匹配演算法(shape context matching algorithm) [3]、U-Net 模型[4]，以及外觀合併策略(appearance merging strategy)來將變形服裝合成至目標人物身上。然而，該方法僅採用自由度(degrees of freedom)有限的薄板樣條(thin plate spine, TPS)變換來進行服裝變形，這使得該方法在對具有較大幾何變化的影像進行轉換的時候容易生成出不自然的結果[5]。接著，Sun *et al.* [6]提出一種基於最佳特徵線性分配(optimal features linear assignment)的外觀流(appearance flow)

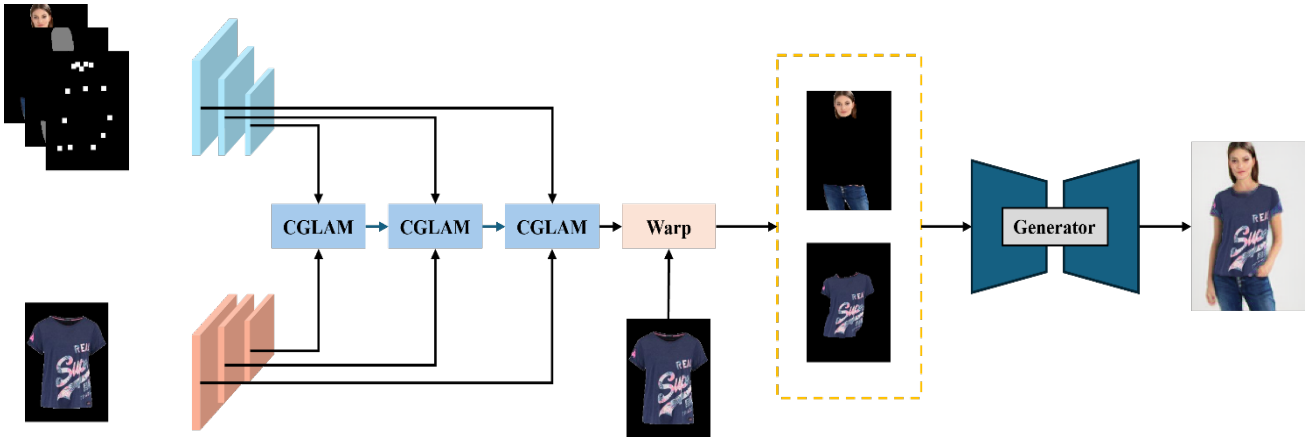
估計網路(Warping-Flow)來實現 VTON。首先，該方法會採用兩個特徵金字塔網路(feature pyramid network, FPN)[7]來分別對目標人體姿態(target person pose)和店內服裝(in-shop garment)進行提取特徵。接著，該方法透過局部特徵聚合模組(local context feature aggregation module, LCFA)來融合不同尺度的特徵，並且用其來準確地估計外觀流。然而，該方法在外觀流估計上並沒有將對齊目標服裝形狀和局部服裝紋理進行區分，這使得服裝在變形時容易產生不必要的偽影(undesired artifacts)[8]。

3. 研究方法

為解決上述提到的問題，本研究提出一種基於可組裝的串聯外觀流網路(cascade appearance flow assemble network, CAFANet)來進行 VTON。以下的內容將對本研究提出的模型架構進行更加詳細的說明

3.1 整體模型

為使服裝能夠有效地變形，因此本研究藉由串聯全域-局部對齊模組(cascade global-local alignment module, CGLAM)來建構出一種基於全域-局部外觀流估計的 CAFANet 模型來實現 VTON，如圖一所示。首先，該方法會透過 FPN 模型對目標人物姿態和店內服裝影像分別進行提取不同尺度的特徵。接著，CAFANet 會透過 CGLAM 對先提藉由 FPN 模型提取的不同尺度特徵來分別對服裝形狀和服裝紋理之外觀流進行估計，以此來避免模型估計的外觀流使得服裝紋理產生扭曲變形。最後，本研究會透過 U-Net 模型架構來對經由外觀流變形的服裝和服裝無關的人物影像(clothing-agnostic person representation)進行服裝合成，以此來實現 VTON。

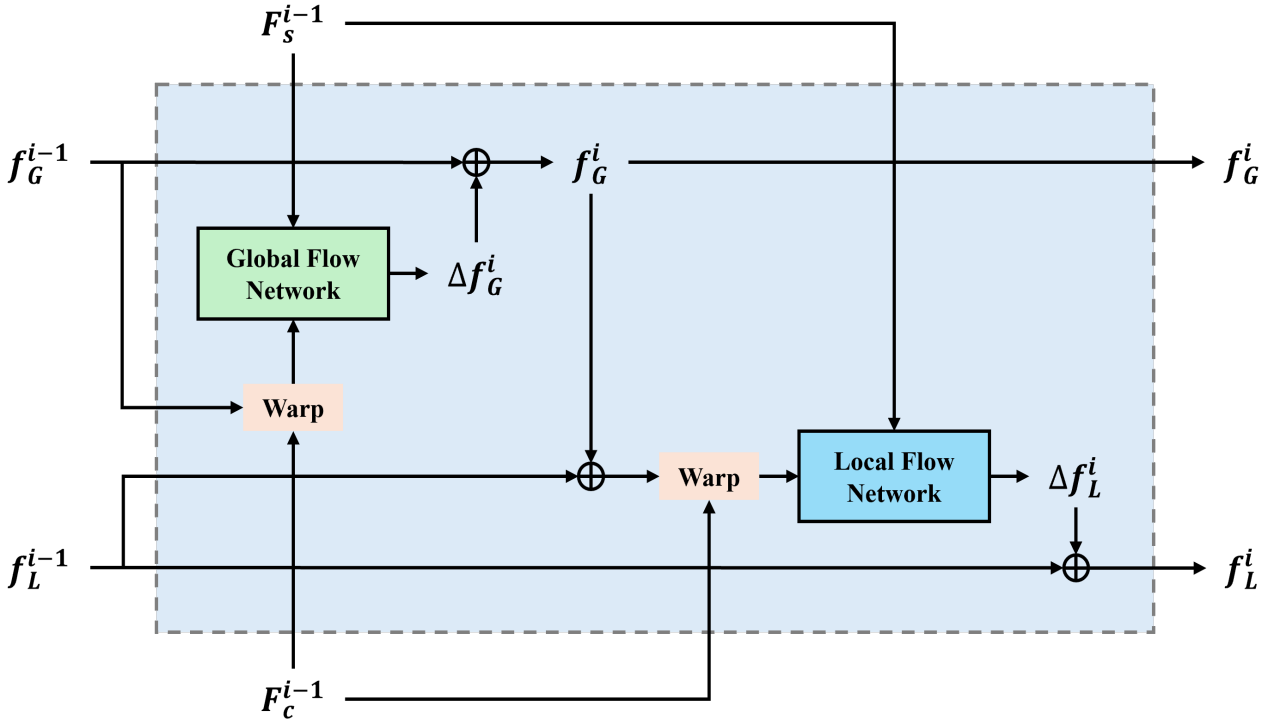


圖一、本研究所提出的 CAFANet 模型

3.2 串聯全域-局部對齊模組(Cascade Global-Local Alignment Module, CGLAM)

為使得本研究提出的 VTON 模型能夠有效地估計外觀流，並且同時避免在對齊服裝形狀和服裝紋理上導致的服裝紋理失真的問題。因此，本研究提出一種能夠對服裝形狀和服裝紋理分別進行估計外觀流的 CGLAM 模組，如圖二所示。首先，CGLAM 模組會將 FPN 網路提取出的人物姿態特徵F_s和店內服裝特徵F_c輸入至全域外觀流網路(global flow network, GFN)和局部外觀流網路(local flow network, LFN)進行服裝形狀和服裝紋理之外觀流估計。其中，CGLAM 會透過先前服裝形狀之外觀流f_G^{i-1}對店內服裝特徵F_c^{i-1}進行變形，並將其與人物姿態特徵F_s^{i-1}輸入至 GFN，以此來估計服裝形狀之外觀流變量Δf_G。接著，本研究會使用先前服裝紋理之外觀流f_L^{i-1}對經由 GFN 微調後的服裝形狀之外觀流f_G^i再次進行微調，以此來使模型能

夠對服裝的形狀和紋理進行準確地估計。最終，CGLAM 模組會藉由微調後的外觀流進行變形的店內服裝特徵F_c^{i-1}與人物姿態特徵F_s^{i-1}輸入至 LFN 來估計服裝紋理之外觀流變量Δf_L，並微調先前服裝紋理之外觀流f_L^{i-1}來獲取該階段的服裝紋理之外觀流f_L^i。



圖二、本研究所提出的 CGLAM 模組

4. 實驗結果與分析

4.1. 實驗環境及模型參數設置

為評估 CAFANet 模型的影像合成能力，因此本研究會透過使用 VITON 資料庫[2]來實現 VTON。該資料庫總共具有 16,235 張解析度(resolution)為 256 x192 的正面人物影像，以及相應的上衣服裝影像。然而，為與近年來的方法進行比較，本研究會遵循過往的實驗設置來將資料庫劃分成具有 14,221 個影像對(image pairs)的訓練集，以及具有 2,032 個影象對的測試集來進行模型訓練。在模型訓練上，本研究將採用 AdamW[9]作為模型的優化器(optimizer)，並且同時將其學習率(learning rate)設置為 0.00005 來進行 VTON 模型的訓練。其中，AdamW 優化器的超參數β_1和β_2分別設置為 0.5 和 0.999。然而，在其餘的超參數設置上，本研究將批次量(batch size)設置為 8、影像尺寸(image size)設置為 256 x192，以及循環次數(epochs)設置為 50。此外，本研究所提出的模型架構均採用 PyTorch DL 框架，並且於 Intel® Core™ i7-12700 CPU 和 Nvidia RTX 3090 圖形處理器(graphics processing unit, GPU)的硬體環境上進行訓練。

4.2. 實驗結果

在影像合成能力的評估上，本研究會採用 Frechet Inception Distance (FID)指標來進行評估。該指標透過計算生成影像和原始影像在特徵空間中的距離來評估兩者間的影像品質是否相似。當 FID 的數值越小時，則兩者特徵間的距離越短，並且相似度高。反之，當 FID 的數值越大時，則兩者特徵間的距離越遠，並且相似度越低。根據實驗結果顯示，本研究提出的 CAFANet 模型架構在 VITON 資料庫上的 FID 指標僅具有 25.34，如表一所示。與過往較佳的

VTON 研究相比，本研究提出的 CAFANet 模型架構在 FID 指標上比 Warping-Flow 的 FID 數值低 3.25。由此可知，對服裝形狀和服裝紋理之外觀流分別進行估計不但能夠有效地提升 CAFANet 模型的服裝變形能力，並且還能同時增強其服裝合成能力。

表一、與現有的虛擬試穿模型在 VITON 資料集上進行 FID 的比較

Methods	FID (↓)
VITON (2018) [2]	55.71
Clothflow (2019) [5]	37.54
Flow-Style (2022) [10]	30.17
Warping-Flow (2024) [6]	28.59
CAFANet (This work)	**25.34**

5. 結論

本研究提出一種能夠對服裝形狀和服裝紋理之外觀流分別進行估計的 CGLAM 模組，透過 CGLAM 模組不僅能夠避免模型產生出過度變形服裝紋理的外觀流，還能夠同時增強模型在服裝合成上的能力。因此，本研究基於該模組進一步提出 CAFANet 模型架構來進行 VTON，並且藉由 VITON 資料庫來對該模型架構的影像合成能力進行有效地評估。根據實驗結果顯示，本研究所提出的 CAFANet 模型與過往的 VTON 研究相比具有較佳的 FID 數值，並且 FID 數值更是比過往較佳研究低 3.25，這使得本研究所提出的 VTON 模型在應用至現今的線上服飾購物服務上具有極大的優勢。

6. 參考文獻

[1] W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, and J. Liu, "Fashion meets computer vision: a survey," *ACM Computing Surveys*, vol. 54, no. 4, pp. 1-41, 2021.

[2] X. Han, Z. Wu, Z. Wu, R. Yu, and L.S. Davis, "VITON: an image-based virtual try-on network," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7543-7552.

[3] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509-522, 2002.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234-241.

[5] X. Han, X. Hu, W. Huang, and M.R. Scott, "ClothFlow: a flow-based model for clothed person generation," *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10471-10480.

[6] K. Sun, J. Tao, P. Zhang, and J. Zhang, "Appearance flow estimation for online virtual clothing warping via optimal feature linear assignment," *Image and Vision Computing*, vol. 142, pp. 104899, 2024.

[7] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie, "Feature pyramid networks for object detection," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117-2125.

[8] H. Rawal, M. J. Ahmad, and F. Zaman, "GC-VTON: predicting globally consistent and occlusion aware local flows with neighborhood integrity preservation for virtual try-on," *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5251-5260.

[9] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *IEEE International Conference on Learning Representations*, 2019.

[10] S. He, Y.Z. Song, and T. Xiang, "Style-based global appearance flow for virtual try-on," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3470-3479.