前瞻科技與管理　12 卷 2 期，29-49 頁（2024 年 5 月）
Journal of Advanced Technology and Management Vol. 12, No. 2, pp. 29-49 (May, 2024)
DOI:10.6193/JATM.202405_12(2).0003

使用人體姿態遷移技術、臉部姿勢轉換技術以及影像修復技術之舞蹈影片生成系統

鄭旭詠 [1,*]　余執彰 [2]

[1] 國立中央大學資訊工程學系教授
[2] 中原大學資訊工程學系教授

摘 要

本篇論文利用生成對抗網路建立了一個舞蹈影片生成系統，可將單張影像和目標舞蹈影片輸入，使照片中的人物跳舞。系統主要採用人體姿態遷移技術以及臉部姿勢轉換技術，讓電腦生成符合目標姿態的人物影像，並修復由背景切割導致的空洞。此外，系統還採用多尺度區域提取器捕捉身體特徵，並將區域風格損失納入損失函數。在臉部姿勢轉移方面，採用漸進式人臉角度轉換架構，並加入遮罩鑑別器以提高生成影像品質。最後，針對修復過程中的顏色偏移問題，系統使用基於 CIEDE2000 色差公式的感知顏色損失函數進行處理，使修復結果更符合人類視覺感知。

關鍵詞：人工智慧、生成對抗網路、生成模型、姿勢轉移、影像修復

[*] 通訊作者：鄭旭詠
電子郵件：breeze.cheng@gmail.com
（收件日期：2023 年 12 月 4 日；修正日期：2024 年 3 月 27 日；接受日期：2024 年 4 月 1 日）

Journal of Advanced Technology and Management Vol. 12, No. 2, pp. 29-49 (May, 2024)
DOI:10.6193/JATM.202405_12(2).0003

Dance Video Generation System Using Human Pose Transfer, Facial Posture Transfer and Image Inpainting Techniques

Hsu-Yung Cheng[1,*], Chih-Chang Yu[2]

[1]Professor, Department of Computer Science and Information, National Central University
[2]Professor, Department of Information and Computer Engineering, Chung Yuan Christian University

Abstract

This paper builds a dance video generation system using a generative adversarial network, which can input a single image and a target dance video to make the people in the photo dance. The system mainly adopts human posture transfer technique and facial posture transfer technique to allow the computer to generate an image of a person that matches the target poses. Also, image inpainting technique repairs the holes in the picture caused by the characters being cut out from the background. In addition, the system uses a multi-scale region extractor to capture body features and incorporates region style loss into the loss function. For face pose transfer, a progressive face angle transformation framework is adopted and a mask discriminator is added to improve the quality of the generated image. Finally, for the color shift problem during the inpainting process, the system uses the perceptual color loss function based on the CIEDE2000 color difference formula to handle the problem, so that the inpainting results match human visual perception better.

Keywords: artificial intelligence, generative adversarial network, generative model, pose transfer, image inpainting

* Corresponding Author: Hsu-Yung Cheng
E-mail: breeze.cheng@gmail.com

RECAST
Ainosco Press

壹、緒論

自從生成對抗網路（Generative Adversarial Network, GAN）出現以來，影像和視訊生成已成為一個熱門的研究主題（Goodfellow et al., 2014）。近年來，GAN 得到了深入的探索。該領域值得注意的研究之一是在條件設定中應用 GAN，透過對輸入影像進行條件處理並產生相應的輸出影像來執行影像到影像的轉換（Isola, Zhu, Zhou, and Efros, 2017; Lassner, Pons-Moll, and Gehler, 2017; Lin, Xia, Qin, Chen, and Liu, 2018）。此類技術可用於產生草圖到圖像或照片到繪畫的轉換。Lin et al.（2018）提出了使用不成對資料和對偶學習（Dual Learning）來細粒度（Fine Granularity）控制生成影像的想法。基於當前多媒體處理領域的需求，我們面臨著如何有效結合人物的身體姿勢、臉部表情和背景影像，以生成高品質舞蹈影片的挑戰。在這過程中，我們需要解決多個問題，包括資料集的選擇與平衡、技術上的挑戰如人臉姿勢的漸進式轉換，以及如何評估生成影片的品質等。同時，我們也探討了這項技術的應用範圍，例如是否適用於虛擬現實等其他領域，這些問題不僅指導著我們研究的動機和目標，同時也為未來相關領域的研究提供了方向和啟示。

本篇論文的目標是建立一個影片生成系統，使用者輸入單張影像和目標的舞蹈影片，即可讓照片中的人物做出舞蹈動作。在此生成系統中主要提出三個技術：第一個是人體姿勢轉移（Pose Transfer）技術，藉由人物影像與目標姿態，讓電腦自動生成出符合目標姿態的人物影像；第二個則是臉部姿勢轉換技術，以解決臉部五官輪廓模糊以及不自然的問題，避免整個舞蹈過程中若臉部影像總是面對同一個角度會造成在跳舞時有生硬不流暢的感覺；第三個是影像修補技術，在將畫面中的人物從背景切割出來並產生不同動作之後，背景有部分會產生出空洞的情況，因此需要使用影像修復（Image Inpainting）技術讓人物的背景自動修補成原本的背景以補足原本人物位置的空洞部分。最後再將生成的跳舞姿勢人物影像放置於修補好的背景影像之上，即可讓輸入的來源影像上的人物在原本的照片背景影像上做出跳舞的動作，並可將連續的舞蹈動作影像製作成影片。

基於上述實驗結果，我們成功地開發了一個用於舞蹈影片生成的系統框架。首先，我們在舞蹈身體姿勢轉換生成模型實驗中，提出了一個具有多尺度區域提取器（Multiple Scale Region Extractor）和姿勢注意力轉移網路（Pose Attentional Transfer Network）的系統，能夠在不同資料集上有效生成高品質的舞蹈影像。其次，在人臉姿勢轉移生成模型實驗中，我們利用了卡內基美隆大學的 Multi-PIE Face 人臉資料集，成功實現了人臉的漸進式轉換，提高了生成品質。最後，在背景修補生成模型實驗中，我們通過訓練和評估背景影像修復模型，成功地修復了受損的影像部分，這些成果為我們的研究提供了良好的基礎，使我們能夠在舞蹈影片生成方面取得令人滿意的結果。

貳、文獻探討

一、姿勢轉移

姿勢轉移是從影像翻譯延伸出來的研究領域，其目標是在給定來源影像的情況下產生指定目標姿勢的影像，它可以應用於所需人體姿勢的影片產生和人員重新識別的資料增強（Ma, Jia, Sun, Schiele, Tuytelaars, and Van Gool, 2017; Ma, Sun, Georgoulis, Van Gool, Schiele, and Fritz, 2018; Neverova, Alp Güler, and Kokkinos, 2018; Siarohin, Sangineto, Lathuilière, and Sebe, 2018; Zhu, Huang, Shi, Yu, Wang, and Bai, 2019）。由於來源主體和目標姿勢之間的變形，姿勢轉移的任務更具挑戰性。此外，不同視角下不同姿勢的人類外觀也存在顯著差異。

Ma et al.（2017）首先嘗試透過將問題分為兩個階段來處理姿勢轉換問題，第一階段將學習粗略的整體人體結構，第二階段將產生精細的外觀細節。解開的人物影像生成也提出了一個兩階段重建流程來處理這個問題，多分支重建網路學習前景、背景和姿態因素的解纏表示（Disentangled Representation），並將這三個因素編碼為嵌入特徵。然後，以對抗方式學習三個對應的映射函數，將高斯雜訊映射到每個因子的學習嵌入特徵空間。可變形 GAN（Siarohin et al., 2018）在 GAN 的生成器中引入了可變形跳躍連接。此外，還提出了最近鄰損失來取代傳統的 L1 和 L2 損失，以將生成影像的細節與目標影像進行匹配。Zhu et al.（2019）提出了一個基於姿勢注意力轉移區塊的框架，用於在到達目標之前通過一系列中間姿勢表示來轉移條件姿勢。該方案透過允許每個傳輸轉換區塊執行局部傳輸轉換，來避免捕獲全局等級的複雜結構（Complex Structure of the Manifold on the Global Level）的挑戰。注意力機制逐步引導可變形轉移過程，以同時優化外觀和姿勢。與先前在 DeepFashion（Liu, Luo, Qiu, Wang, and Tang, 2016） 和 Market-1501（Zheng, Shen, Tian, Wang, Wang, and Tian, 2015）資料集上的工作相比，該設計在定性和定量上都表現出了卓越的性能。然而，當目標姿勢的身體運動較大時，他們的系統無法產生令人滿意的結果。

二、臉部姿勢轉換

關於人臉角度的轉換方面的研究，這是一個具有挑戰性的生成學習問題，兩張人臉影像之間的角度差異越大，要生成準確的影像就越困難。近年來有一些學者提出了使用深度學習的方法來解決，早期的研究都是將人臉轉正，最具代表性的研究為 TP-GAN（Huang, Zhang, Li, and He, 2017），而近來發展的方法可以做到任意角度的旋轉，比如完整表示生成對抗網路（Complete Representation Generative Adversarial Network）（Tian, Peng, Zhao, Zhang, and Metaxas, 2018）和配對代理人姿勢引導的生成對抗網路（Couple-Agent Pose-Guided Generative Adversarial Network）（Hu, Wu, Yu, He, and Sun, 2018）。本論文的目標是希望讓人臉轉換至準確的目標角度，且生成出來的影像可以有較好的品質。

在姿勢引導的真實感臉部旋轉（Pose-Guided Photorealistic Face Rotation）論文中，提出了配對代理人姿勢引導 GAN 的方法（Hu et al., 2018），它不只能將人臉轉換成任意角度，其中提出的配對代理人鑑別器（Couple-Agent Discriminator），可以有效的結合臉部的姿勢

和區域結構（Local Structure），加強了生成影像的真實感，這使得配對代理人姿勢引導的 GAN 生成出來的任意角度人臉影像都具有真實感並且能夠保留了臉部的特徵。配對代理人姿勢引導的 GAN 架構是由兩個部分組成，分別為姿勢引導生成器（Pose-Guided Generator）和配對代理人鑑別器。在姿勢引導生成器的部分，為了讓人臉影像能夠轉換成任意角度，生成器的輸入需要包含姿勢資訊，而自適應的生成器可以從姿態嵌入表示（Pose Embedding）中學習所需的訊息，也就是說旋轉角度是藉由模型學習而來的，而非事先給定的資訊。姿態嵌入表示是透過臉部標誌偵測器（Facial Landmark Detector）和一系列的影像處理得到的，其使用的臉部標誌偵測器是 LightCNN（Convolutional Neural Networks）（Wu, He, Sun, and Tan, 2018），藉由這個偵測器，可以取得五個臉部標誌的座標點，分別為左右眼、左右嘴角以及鼻子，然後將這五個座標點轉換成熱圖（Heatmaps），每個熱圖上的臉部標誌座標點，都有使用標準差為 2 的高斯分布做處理，獲得的影像處理結果就稱為姿態嵌入表示。因為 U-Net 架構在影像至影像（Image-to-Image）轉換方面的成功，此生成模型由一個下採樣編碼器（Down-Sampling Encoder）與一個帶有長跳躍連接（Skip Connections）的上採樣解碼器（Up-Sampling Decoder）組成，用於多尺度特徵融合。這種架構保留了上下文和紋理訊息，這對於除去人為的偽影（Artifact）和填充紋理有很大的幫助。此系統生成模型的輸入為原始影像 I^a、原始影像的姿態嵌入表示 P^a 和目標影像的姿態嵌入表示 P^b 串連（Concatenate）起來的內容，經過生成模型產生出生成影像 $\hat{I}^b$，在姿態嵌入表示的引導下，生成器可以合成出指定角度的人臉影像。

三、背景修補

修補原本照片中人物所在的背景資訊，需要用到影像修復的技術。影像修復是電腦視覺中的一項具有挑戰性的任務，它可以用於恢復影像中損壞或損壞的區域。先前的研究（Ballester, Bertalmio, Caselles, Sapiro, and Verdera, 2001; Bertalmio, Sapiro, Caselles, and Ballester, 2000; Bertalmio, Vese, Sapiro, and Osher, 2003）都是從基於擴散的方法開始的，但它們僅適用於具有小損壞區域的圖像。後來提出了許多基於樣本的方法（Barnes, Shechtman, Finkelstein, and Goldman, 2009; Barnes, Shechtman, Goldman, and Finkelstein, 2010; Chen et al., 2016），透過分析影像的背景區域來獲得合適的補丁來填充損壞的區域。PatchMatch（Barnes et al., 2009, 2010）是基於範例的方法中最先進的方法，著名的影像編輯軟體 Adobe Photoshop 使用它來提供其上下文感知填充功能。當損壞區域與背景區域相似或具有重複的結構特徵時，此類方法可以獲得良好的結果。這些傳統方法的瓶頸在於只能應用於損壞區域較小的影像，無法重建具有語意的複雜結構，且演算法時間複雜度較高。

隨著人工智慧的蓬勃發展，最近的研究（Iizuka, Simo-Serra, and Ishikawa, 2017; Li et al., 2020; Li, Wang, Cheng, Wu, Gan, and Fang, 2019; Liu, Reda, Shih, Wang, Tao, and Catanzaro, 2018; Wang, Zhang, Niu, Ling, Yang, and Zhang, 2021; Wang, Zhang, and Zhang, 2021; Yi, Tang, Azizi, Jang, and Xu, 2020; Yu et al., 2020; Yu, Lin, Yang, Shen, Lu, and Huang, 2018, 2019; Zheng, Cham, and Cai, 2021）發現基於深度學習的方法可以在影像修復方面取得更好的效果。與基於擴散和基於樣本的方法等傳統方法相比，基於深度學習的方法利用卷積神經網路（Convolutional Neural Networks, CNN）從大量數據中學習到的特徵來重建更有意義、更

合理的結構和紋理特徵。Iizuka et al.（2017）提出使用兩個判別器分別監視局部特徵和全局特徵，以確保修復影像保持全局和局部一致性。上下文注意力（Yu et al., 2018）首先使用兩階段產生網路產生粗略結果，然後使用上下文注意機制，利用背景區域的紋理特徵產生最終結果。部分卷積（Partial Convolution）（Liu et al., 2018）解決了正常像素和損壞像素以相同方式處理的問題，並提出了更新遮罩（Updating Mask）的機制。門控卷積（Gated Convolution）（Yu et al., 2019）進一步利用部分卷積（Liu et al., 2018）的概念，提出了動態遮罩更新機制，以解決遮罩在網路深層消失的問題。Li et al.（2020）提出了互動式分離網絡，將特徵逐步分解為兩個流（Streams）並將它們融合，從而減少對影像解析度成分（High-Resolution Component）的侵蝕並保持語義表示。

參、實驗方法

一、基於漸進式框架、多尺度區域提取器與可學習區域正規化之舞蹈姿勢轉換

在本研究中，我們提出了一個舞蹈姿勢傳輸系統，它改進了 Zhu et al.（2019）的漸進框架，所提出的舞蹈姿勢轉換系統具有處理較大身體動作的能力，可用於生成舞蹈影片。我們提出了多尺度區域提取器，以根據身體關鍵點更好地捕獲每個身體區域的特徵。此外，可學習區域正規化（Learnable Region Normalization, RN-L）（Yu et al., 2020）被整合到所提出的框架中，以提高生成結果的品質。透過所提出的設計，尤其是當上肢和下肢的運動和角度較大時，生成的影像得到顯著改善。

本論文所提出的舞蹈姿勢轉換架構圖系統框架如圖 1 所示，目標是基於單一來源影像產生指定目標姿勢的舞蹈影片。在我們的實作中，應用 OpenPose（Cao, Simon, Wei, and Sheikh, 2017）從來源影像和目標舞蹈影片的每一幀中提取人體的 18 個關節，提取的來源姿態 P_S 和目標姿態 P_T 的關鍵點用作生成系統的輸入。編碼器使用來源影像 I_S 產生來源影像的初始特徵圖，表示為 F_0^I。來源姿態 P_S 和目標姿態 P_T 連接在一起，並作為另一個解碼器的輸入來產生初始特徵圖姿勢特徵 F_0^P。姿勢注意力轉移網路有兩條路徑，影像路徑和姿勢路徑。對於姿勢注意力轉移網路中的每個區塊，兩條路徑分別將 F_{t-1}^I 更新為 F_t^I，將 F_{t-1}^P 更新為 F_t^P。在輸出端，最終的影像特徵 F_T^I 用於解碼器獲得輸出影像 I_T，並丟棄最終的姿態表示 F_T^P。姿勢注意力轉移網路的詳細資訊可在 Zhu et al.（2019）的研究中找到。

具有長跳躍連接的 U-net 架構被廣泛用於減少層間資料傳遞過程中的空間資訊損失並穩定訓練和收斂（Ronneberger, Fischer, and Brox, 2015）。然而，姿勢轉移後，來源影像和特徵圖中的區域並不對應於解碼器對應層中的相同像素，直接透過長跳躍連接連接特徵圖在姿態轉移任務中是不合適的，因此我們提出了一個多尺度區域提取器來解決這個問題。根據人體 18 個關節，多尺度區域提取器定義了頭部、軀幹和四肢 10 個區域，對於來源姿態中的每個區域，從編碼器的所有中間層 F_{all}^I 中提取相應的多尺度特徵表示。我們使用區域的邊界框來裁剪已經過多次子採樣的卷積特徵圖，為了對齊特徵圖上的邊界框，矩形的角點被投影到

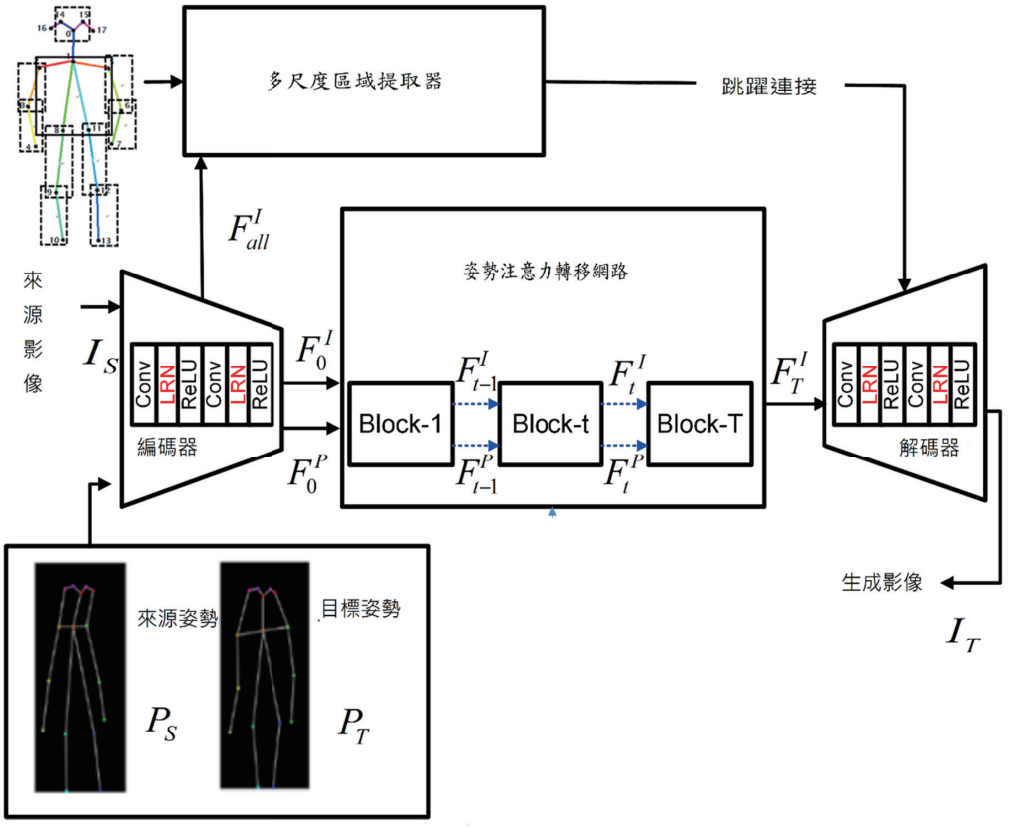


圖 1　舞蹈姿勢轉換架構圖

資料來源：自行繪製。

特徵圖中的像素上，使得圖像域中的該角點最接近該特徵圖像素的感受野（Receptive Field）的中心。根據 He, Zhang, Ren, and Sun（2014）計算特定區域的感受野和對應特徵圖的對應，然後，使用雙線性取樣器調整擷取的多尺度特徵的大小，使其尺寸與特徵圖中目標姿態的相應區域相符。特徵區域遮罩中目標姿態的 10 個區域被相應調整大小的多尺度特徵表示替換，以形成部分特徵遮罩，產生的部分特徵遮罩被連接到解碼器中對應的卷積層。透過這種方式，即使身體形狀變形，每個身體區域的細粒度細節也能在姿勢轉移過程中得以保留。

Yu et al.（2020）提出了區域正規化（Region Normalization）的概念來改善影像修復的結果，為了取代批量正規化（Batch Normalization）或實例正規化（Instance Normalization, IN），區域正規化根據輸入遮罩將空間像素劃分為不同的區域，並計算每個區域的均值和方差以進行正規化。這樣，標準化就不會受到損壞區域的影響，因此可以避免不期望的平均值和變異數偏移。RN-L 的目的是自動偵測損壞的區域並獲得區域遮罩來執行區域正規化。我們應用 RN-L 的概念來學習生成影像中需要改進的區域作為損壞區域。損壞區域和未損壞區域的歸一化是分別執行的，並使用全域仿射變換來增強融合。RN-L 中使用特徵的空間關係產生空間響應圖。沿著通道軸進行最大池化和平均池化，分別獲得特徵描述子 F_{max} 和 F_{avg}。然後，將兩個池化結果連接起來。RN-L 以 Sigmoid 活化函數 σ 在兩個映射上進行卷積，得到空間響應映射 M_{sr}，如下式所示。

$$M_{sr} = \sigma\big(Conv([F_{max}, F_{avg}])\big) \tag{1}$$

為空間響應圖設定閾值 t 以獲得區域遮罩 M。基於遮罩 M，區域正規化對輸入特徵 F 進行正規化，然後執行像素級仿射變換（Affine Transformation）。透過對空間響應圖 M_{sr} 進行卷積得到仿射參數 γ 和 β。

$$\gamma = Conv(M_{sr}),\ \beta = Conv(M_{sr}) \tag{2}$$

在仿射變換中，γ 和 β 的值沿著通道維度擴展，全域空間資訊可以在空間響應圖 M_{sr} 中找到，對它進行卷積可以學習全局表示，有助於損壞和未損壞區域的融合。

如下式所示，完整損失函數由對抗性損失 L_{GAN}、像素級 $L1$ 損失 L_{L1}、感知損失 L_{pec} 和區域風格損失 L_{RStyle} 組成。

$$L_{full} = arg\min_{G}\max_{D} L_{GAN} + L_{L1} + L_{pec} + L_{RStyle} \tag{3}$$

L_{GAN} 項如下式所示，其中 I_G 表示真實影像，兩個判別器 D_A 和 D_S 分別是外觀判別器和姿勢結構判別器。

$$\begin{aligned} L_{GAN} = {} & E_{I_S,\, I_G,\, P_T \sim Real\ Data}\Big\{\log\big[D_A(I_S, I_G)\cdot D_S(P_T, I_G)\big]\Big\} \\ & + E_{I_T \sim Fake\ Data;\ I_S,\, P_T \sim Real\ Data}\Big\{\log\big[\big(1 - D_A(I_S, I_T)\big)\cdot\big(1 - D_S(P_T, I_T)\big)\big]\Big\} \end{aligned} \tag{4}$$

L_{L1} 損失是產生的影像 I_T 和真實影像 I_G 之間的像素級 $L1$ 距離的總和。儘管它往往會在生成的圖像中引入一定程度的模糊副作用，但 L_{L1} 在生成類似真實圖像的過程中仍然發揮著重要作用。感知損失 L_{pec} 旨在減少生成影像的失真，它採用產生的影像 I_T 和真實影像 I_G 的特徵圖之間的 $L2$ 距離的平方。在下式中，φ_j 表示在 ImageNet 上預訓練的 Visual Geometry Group（VGG）模型的第 j 層的輸出，H_j, W_j, C_j 分別表示 φ_j 的空間高度、寬度和深度。

$$L_{pec} = \frac{1}{C_j H_{ij} W_{ij}} \sum_{c=1}^{C_j} \sum_{y=1}^{H_j} \sum_{x=1}^{W_j} \left|\left|\varphi_j(I_G)_{x,y,c} - \varphi_j(I_t)_{x,y,c}\right|\right|_2 \tag{5}$$

除了上述損失項之外，我們還為所提出的系統中的多尺度區域提取器機制新增了區域樣式損失（L_{RStyle}）項目。如下式所示，L_{RStyle} 利用每個區域的格拉姆（Gram）矩陣來計算損失。

$$L_{RStyle} = \frac{1}{10C_jH_{ij}W_{ij}} \sum_{i=1}^{10} \sum_{c=1}^{C_j} \sum_{y=1}^{H_{ij}} \sum_{x=1}^{W_{ij}} \left|\left| G_j(R_i^G)_{x,y,c} - G_j(R_i^T)_{x,y,c} \right|\right|_2 \quad (6)$$

根據 Huang et al.（2017），格拉姆矩陣的定義如下式所示。

$$G_j(A) = \varphi_j(A)^T \varphi_j(A) \quad (7)$$

它傳達了 j 層的風格表示。符號 $R_1^G \sim R_{10}^G$ 表示來自真實影像 I_G 的 10 個身體區域；$R_1^T \sim R_{10}^T$ 表示生成影像 I_T 中的 10 個身體區域。H_{ij} 和 W_{ij} 分別表示 $\varphi_j(R_i)$ 的空間高度和寬度。

二、臉部姿勢轉換

圖 2 為我們所使用之臉部影像角度轉換與加強生成模型。本論文所提出的臉部姿勢轉換生成模型的輸入為原始影像 I^a、原始影像的姿態嵌入表示 P^a 和目標影像的姿態嵌入表示 P^b 串連起來的內容，經過生成模型產生出生成影像 $\hat{I}^b$，在姿態嵌入表示的引導下，生成器可以合成出指定角度的人臉影像。在配對代理人鑑別器的部分，它有兩個鑑別器 Agent 1 跟 Agent 2。Agent 1 是條件判別器，它是基於 CNN 框架實作而成，它把原始影像 I^a 當作條件並跟生成影像 $\hat{I}^b$ 或是目標影像 I^b 組成 pairs 當作輸入，不只可以區分生成影像與真實影像，還可以學習旋轉角度的區別。Agent 2 也是條件判別器，其架構跟 Agent 1 一樣，與 Agent 1 不同的是，它採姿態嵌入表示 P^b 當作條件，並且跟生成影像 $\hat{I}^b$ 或是目標影像 I^b 組成 pairs 當作輸入，這種成對輸入的設計提升了條件判別器區分臉部結構的多樣性和獲取局部感知訊息的能力。本論文提出的改進方法不同之處是將人臉影像和姿態嵌入表示串連之後才進行特徵提取，我們設置了兩個編碼器，一個是專門用來提取人臉影像的特徵，另一個專門用來提

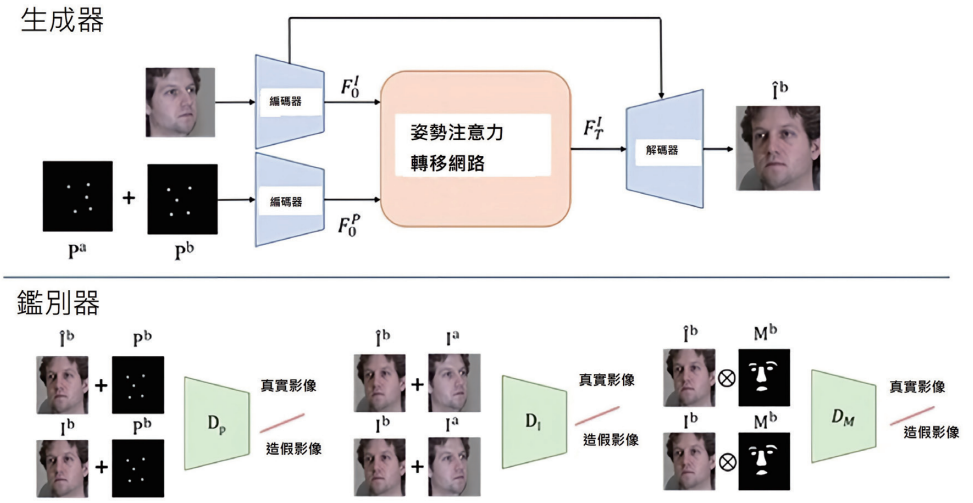


圖 2　臉部姿勢轉換生成模型架構圖

資料來源：自行繪製。

取姿態嵌入表示的特徵，之後再藉由前面提到的姿勢注意力轉移網路進行漸進式人臉轉換，加入此方法不但能讓人臉特徵進行像素塊的遷移，還能透過注意力的機制強調出影像中重要的特徵。判別器的部分，除了使用 Hu et al.（2018）提出的配對代理人鑑別器，還參考了 Yin, Jiang, Robinson, and Fu（2020）與 Cao et al.（2017）提出的臉部注意力機制概念，在此加入遮罩鑑別器用來加強重點區域的生成品質。

應用多工串聯卷積網路（Multi-Task Cascaded Convolutional Networks）（Zhang, Zhang, Li, and Qiao, 2016）作為提取臉部特徵之特徵提取器，這個網路模型可偵測出人臉與相對應特徵的座標位置。它能夠應用於各種複雜條件下的場景，可以偵測出場景中是否有出現人臉，以及人臉上的五個特徵點，分別為左邊眼睛、右邊眼睛、左側嘴角、右側嘴角和鼻子。多工串聯卷積網路模型運作流程包含四個部分，依序為建立影像金字塔、提案網路（Proposal Networks, P-Net）、精煉網路（Refine Networks, R-Net）以及輸出網路（Output Networks, O-Net）。首先要做影像的前處理，透過雙線性插值法產生出不同尺度的影像，然後建立出影像金字塔，並作為後面三個階段的網路模型的輸入。第一個階段是 P-Net，它的架構為全 CNN，主要的功能是盡可能的產生有包含人臉的候選矩形框。取得人臉的候選矩形框還有它的邊界框回歸向量後，可以採用非極大值抑制（Non-Maximum Suppression）來合併高度重疊的人臉候選矩形框，以及可透過預測的邊界框回歸向量，計算它跟真實影像之間的 $L2$ 損失，用來校準人臉候選矩形框的座標。

第二階段是 R-Net，它的架構是單純的 CNN，會先將 P-Net 認為可能包含人臉的邊界框，利用雙線性插值法調整大小，並輸入到 R-Net 中，進一步過濾掉大量錯誤的候選者，它跟 P-Net 一樣，都是藉由計算邊界框回歸向量之間的 $L2$ 損失和非極大值抑制來篩選與校準邊界矩形框的座標。

最後一個階段是 O-Net，它的架構和 R-Net 很相似，但是使用深度更深的網路，在這個階段，目標是更詳細的描述臉部特徵資訊，可以藉由它獲得臉部五個特徵點的座標。首先，會將 R-Net 認為可能包含人臉的邊界框，利用雙線性插值法調整大小並且輸入到 O-Net 中，進行人臉偵測與臉部特徵點的提取，臉部特徵點的定位方法和邊框類似，都是透過計算歐幾里德距離來調整，人臉特徵點座標有五個，分別為左右眼睛、鼻子和左右嘴角。

取得五個人臉特徵點座標後，將這五個人臉特徵點轉換成熱圖，處理方式與配對代理人姿勢引導的 GAN 一樣，每個熱圖上的人臉特徵點座標點，都使用標準差為 2 的高斯分布做處理，最後獲得的影像處理結果就稱為姿態嵌入表示。接著產生鑑別器所需要使用到的遮罩，在遮罩生成的部分是參考 Bilateral Segmentation Network（BiSeNet）（Yu, Wang, Peng, Gao, Yu, and Sang, 2018）的切割模型，他是一個與場景分割有關的網路模型，使用此切割模型再利用 CelebAMask-HQ 資料集（Lee, Liu, Wu, and Luo, 2020）來進行訓練。CelebAMask-HQ 資料集是一個人臉資料集，裡面的每一張人臉影像都帶有 19 個臉部區域的遮罩標記，比如眼睛區域、鼻子區域，以及嘴巴區域。因此採用此資料及訓練之後即可以將人臉切割為多個部分，例如帽子、頭髮、眉毛、眼睛、鼻子、嘴巴和脖子等。

在編碼器的部分，本論文的架構使用了兩個編碼器，這兩個編碼器的架構都是由三層卷積所組成。其中一個編碼器是用來取得人臉影像的特徵，所以它的輸入就是原始的人臉影像，另外一個編碼器是用來取得姿態嵌入表示的特徵，因此它的輸入為原始影像的姿態嵌入表示

和目標影像的姿態嵌入表示。接下來會將獲得的特徵圖輸入到姿勢注意力轉移網路中進行漸進式人臉角度轉換。在解碼器的部分，架構上可以分為三個部分，第一部分是單純的反卷積層，第二部分是反卷積堆疊，是由二個殘差區塊與反卷積層組成的，而第三部分則是單純的卷積層。由於自動編碼器雖然能夠有效的找出資料間的高度相關性，但是不可避免的過程中會遺失部分訊息，此外隨著網路深度的增加，也會導致原始影像保留下來的細節越來越少。為了彌補這個問題，在這裡也採用了 U-Net 架構中跳躍連接的設計，將原始影像與編碼器中各個卷積層的特徵圖接到解碼器的對應層上，透過底層特徵與高層特徵的融合，可以讓解碼器在做反卷積時獲得更多訊息，盡可能的學習原始影像中的細節，生成更清晰的人臉影像。

鑑別模型由三個鑑別器組成，其中兩個鑑別器是參考配對代理人姿勢引導的 GAN 所提出的配對代理人鑑別，另外一個鑑別器是參考論文提出的遮罩鑑別器。三個鑑別器的模型架構皆相同，都是基於 CNN 來實作而成的，最後一層卷積層使用的激活函數為 Sigmoid，而它們彼此間的主要差異在於其輸入的內容不一樣。在配對代理人鑑別中的 DI 鑑別器主要是用來區分生成影像與真實影像，以及學習旋轉角度的區別，它的輸入是生成影像跟原始影像所串連而成，或是目標影像跟原始影像所串連而成。在配對代理人鑑別中的 DP 鑑別器是用來提升區分臉部結構的多樣性和獲取局部感知訊息的能力，它的輸入是生成影像跟目標影像的姿態嵌入表示所串連而成，或是目標影像跟目標影像的姿態嵌入表示所串連而成。而遮罩鑑別器的主要功能是加強人臉重點區域的生成品質，不論是在影像生成領域中，或是人臉識別領域，人臉五官的清晰度都是很重要的，因此我們在研究中選了眼睛、鼻子和嘴巴，作為要加強的區域。因此會使用到前述的遮罩資料，遮罩鑑別器的輸入有兩種，一是輸入生成影像跟遮罩影像乘積的結果，二是輸入目標影像與遮罩影像乘積的結果，也就是說遮罩鑑別器要根據眼睛、鼻子和嘴巴的影像去進行鑑別的任務。

在損失函數的部分，我們採用的損失函數由多尺度逐像素 $L1$ 損失（Multi-Scale Pixel-Wise $L1$ Loss）、條件對抗性損失（Conditional Adversarial Loss）與身分保留損失（Identity Preserving Loss）三個部分所組成。另外，下列方程式中的損失函數權重根據 Yin et al.（2020）所建議的設定，設為 $\lambda_1 = 10, \lambda_2 = 0.1, \lambda_3 = 0.1, \lambda_4 = 0.02$。

$$\min_{\theta_G} \max_{\theta_{ii},\theta_{pe}} L = \lambda_1 L_{pix} + \lambda_2 L_{adv}^{ii} + \lambda_3 L_{adv}^{pe} + \lambda_4 L_{ip} \tag{8}$$

損失函數的第一部分是多尺度逐像素 $L1$ 損失，在生成影像上採用多尺度逐像素 $L1$ 損失可用來約束內容的一致性，多尺度逐像素 $L1$ 損失將目標影像與生成影像之間對應的像素點相減並取絕對值，接著再把它們加總取平均，如下列數學式所示，其中 S 代表各種尺度大小的影像，在這裡我們選用了 32 × 32、64 × 64、128 × 128 像素大小的影像，W_s 和 H_s 分別代表對應大小的影像的寬度與高度，C 則是影像的通道數。多尺度逐像素 $L1$ 損失會讓生成影像變的相對比較平滑，造成一定程度的模糊化，但可以優化加速和重建全域訊息。

$$L_{pix} = \frac{1}{S}\sum_{s=1}^{S} \frac{1}{W_s H_s C} \sum_{w,h,c=1}^{W_s,H_s,C} \left| \hat{I}_{s,w,h,c}^{b} - I_{s,w,h,c}^{b} \right| \tag{9}$$

損失函數的第二部分是條件對抗性損失中的 L_{adv}^{ii} 和 L_{adv}^{pe}，分別是用來保存資訊和重建區域結構資訊的，它們都能提升生成影像視覺上的效果，此外還可以減緩多尺度逐像素 $L1$ 損失造成的平滑模糊化問題，如下列方程式所示。$D_{\theta_{ii}}$ 是用來區分生成影像 $\{\hat{I}^b, I^a\}$ 與真實影像 $\{I^b, I^a\}$ 的鑑別器，而 $D_{\theta_{ii}}$ 是用來區分生成影像 $\{\hat{I}^b, P^b\}$ 與真實影像 $\{I^b, P^b\}$ 的鑑別器。

$$L_{adv}^{ii} = E_{I^b \sim P(I^b)}\left[\log D_{\theta_{ii}}(I^b, I^a)\right] + E_{\hat{I}^b \sim P(\hat{I}^b)}\left[\log\left(1 - D_{\theta_{ii}}(\hat{I}^b, I^a)\right)\right] \quad (10)$$

$$L_{adv}^{pe} = E_{I^b \sim P(I^b)}\left[\log D_{\theta_{pe}}(I^b, I^a)\right] + E_{\hat{I}^b \sim P(\hat{I}^b)}\left[\log\left(1 - D_{\theta_{pe}}(\hat{I}^b, I^a)\right)\right] \quad (11)$$

損失函數的第三部分是身分保留損失，身分保留損失的用途就是希望能保留生成影像的身分資訊，如下式所示。D_{ip} 是特徵提取器，在這裡使用的是 VGG16，它可以提取出生成影像跟目標影像各自的特徵，並希望它們的特徵越接近越好，其中 $D_{ip}^{p}(\cdot)$ 代表最後一層 Pooling Layer 的輸出，$D_{ip}^{fc}(\cdot)$ 代表全連接層的輸出。

$$L_{ip} = \left|\left| D_{iP}^{p}(\hat{I}^b) - D_{ip}^{p}(I^b)\right|\right|_{\mathrm{F}}^{2} + \left|\left| D_{ip}^{fc}(\hat{I}^b) - D_{ip}^{fc}(I^b)\right|\right|_{2}^{2} \quad (12)$$

三、背景修補

本論文中所使用的背景修補模型架構如圖 3 所示，此架構有兩個階段，從粗略到精細執行影像修復。第一階段分為編碼器、殘差區塊和解碼器。如圖 3 所示，編碼器包括三個卷積層，每層卷積運算後，它們都透過基本區域正規化（Region Normalization-Basic, RN-B）進行正規化（Yu et al., 2020），並透過 Rectified Linear Unit（ReLU）函數活化。然後，透過具有相同架構的八個殘差區塊對編碼結果進行進一步處理和修復。每個殘差區塊包含兩個膨脹卷積層，透過它們可以增加卷積的感受野。每層卷積運算後，它們會透過 RN-L 進行正規化，並進行上採樣。在殘差區塊中，第一個卷積層的輸入將連接到第二個卷積層的輸出，這樣的殘差連接可以避免梯度消失的問題。最後，解碼器與編碼器相同，由三個通用卷積層組成，將 256 維解碼為 3 維 RGB 影像。編碼器中的基本區域正規化和殘差塊的 RN-L 可以提高第一階段的模型效能，然而與 Yu et al.（2020）的原始架構相比，我們將解碼器的正規化方法從 RN-L 修改為 IN 以緩解顏色偏移問題，在解碼器中使用 IN 而不是使用 RN-L 可以改善遮罩區域和背景區域的顏色融合。此外，我們添加從編碼器到解碼器的殘差連接以減輕潛在的梯度消失問題。

我們基於 Yu et al.（2019）的第二階段設計了所提出架構的第二階段，並進行了一些修改，可分為三個部分，分別為編碼器的分支一、編碼器的分支二、和解碼器，如圖 3 下半部所示。輸入是遮罩和 3 維 RGB 影像第一階段後的輸出，此遮罩僅在執行上下文注意時使用，它利用背景的紋理資訊來增強最終結果。在第二階段編碼器的分支一中，情境注意力的目標是從高維度特徵圖的背景區域中找到與修復區域相似的補丁，並使用它們來重建修復

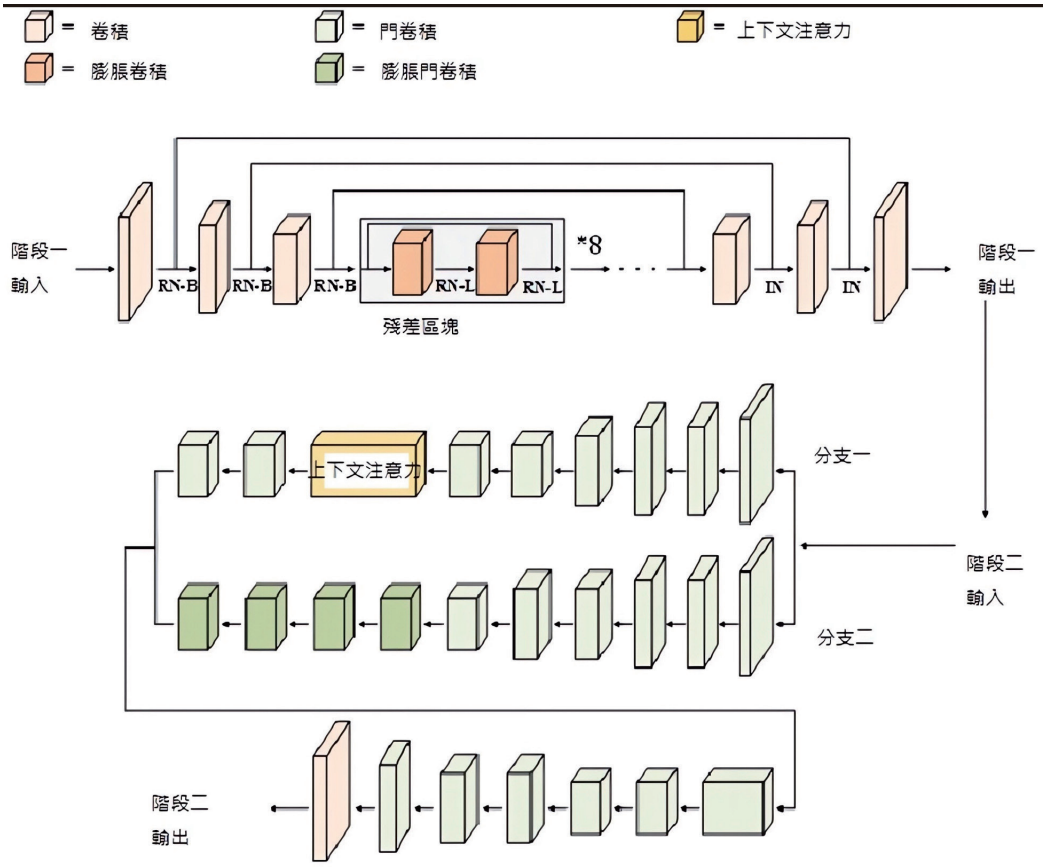


圖 3　臉部姿勢轉換生成模型架構圖

資料來源：自行繪製。

區域的特徵。在第二級編碼器的分支二中，第一級的圖像透過網路再次修復。將第二級編碼器分為兩個分支的原因如下：分支一中上下文注意力重建的輸出由背景區域特徵組成，然而背景區域特徵不一定具有完全正確的紋理，可能只是相對相似，透過分支一重構的輸出不能直接作為最終結果，因此，分支一的輸出與分支二的結果連接起來形成第二級解碼器的輸入。第二階段解碼器執行上取樣以產生網路的最終輸出。在訓練過程中，我們在模型中應用 PatchGAN 判別器和完整損失函數，完整的損失函數包括感知顏色損失 L_{perc_color} 和對抗性損失 L_{GAN}，其可表示為下式所示。其中，p_d 表示真實資料分布，p_z 表示假資料分布。λ_{GAN} 和 λ_{perc_color} 表示調整兩個損失項的權重因子，在感知顏色損失函數中，我們使用基於 CIEDE2000 色差公式的感知顏色損失函數來取代 $L1$ 損失。因為使用經過感知顏色損失訓練的模型可以有效解決顏色偏移問題，使得修復結果的顏色在人類視覺觀察下更加合理。

$$L_{GAN} = E_{x\sim p_d}\log D(x) + E_{z\sim p_z}\log(1 - D(G(z))) \tag{13}$$

$$L_{total} = \lambda_{GAN}L_{GAN} + \lambda_{perc\_color}L_{perc\_color} \tag{14}$$

肆、實驗結果

在此節中，將分別對上述的舞蹈姿勢轉換生成、臉部姿勢轉換生成，以及背景修補方法進行實驗，並呈現出實驗結果。

一、舞蹈身體姿勢轉換生成模型實驗

我們所提出的系統在編碼器和解碼器中使用三個區塊，在姿勢注意力轉移網路中使用五個區塊。編碼器和解碼器中的一個區塊由 Convolution（Conv）+ RN-L + ReLU 組成。使用兩個不同的資料集來訓練和評估所提出的框架。第一個資料集是 DeepFashion（Liu et al., 2016），此資料集包含用於展示服裝商品的圖像。我們使用「店內服裝檢索基準」子集，與其他子集相比，它具有更多種類的服裝類型，每件服裝都有四種不同的視圖，包括正面、側面、背面和全身視圖。在 DeepFashion 資料集中，每位時裝模特兒穿著不同的衣服擺出相同的姿勢，姿勢的多樣性不足，另外，肢體動作也相對較小。因此，我們收集了另一個名為 DancePose 的資料集。DancePose 資料集包含視訊幀，其中包括不同的舞蹈風格，如鎖舞、流行舞、電動舞、街舞、自由式、爵士舞和芭蕾舞，我們使用 DeepLab V3（Chen, Papandreou, Schroff, and Adam, 2017）進行語意分割來定位每個畫面中的人體區域。兩個實驗資料集的大小如下，DeepFashion 資料集內有 48,647 張訓練影像，4,038 張測試影像，DancePose 資料集內有 5,280 張訓練影像，10,987 張測試影像。由於 DancePose 資料集的訓練影像數量相對較少，在 DancePose 資料集上進行訓練和測試時，我們使用 DeepFashion 預先訓練的模型，並使用 DeepFashion 和 DancePose 的混合資料進行微調。圖 4 展示了我們所提出的舞蹈姿勢轉換結果，圖 4（a）為原始影像，圖 4（b）為使用 OpenPose 從原始影像所提取的身體姿勢節點，圖 4（c）為目標姿勢影像，圖 4（d）為使用 OpenPose 從目標影像所提取的目標身體姿勢節點，圖 4（e）為本系統所生成的輸出影像，即讓原始影像中的人物作出目標姿勢的動作。

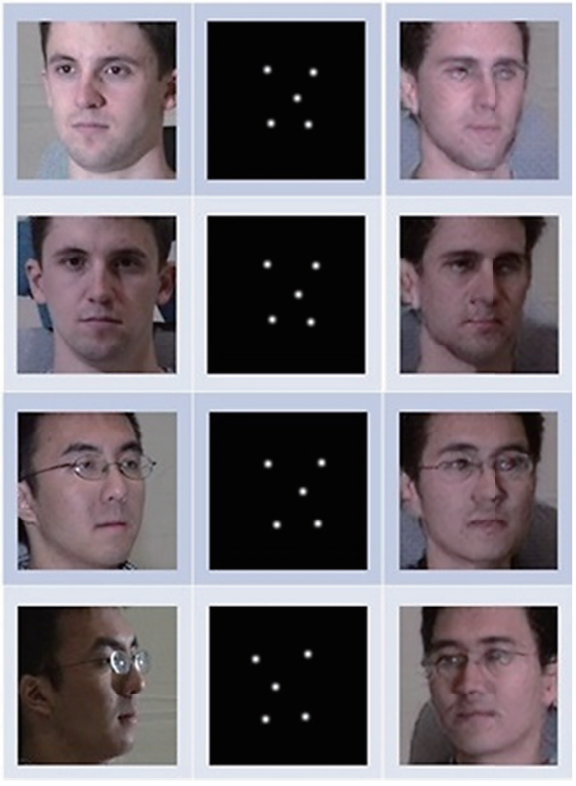
二、人臉姿勢轉移生成模型實驗

在人臉姿勢轉移的實驗中，我們採用由卡內基美隆大學所開放的 Multi-PIE Face 人臉資料集，此資料集內包含超過 750,000 張影像，共有 337 位不同的人，每個人採用 15 種不同的角度拍攝，並配合 19 種光照變化。此資料集拍攝方式為，架設 15 個不同的相機，其中 13 個相機架設於與頭部等高的高度，彼此之間間隔 15 度，另外兩個相機位於頭部上方以模擬監控影像的角度。Multi-PIE Face 資料集是由四個部分組成的，主要差異在於每個部分包含的臉部表情不同，第一個部分的表情包含自然和微笑；第二個部分的表情包含自然、驚訝和瞇眼；第三個部分的表情包含自然、微笑和厭惡；第四個部分的表情包含驚嚇和兩種自然表情。基於我們臉部角度轉換的需求，選擇表情較為常見的第一個部分，作為我們的訓練集，將資料集當中的五分之三的資料量作為訓練資料，而剩下的五分之二當作測試資料，其中訓練樣本包含 53,088 張影像，測試樣本有 16,677 張影像。圖 5 為本論文的人臉姿勢轉移結果，圖 5（a）為原始人臉影像，圖 5（b）為目標人臉姿勢節點，圖 5（c）為本論文之系統所產生的人臉輸出影像，可將圖 5（a）中的人臉轉換成圖 5（b）所指定的姿勢角度。

（a） （b） （c） （d） （e）

圖 4　身體姿勢轉換實驗結果

資料來源：自行繪製。



（a） （b） （c）

圖 5　臉部姿勢轉換實驗結果

資料來源：自行繪製。

三、背景修補生成模型實驗

在背景修補生成模型實驗中，我們在 Places2（Zhou, Lapedriza, Khosla, Oliva, and Torralba, 2018）資料集上訓練和評估我們的背景影像修復模型，使用的遮罩資料集是 NVIDIA 不規則遮罩資料集（Liu et al., 2018），它包含 12,000 個遮罩圖像，每個遮罩圖像覆蓋 0 ~ 60% 的影像區域。圖 6 為我們的背景修補實驗結果影像，圖 6（a）是將原始影像使用遮罩遮蓋部分畫面的輸入影像，圖 6（b）是經由我們的系統所修補後的輸出影像。

（a） （b）

圖 6 背景修補實驗結果

資料來源：自行繪製。

四、舞蹈影片生成實驗

在此實驗當中，我們串接上述三個模組，包含身體姿勢遷移、臉部姿勢轉移和背景修

補，產生出最後的舞蹈影片。圖 7 為我們的舞蹈影片生成實驗結果，擷取影片中的部分影格（Frame），由圖 7 可觀察到本論文所生成的影片能夠將人物轉換成不同的舞蹈動作，並生成出使其在原影像背景中跳舞的影片。

（a） （b） （c） （d）

圖 7　舞蹈影片生成實驗結果

資料來源：自行繪製。

伍、結論

此論文中提出了一個用於舞蹈影片生成的系統框架。在身體姿勢轉移生成模型的部分包含設計了多尺度區域提取器，根據身體關鍵點捕獲每個身體區域的特徵，它解決了透過捷徑傳遞數據時姿勢傳輸中由於身體大運動造成的形狀變形問題，以減少 U-net 的空間資訊損失，使用多尺度區域提取器將透過所有區域的風格表示計算出的區域風格損失添加到損失函數中。此外，RN-L 被整合在所提出的框架中，以自動學習損壞區域並獲得區域遮罩來執行區域正規化。我們所提出的系統，在舞蹈姿勢中有較大的身體動作時，仍可以生成高品質的舞蹈影像。在人臉角度轉移的部分，我們使用了姿勢注意力轉移網路模型實現漸進式的人臉角度轉換，並配合不同的鑑別器，其中一個是用來學習旋轉角度的區別，一個是用來提升區分臉部結構的多樣性和獲取局部感知訊息的能力，而最後一個主要的功能是加強人臉重點區域的生成品質。基於本模型的架構下，讓使用者可以將原始人臉旋轉至目標角度。在背景影像修補的部分，我們提出了一種具有區域正規化和上下文注意的兩階段影像修復架構，該系統以損壞的圖像和遮罩作為輸入，執行從粗到細的圖像修復，我們結合了上下文注意力和區域標準化的優點來設計所提出的框架。此外，還修改了區域標準化的實作細節，以提高修復品質。此外，我們使用基於 CIEDE2000 色差公式的感知顏色損失函數來解決顏色偏移問題，使得修復結果的顏色更貼近人類視覺的感知。

參考文獻

Ballester C., Bertalmio M., Caselles V., Sapiro G., and Verdera J., 2001, "Filling-In by Joint Interpolation of Vector Fields and Gray Levels," *IEEE Transactions on Image Processing*, 10(8), 1200-1211. doi:10.1109/83.935036

Barnes C., Shechtman E., Finkelstein A., and Goldman D. B., 2009, "PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing," *ACM Transactions on Graphics*, 28(3), 24. doi:10.1145/1531326.1531330

Barnes C., Shechtman E., Goldman D. B., and Finkelstein A., 2010, "The Generalized PatchMatch Correspondence Algorithm," in *Computer Vision—ECCV 2010*, Berlin, Germany: Springer, 29-43. doi:10.1007/978-3-642-15558-1_3

Bertalmio M., Sapiro G., Caselles V., and Ballester C., 2000, "Image Inpainting," in *SIGGRAPH'00: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, New York, NY: ACM Press, 417-424. doi:10.1145/344779.344972

Bertalmio M., Vese L., Sapiro G., and Osher S., 2003, "Simultaneous Structure and Texture Image Inpainting," *IEEE Transactions on Image Processing*, 12(8), 882-889. doi:10.1109/TIP.2003.815261

Cao Z., Simon T., Wei S.-E., and Sheikh Y., 2017, "Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA: IEEE Computer Society Press, 1302-1310. doi:10.1109/CVPR.2017.143

Chen L.-C., Papandreou G., Schroff F., and Adam H., 2017, "Rethinking Atrous Convolution for Semantic Image Segmentation," https://doi.org/10.48550/arXiv.1706.05587 (accessed February 18, 2023).

Chen Z., Dai C., Jiang L., Sheng B., Zhang J., Lin W., and Yuan Y., 2016, "Structure-Aware Image Inpainting Using Patch Scale Optimization," *Journal of Visual Communication and Image Representation*, 40(A), 312-323. doi:10.1016/j.jvcir.2016.06.029

Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., and Bengio Y., 2014, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems 27: 28th Annual Conference on Neural Information Processing Systems 2014*, Red Hook, NY: Curran Associates, 2672-2680.

He K., Zhang X., Ren S., and Sun J., 2014, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in *Computer Vision—ECCV 2014*, Cham, Germany: Springer, 346-361. doi:10.1007/978-3-319-10578-9_23

Hu Y., Wu X., Yu B., He R., and Sun Z., 2018, "Pose-Guided Photorealistic Face Rotation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA: IEEE Computer Society Press, 8398-8406. doi:10.1109/CVPR.2018.00876

Huang R., Zhang S., Li T., and He R., 2017, "Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Los Alamitos, CA: IEEE Computer Society Press, 2458-2467. doi:10.1109/ICCV.2017.267

Iizuka S., Simo-Serra E., and Ishikawa H., 2017, "Globally and Locally Consistent Image Completion," *ACM Transactions on Graphics*, 36(4), 107. doi:10.1145/3072959.3073659

Isola P., Zhu J.-Y., Zhou T., and Efros A. A., 2017, "Image-to-Image Translation with Conditional Adversarial Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA: IEEE Computer Society Press, 5967-5976. doi:10.1109/CVPR.2017.632

Lassner C., Pons-Moll G., and Gehler P. V., 2017, "A Generative Model of People in Clothing," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Los Alamitos, CA: IEEE Computer Society Press, 853-862. doi:10.1109/ICCV.2017.98

Lee C.-H., Liu Z., Wu L., and Luo P., 2020, "MaskGAN: Towards Diverse and Interactive Facial Image Manipulation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA: IEEE Computer Society Press, 5548-5557. doi:10.1109/CVPR42600.2020.00559

Li S., Lu L., Zhang Z., Cheng X., Xu K., Yu W., He G., Zhou J., and Yang Z., 2020, "Interactive Separation Network for Image Inpainting," in *2020 IEEE International Conference on Image Processing (ICIP)*, Los Alamitos, CA: IEEE Computer Society Press, 1008-1012. doi:10.1109/ICIP40778.2020.9191263

Li X., Wang L., Cheng Q., Wu P., Gan W., and Fang L., 2019, "Cloud Removal in Remote Sensing Images Using Nonnegative Matrix Factorization and Error Correction," *ISPRS Journal of Photogrammetry and Remote Sensing*, 148, 103-113. doi:10.1016/j.isprsjprs.2018.12.013

Lin J., Xia Y., Qin T., Chen Z., and Liu T.-Y., 2018, "Conditional Image-to-Image Translation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA: IEEE Computer Society Press, 5524-5532. doi:10.1109/CVPR.2018.00579

Liu G., Reda F. A., Shih K. J., Wang T.-C., Tao A., and Catanzaro B., 2018, "Image Inpainting for Irregular Holes Using Partial Convolutions," in *Computer Vision—ECCV 2018*, Cham, Germany: Springer, 89-105. doi:10.1007/978-3-030-01252-6_6

Liu Z., Luo P., Qiu S., Wang X., and Tang X., 2016, "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Los Alamitos, CA: IEEE Computer Society Press, 1096-1104. doi:10.1109/CVPR.2016.124

Ma L., Jia X., Sun Q., Schiele B., Tuytelaars T., and Van Gool L., 2017, "Pose guided person image generation," in *Advances in Neural Information Processing Systems 30: 31st Annual Conference on Neural Information Processing Systems (NIPS 2017)*, Red Hook, NY: Curran Associ-

ates, 406-416.

Ma L., Sun Q., Georgoulis S., Van Gool L., Schiele B., and Fritz M., 2018, "Disentangled Person Image Generation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR*), Los Alamitos, CA: IEEE Computer Society Press, 99-108. doi:10.1109/CVPR.2018.00018

Neverova N., Alp Güler R., and Kokkinos I., 2018, "Dense Pose Transfer," in *Computer Vision—ECCV 2018*, Cham, Germany: Springer, 128-143. doi:10.1007/978-3-030-01219-9_8

Ronneberger O., Fischer P., and Brox T., 2015, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, Cham, Germany: Springer, 234-241. doi:10.1007/978-3-319-24574-4_28

Siarohin A., Sangineto E., Lathuilière S., and Sebe N., 2018, "Deformable GANs for Pose-Based Human Image Generation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA: IEEE Computer Society Press, 3408-3416. doi:10.1109/CVPR.2018.00359

Tian Y., Peng X., Zhao L., Zhang S., and Metaxas D. N., 2018, "CR-GAN: Learning Complete Representations for Multi-View Generation", in *IJCAI'18: Proceedings of the 27th International Joint Conference on Artificial Intelligence*, Palo Alto, CA: Association for the Advancement of Artificial Intelligence, 942-948. doi:10.24963/ijcai.2018/131

Wang N., Zhang Y., and Zhang L., 2021, "Dynamic Selection Network for Image Inpainting," *IEEE Transactions on Image Processing*, 30, 1784-1798. doi:10.1109/TIP.2020.3048629

Wang W., Zhang J., Niu L., Ling H., Yang X., and Zhang L., 2021, "Parallel Multi-Resolution Fusion Network for Image Inpainting," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Los Alamitos, CA: IEEE Computer Society Press, 14539-14548. doi:10.1109/ICCV48922.2021.01429

Wu X., He R., Sun Z., and Tan T., 2018, "A Light CNN for Deep Face Representation with Noisy Labels," *IEEE Transactions on Information Forensics and Security*, 13(11), 2884-2896. doi:10.1109/TIFS.2018.2833032

Yi Z., Tang Q., Azizi S., Jang D., and Xu Z., 2020, "Contextual Residual Aggregation for Ultra High-Resolution Image Inpainting," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA: IEEE Computer Society Press, 7508-7517. doi:10.1109/CVPR42600.2020.00753

Yin Y., Jiang S., Robinson J. P., and Fu Y., 2020, "Dual-Attention GAN for Large-Pose Face Frontalization," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, Los Alamitos, CA: IEEE Computer Society Press, 249-256. doi:10.1109/FG47880.2020.00004

Yu C., Wang J., Peng C., Gao C., Yu G., and Sang N., 2018, "BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation," in *Computer Vision—ECCV 2018*, Cham,

Germany: Springer, 334-393. doi:10.1007/978-3-030-01261-8_20

Yu J., Lin Z., Yang J., Shen X., Lu X., and Huang T. S., 2018, "Generative Image Inpainting with Contextual Attention," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA: IEEE Computer Society Press, 5505-5514. doi:10.1109/CVPR.2018.00577

Yu J., Lin Z., Yang J., Shen X., Lu X., and Huang T., 2019, "Free-Form Image Inpainting with Gated Convolution," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Los Alamitos, CA: IEEE Computer Society Press, 4470-4479. doi:10.1109/ICCV.2019.00457

Yu T., Guo Z., Jin X., Wu S., Chen Z., Li W., Zhang Z., and Liu S., 2020, "Region Normalization for Image Inpainting," *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7), 12733-12740. doi:10.1609/aaai.v34i07.6967

Zhang K., Zhang Z., Li Z., and Qiao Y., 2016, "Joint Face Detection and Alignment Using Multi-Task Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, 23(10), 1499-1503. doi:10.1109/LSP.2016.2603342

Zheng C., Cham T.-J., and Cai J., 2021, "Pluralistic Free-Form Image Completion," *International Journal of Computer Vision*, 129, 2786-2805. doi:10.1007/s11263-021-01502-7

Zheng L., Shen L., Tian L., Wang S., Wang J., and Tian Q., 2015, "Scalable Person Re-Identification: A Benchmark," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Los Alamitos, CA: IEEE Computer Society Press, 1116-1124. doi:10.1109/ICCV.2015.133

Zhou B., Lapedriza A., Khosla A., Oliva A., and Torralba A., 2018, "Places: A 10 Million Image Database for Scene Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6), 1452-1464. doi:10.1109/TPAMI.2017.2723009

Zhu Z., Huang T., Shi B., Yu M., Wang B., and Bai X., 2019, "Progressive Pose Attention Transfer for Person Image Generation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA: IEEE Computer Society Press, 2342-2351. doi:10.1109/CVPR.2019.00245