

中華民國第 60 屆中小學科學展覽會 作品說明書

高級中等學校組 電腦與資訊學科

第二名

052501

利用深度學習改善自拍人像構圖

學校名稱：臺北市立第一女子高級中學

作者： 高二 陳 新 高二 朱致伶	指導老師： 黃芳蘭
---------------------------------	------------------

關鍵詞：深度學習、影像處理

摘要

手機自拍像中常有人臉佔照片整體面積比例過大、構圖不佳的問題，針對此問題本研究運用深度學習提出改善方法。所設計的方法，主要包含四大部分：圖像擷取、圖像外擴、圖像填補以及圖像結合。圖像擷取中，我們運用 Mask_RCNN 和 Closed_Form_Matting 兩個演算法，結合兩者優點來生成遮罩，正確率可達 87.5%；然後，藉由更改模型架構以優化與對訓練集做不同的前置處理，比較不同模型成效並修正，可對人像自動進行適當的外擴；此外，在圖像填補中，我們設計差集填補的方式來提升填補模型準確率；最後，自動將外擴後人像與填補後背景做圖像結合。實驗結果顯示，所設計的方法可有效地對圖像擷取與圖像外擴模型做優化，並將自拍像中過大的人像縮小至合理大小，形成結構均衡的照片。

壹、研究動機

自拍是現代人們日常生活常運用到的拍攝手法，只要憑藉手機即可獨自拍攝。市面上更是推出各式各樣的照片編輯器、濾鏡，以滿足廣大使用者的需求。

由於自拍是近幾年興起的拍攝方式，我們認為這個領域尚有發展空間。自拍照的一大缺失在於相機與人像本身的距離受極大的限制，因而導致臉部過大使圖像不勻稱甚至人像些微失真。目前民眾普遍以第三者拍攝或使用自拍棒解決此問題，可知對於此問題市面上有一定的需求，而鮮少有對自拍照本身直接修圖、校正的應用軟體。若可以依比例適度縮小攝影者的影像，使照片中的人像大小合理，便能以另一種方法解決大多數自拍者的困擾。

貳、研究目的

綜合上述討論，本研究之研究目的為「創建一完善系統將自拍個人照中的人像大小調整至勻稱比例」，將問題簡化至單人自拍照的情況，可細分為下列項目：

- 一、 研究圖像擷取技術，並自行編撰程式結合 Mask_RCNN 和 Closed Form Matting 的輸出遮罩，達到最佳人像切割效果。
- 二、 研究圖像外擴技術並尋找合適資料集訓練外擴模型，更改模型架構以加入人物解析圖一同作為輸入，比較不同資料集、訓練時間與優化後模型對生成效果的影響。
- 三、 研究圖像填補技術，提出適合本研究的填補方式，測試填補模型。
- 四、 撰寫圖像結合模型以有效結合填補後的圖像前景（人像）與外擴後的背景。
- 五、 撰寫程式整合以上技術，建構一完善系統達成「將自拍個人照中的人像大小調整至

勻稱比例」的研究目的。

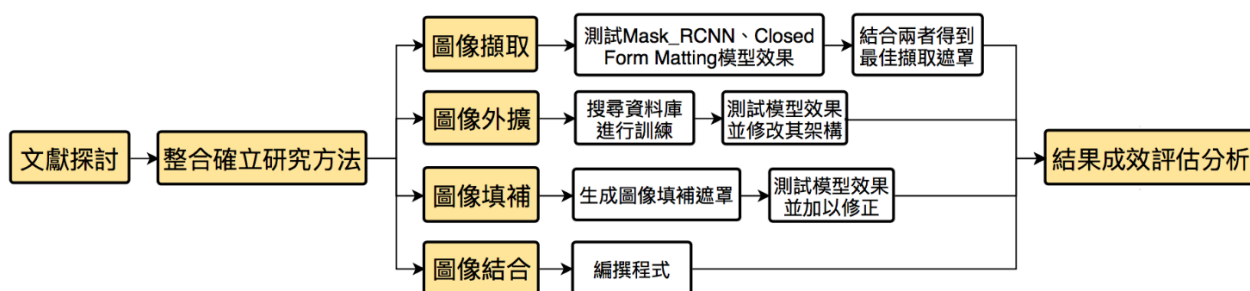
參、研究設備及器材

- 一、硬體：筆記型電腦（GPU： GTX 1080Ti 、12GB RAM）、GPU： Nvidia Tesla k80。
- 二、軟體及工具：Python（3.6.5、2.7.17）、torch（1.0.1.post2、0.4）、torchvision（0.2.2.post）、opencv-python（4.0.0.21）、tensorflow（1.15.0、2.0.0a0）、cuda（10.1）、相關套件：numpy、scipy、tqdm、pillow。

肆、研究過程及方法

一、研究流程與架構

以下是我們根據本研究之研究目的—「將自拍個人照中的人像大小調整至勻稱比例」所建構的研究流程。我們分為圖像擷取、圖像外擴、圖像填補、圖像結合四個部份分別深入探討，最後再將其整合成完整的實驗操作流程。



圖一、研究流程示意圖

二、研究文獻探討

（一）卷積神經網路（Convolutional Neural Network, CNN）

卷積神經網路是參考人的大腦視覺組織建立的深度學習模型，多運用於影像辨識上，其功能為擷取圖片特徵，找出明顯特徵再進行分類。以下對卷積神經網路包含的專有名詞與其架構主要的四個部分做介紹：

1. 特徵（Feature）

CNN 會比較原圖和生成器生成的圖片之間各個局部，而這些局部就被稱為特徵。比起比較整張圖片，藉由在小範圍相似的位置上比對大略特徵，CNN 能更好地分辨兩張圖片是否相同。

2. 卷積層 (Convolution Layer)

卷積層的目的為歸納輸入圖片中各種可能的特徵，卷積層透過隨機生成的特徵偵測器判斷局部圖片和判別物特徵的相符程度。重複動作以歸納出圖片中各種可能的特徵，再根據每次卷積計算後的值和位置，製作一個新的二維矩陣，而這新的二維矩陣又稱為特徵圖 (Feature Map)。

3. 池化層 (Pooling Layer)

池化層是一個壓縮圖片並保留重要圖片資訊的方法，有助於改善龐大運算所衍伸之問題。池化會選取不同窗口 (Filter)，並在這個窗口涵蓋的範圍中保留一個最大值，藉此保留局部圖片和特徵的相符程度。其中，卷積計算中的 Stride 即窗口的每次池化所移動的格數，重複上述過程，就可以從低階特徵提取出更具象徵性的特徵。

4. 線性整流單元 (Rectified Linear Unit, ReLU)

線性整流單元則是將圖片上的所有負數轉為 0，ReLU 可以把線性運算轉為非線性，讓網路更容易學會非線性特徵。

5. 全連結層 (Fully Connected Layer, FC)

全連結層會集合高階層中篩選過的特徵，並將這些特徵資訊轉化為數值。由於某些特徵的數值可以更有效地判別結果，所以會以權重 (Weight) 的方式調整其對結果的影響程度。

(二) 全卷積神經網路 (Fully Convolutional Network, FCN)

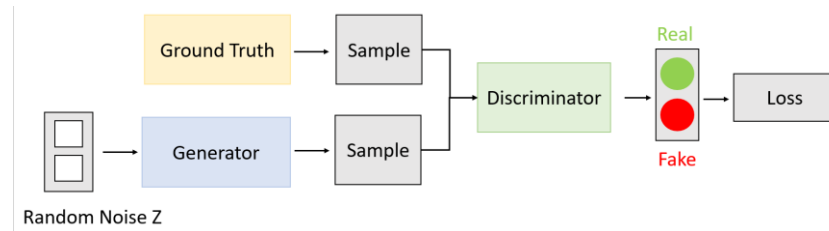
FCN 能夠對影像畫素進行分類，可以接受任意尺寸的輸入影像。利用反卷積層對最後一個卷積層的特征 map 做上取樣 (近似放大影像的動作)，使它恢復到輸入影像尺寸，保留輸入影像的空間資訊。最後在上取樣的特徵圖上進行逐畫素分類，常用於實體分割問題。

(三) 生成對抗網路 (Generative Adversarial Network, GAN)

生成對抗網路簡稱 GAN，其概念啟發自博弈論中的二人零和博弈，GAN 模型中包含兩位博弈者：生成器 (Generator) 和判別器 (Discriminator)。生成器 G 捕捉樣本資料的分布，利用符合某一分布 (均勻分布、高斯分布等等) 的雜訊 Z 生成一個類似真實訓練資料的樣本，目標越像真實樣本越好；而判別器 D 則是一個二分類器，估計一個樣本來自於真實訓練資料的機率，若其判斷樣本來自真實訓練資料，就輸出大機率，反之則為小機率，目標是能判斷每一個樣本是真是假。藉由生成器和判別器之間不斷的對抗，可以讓生成器達到最佳的學習效果。上述過程如下列公式及下圖所

示：

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log 1 - D(G(z))]$$



圖三、生成對抗網路 (GAN) 模型示意圖 [11]

三、研究方法設計

(一) 圖像擷取

運用圖像擷取技術擷取自拍照中的人像，以利將人像與背景分開做後製。此技術需符合以下兩個條件：框選範圍正確、框選邊框細緻，我們依據這兩項條件尋找適合模型測試，並且自行設計程式將兩模型—Mask_RCNN、Closed Form Solution to Natural Image Matting 結合在一起，分別取其優點以達最佳效果。

1. Mask_RCNN

(1) 模型介紹

Mask_RCNN [1]是一種深度學習影像分析技術，可以達成影像實體分割 (Instance Segmentation)。實體分割可再分為物體偵測 (Object Detection) 與意義分割 (Semantic Segmentation) 兩部分。我們先初步定義上述三個分割技術：

- I. 物體偵測：分類物件，輸出每個物件最小矩形範圍稱為 Bounding Box。
- II. 意義分割：依照事物類別做區分，輸出分割每個類別的遮罩。故同類不同物件無法被分割開來。
- III. 實體分割：綜合上述技術，偵測影像中的物件並依其外形分割開來。



圖四、圖像分割技術 [6]

(2) Mask_RCNN 的模型可以分成三個部分：

I. 區域建議網路 (Region Proposal Network, RPN)

在原圖通過 CNN 得到 Feature Map 後會進入區域建議網路 (Region

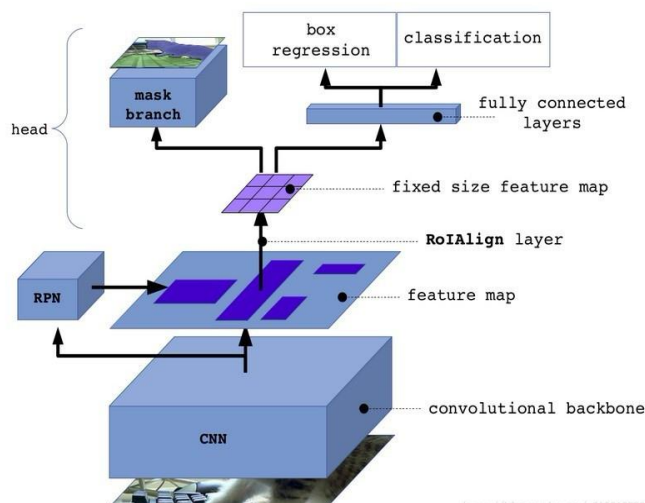
Proposal Network, RPN)。在 RPN 中要判斷物體位置就必須將影像分割成小部分來分析，而這些不同大小與比例的邊界框稱為 Anchor。在走訪圖片時，會計算每個 Anchor 中有無包含物體的機率和其位置與物體的偏移量。之後將有無物體的機率排序，取機率為正值的前 300 個，再運用偏移量調整 Anchor 大小。

II. RoI_Align (Region of Interest Align)

在進行完 RPN 後加入一池化層—RoI_Align，在保留確切的空間位置的下壓縮 Feature Map。RoI_Align 特別的地方在於執行時其 Stride 值不一定要為整數。輸入為固定大小的 Feature Map 判斷該像素有無物件的機率，和 Bounding Box 的位置；輸出為其池化後的結果。由於經過池化後數據減少，可以顯著提升訓練和測試的速度。

III. 生成物體偵測與意義分割遮罩

利用 RoI_Align 生成的 Feature Map 分兩個分支進行。一支是利用 FCN (Fully Convolutional Networks) 逐像素預測，生成意義分割遮罩。運用 FCN 可以保留空間維度的訊息，預防資料遺失。另一分支亦利用 FCN 生成 Bounding Box 和其物件類別。



圖五、Mask_RCNN

2. A Closed Form Solution to Natural Image Matting (Closed Form Matting)

(1) 模型介紹

Closed Form Matting [2]是一種準確切割人像的方法，可以切割出髮絲等細微部位。輸入為要做人像切割的原圖和其 Trimap。從前景和背景的紋路中得到一個

可以判定待確認區為前景或背景的函式，最後生成人像不透明度遮罩 Alpha Map，能將人像有漸層地切割下來。

(2) 製作 Closed Form Matting 的 Trimap

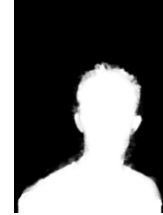
Trimap 是將影像中每像素劃分為三種區域：前景、背景和待確認區（如圖七）。為了自動生成 Trimap，我們運用 Mask_RCNN 切割的人像遮罩來製作。將 Mask_RCNN 的人像遮罩做開運算，開運算即把遮罩中白色區域先內推（Erode）再外移（Dilate），形成邊界區域當作待確認區，並分割前景與背景。

(3) Mask_RCNN 和 Closed Form Matting 的結合

在實驗中 Closed Form Matting 可以精準的切出髮絲等細微部位，但在人像下半部邊界無須漸層處（如衣服和圖像底部）會有誤判的現象，而 Mask_RCNN 正好相反。因此我們只框出人像肩膀以上的部分做 Closed Form Matting，其餘部分則用 Mask_RCNN 所框選的結果以提升正確率（如圖七）。另外，由於 Mask_RCNN 在框選人像下半部時常有無法框出人像底部的現象，因此我們編撰程式偵測人像底部最左和最右處，將人像遮罩向下延伸至圖像底部（如圖八）。



圖七、經調整後的 Trimap



圖八、結合後的擷取遮罩

(二) 圖像外擴 Image Outpainting

圖像外擴是對模型輸入一張不完整的物件圖片，從輸入的圖片中尋找潛在的特徵，對圖像做語意合理且保持紋理一致性的外擴，形成一張內容物完整的圖片。本研究利用此技術達到將部分人像外擴至全身的目的，以便在實際運用上縮小人像後依舊保持照片完整性。

1. 模型

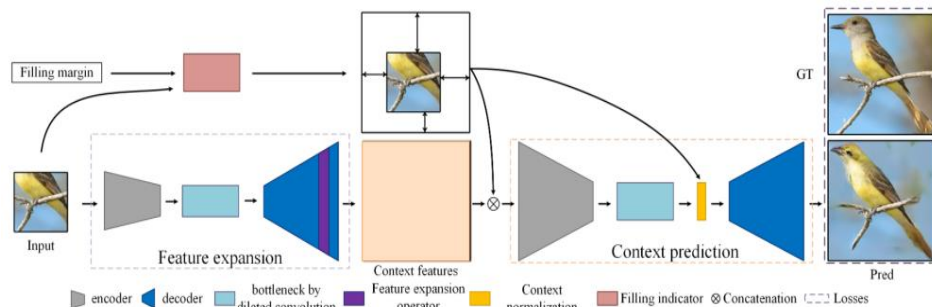
圖像外擴的整體架構為引用自[3]中的模型，採用部分自建資料集，經過重新訓練後測試，其類似生成對抗網路（GAN），由生成器和判別器組成，此模型的生成器-- Semantic Regeneration Network 尤為特別。它可以從小範圍的照片中再生出完整的物件結構，且其再生的物件為任意大小，生成器主要分為兩個部分，以下對兩者分別做介紹：

(1) Feature Expansion Network (FEN)

FEN 的目的是找出不完整圖像中潛在的特徵並提取出來。其以類似 Encoder-decoder Structure 為基礎，也就是輸入一張為 $h \times w \times c$ 的圖像，輸出一張二維 $h' \times w' \times c$ 的特徵圖，利用多層卷積從最接近的區塊提取潛在特徵逐漸向外延伸。

(2) Context Prediction Network (CPN)

CPN 的目的是判斷特徵在已知和未知區域中的分布並將提取的特徵從矩陣分析轉換成影像呈現。判斷是否為未知區域是很重要的環節，因為在像素生成中，對於某些象徵意義較為敏感的物件，來自已知和未知區域的特徵訊息重要性會有所不同。將 FEN 輸出的特徵圖與一張二維遮罩一併輸入，二維遮罩以 0 表示未知區塊、1 表示已知區塊。



圖九、原圖像外擴模型之架構圖[3]

2. 模型訓練

(1) 尋找訓練集

I. ATR_dataset

ATR_dataset 為一個單人全身的照片集合，總共約為 17,000 張，大多用在人物分析 Human Parsing 的訓練上，但其大多為女性且背景、人像動作和衣著過於複雜，導致訓練模型效果不如預期。



圖十、ATR dataset



圖十一、Deep_fashion

II. Deep_fashion

針對背景單一、人像動作簡單、人像衣著清楚且不繁複等條件，Deep_fashion 是個合適的選擇，總共約為 5,500 張。而我們選擇正對著鏡頭且人像為全身的類型，此類型的檔名多為 `_pose.type`，再從中篩選出正面的資料。

以 Deep_fashion 為訓練集的模型和 ATR_dataset 相比，學習效果較佳。

(2) 前置處理

I. 設定訓練集的圖像大小 (Image_shape) 和遮罩大小 (Mask_shape)

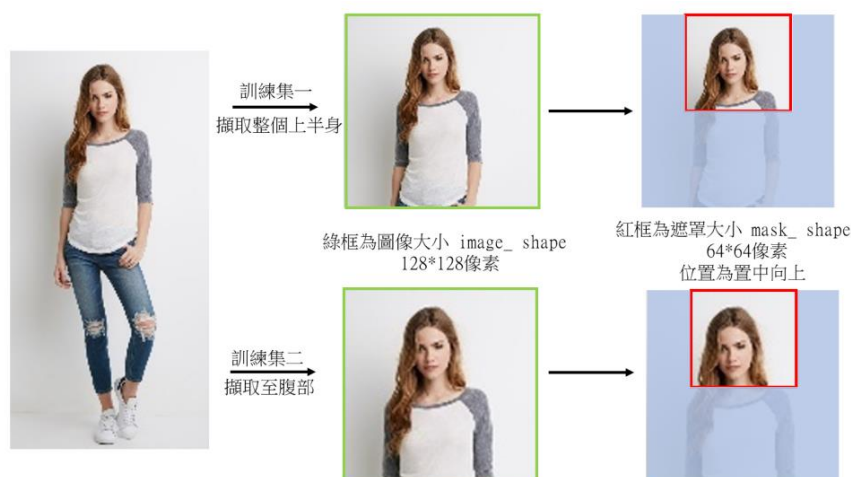
在訓練模型過程中，圖像大小 (Image_shape) 的定義類似實際運用時經過圖像外擴後的人像結果，而遮罩大小 (Mask_shape) 則類似實際運用時輸入的自拍照。除了圖像大小和遮罩大小必須固定的限制以外，還須滿足遮罩大小的長和寬皆為圖像大小長和寬的因數。

在設定圖像大小時，考慮人像縮小程度和模型學習效果，若由自拍中的人像直接生成全身人像，不只訓練困難度極高且在實際運用的縮放過程中並不需。因此我們分別設定兩組訓練集，分別為擷取 Deep_fashion 訓練集的上半身至腹部及臀部，將兩組訓練集的圖像大小皆調整為 128*128 像素 (H×W)。

在設定遮罩大小時，挑選的訓練集本身滿足人像為全身的條件。因自拍照的人像大多只露出胸部以上且肩膀多不完整，所以我們仿照自拍照的樣式做為模型輸入，在多張照片相互比對後設定遮罩大小為 64*64 像素 (H×W)。

II. 遮罩位置的設定

在訓練過程中，遮罩位置的設定影響模型訓練的生成影像。遮罩位置的設定理應包括人像的頭部、兩邊各半的肩膀及部分胸部，故遮罩位置設定為圖像置中向上的位置。



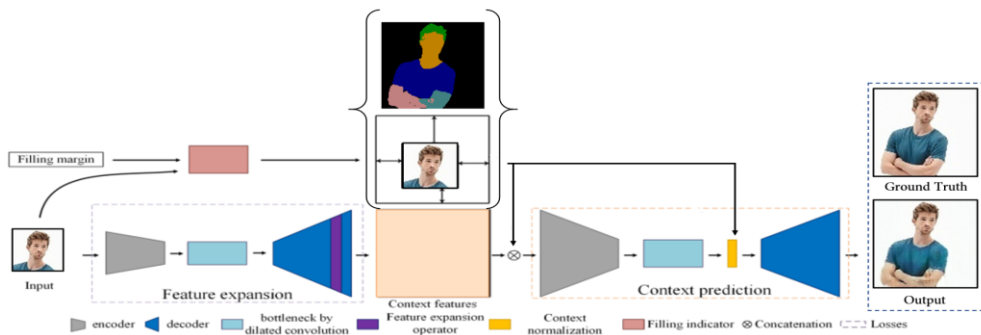
圖十二、Deep_fashion 前置處理示意圖

3. 模型優化—加入「人物解析圖」

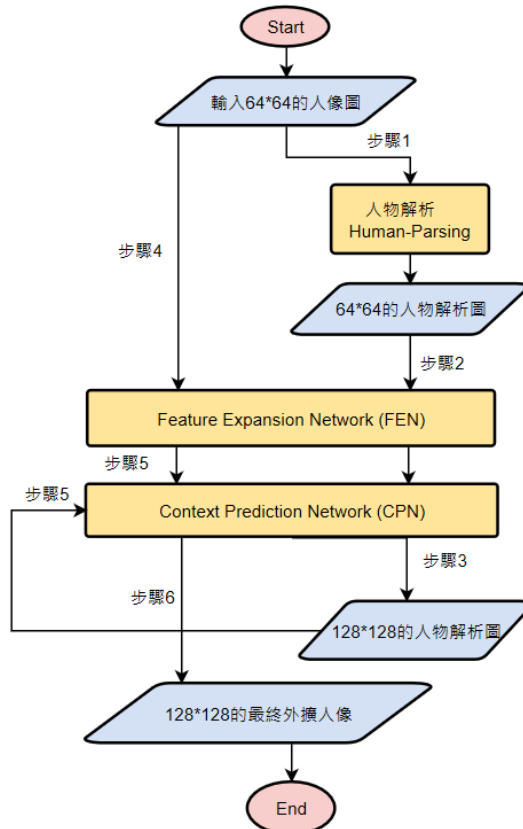
原圖像外擴模型的輸入僅有遮罩大小的原圖，透過模型生成圖像大小的外擴結果

困難度極高且效果不穩定，因此我們透過增加人物解析圖一同作為輸入以提高圖像外擴的準確率與穩定性。

人物解析 Human-Parsing [10] 是語意分割中常見問題，指判斷圖片中人像身體部位或穿著服飾並劃分區塊以歸類。人物解析圖由代表各身體部位、服飾的色塊所組成，其外擴因不需考慮像素的顏色、紋理，只須考慮像素所屬的區塊意義與生成人像的姿勢，和含有人像的原圖外擴相比困難度降低，導致人物解析圖像外擴的準確率與穩定性提高。透過以外擴的人物解析圖與尚未外擴的原圖一同作為輸入，外擴後的人物解析圖可以作為原圖外擴時的指引，使得原圖外擴時不再是毫無依據的生成，大幅提高最終外擴人像的準確率和穩定性。



圖十三、優化後圖像外擴模型之架構圖



圖十四、優化後圖像外擴模型之流程圖

(三) 圖像填補 Image Inpainting

圖像填補是指結合現實情況和圖像意義以重建圖像空缺部分，並同時保持圖像整體紋理風格的統一。主要分為兩種方法：其一為 Exemplar-based Inpainting，例如：PatchMatch，其方法大致為從已知的部分尋找並複製與空缺處看似相似的圖塊，然而此方法不適合運用在較複雜或未包含在圖庫中的圖像。近幾年大多使用另一種方法 Deep CNN-based Inpainting，其概念以卷積神經網路（CNN）為基礎，CNN 可以突破圖像空缺處未包含在圖庫中的限制，再結合生成對抗網路（GAN），藉由生成器和判別器之間的對抗，可以更有效地生成較複雜和不可重複的圖像結構，例如：Global and Local consistent image completion（GL）、Contextual Attention（CA）、Partial Convolution（PConv）。

1. 模型

本文使用的是局部卷積神經網路（Partial Convolution, PConv）作為達到 Inpainting 效果的模型，整體架構為引用自[4]中的模型。局部卷積神經網路其對卷積神經網路做的改善為以下兩項，也是因此局部卷積神經網路會有比較好的學習效果：

(1) 局部卷積層的定義

$$x' = \begin{cases} \mathbf{W}^T(\mathbf{X} \odot \mathbf{M}) \frac{1}{\text{sum}(\mathbf{M})} + b, & \text{if } \text{sum}(\mathbf{M}) > 0 \\ 0, & \text{otherwise} \end{cases}$$

其中 W 為卷積層的權重（Weight），X 為當前像素與潛在特徵的比值，b 為預測值和實際值的偏差（Bias），M 為區分是否為空洞的二元遮罩（Binary Mask），0 代表是空洞也就是要填補的地方，1 則代表已有像素的地方， \odot 則是阿達瑪卷積（Hadamard Product），是一種二元矩陣計算。下面則是對二元遮罩做更詳細的定義和說明：

$$m' = \begin{cases} 1, & \text{if } \text{sum}(\mathbf{M}) > 0 \\ 0, & \text{otherwise} \end{cases}$$

1	1	1	1	1	1	0	0
1	1	1	1	1	1	0	0
1	1	1	1	1	0	0	0
1	1	1	1	0	0	0	0
1	1	1	1	0	0	0	0
0	0	1	1	1	0	0	0
0	0	0	1	1	0	0	0

圖十三、圖示局部卷積神經網路中二維遮罩計算會出現的三種情況

I. 卷積核（Kernel）內皆為 1（即紅框）：

卷積核內的二元遮罩數值皆為 1，也就是說都是已填補過像素，此時可

用一般的卷積層處理。

II.卷積核內有些部分為 1，有些部分為 0（即綠框）：

卷積核內的二元遮罩數值部分為 1、部分為 0，也就是說有些地方是需要填補的，此時卷積網路會從已知的區塊提取特徵作為填補未知區域紋理或顏色的依據，把未知的區域補上。

III.卷積核內皆為 0（即藍框）：

卷積核內的二元遮罩數值皆為 0，即卷積核所框住的範圍皆為空缺需要填補，此時先暫不做處理，等到越後面的卷積層有多資訊時再做填補。透過二元遮罩對要填補的區域做篩選，因此可以確保周圍皆有已填補過的像素，避免卷積核內全為未知區塊卻依然要做填補的窘境，提升填補色塊準確度。

（2）突破未知區塊的形狀限制

在局部卷積神經網路提出之前，先前圖像填補模型的空缺處大多為長方形，且位置大多分布在圖像中心，這些對於空缺處的限制會降低訓練模型的學習效果，因此將遮罩的空缺改為任意形狀、大小和位置，其中還為空缺處加設邊界，更加豐富遮罩的多樣性。

2. 模型訓練

（1）尋找訓練集

考慮到自拍者背景十分多樣，我們選擇 Places2 作為訓練集，總共有約 15,000 張不同地點的拍攝。Places2 有超過一百萬張的景點照片，有 400 個以上的地點分類，每個類別都有 5,000 張以上的照片。



圖十四、Places2 訓練集

3. 差集填補遮罩的生成

由於原圖人像擷取下來後背景空缺區域甚大，模型對於大面積的填補效果並不穩定，且實際運用上會將外擴後人像貼回背景，因此真正需要填補的區域只有「原圖背景空缺處減去外擴後人像遮蓋」的範圍。為了達到縮減填補範圍的目的，我們編撰程式將原圖人像擷取遮罩以及人像外擴後人像擷取遮罩重疊，以找

出需要填補的區域，並把需填補區域標示為黑色，無須填補的為白色，最終生成差集填補遮罩。由於人像擷取遮罩有灰階區域，因此在生成差及填補遮罩時我們將原圖以及外擴後遮罩的灰階區域均判斷成需填補的黑色區域，達到最完整的填補結果。



(a) 原圖人像擷取遮罩 (b) 人像外擴後擷取遮罩 (c) 差集填補遮罩

圖十五、圖像填補遮罩示意圖

(四) 圖像結合

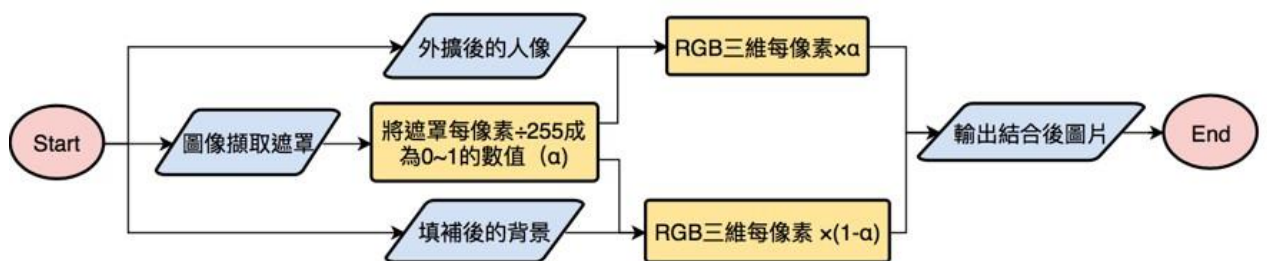
1. 自行編寫程式將外擴後的人像縮小貼回原圖

利用下式將圖像外擴後的人像擷取下來貼回原圖背景。以公式表示，其中 F 為前景，B 為背景，走訪每個像素 i，alpha map (α) 和前景、背景的乘積和為輸出 (I)，其公式為：
$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i$$



(a)原圖 (b)人像遮罩 (c)填補後的背景 (d)外擴後的人像 (e)輸出結果

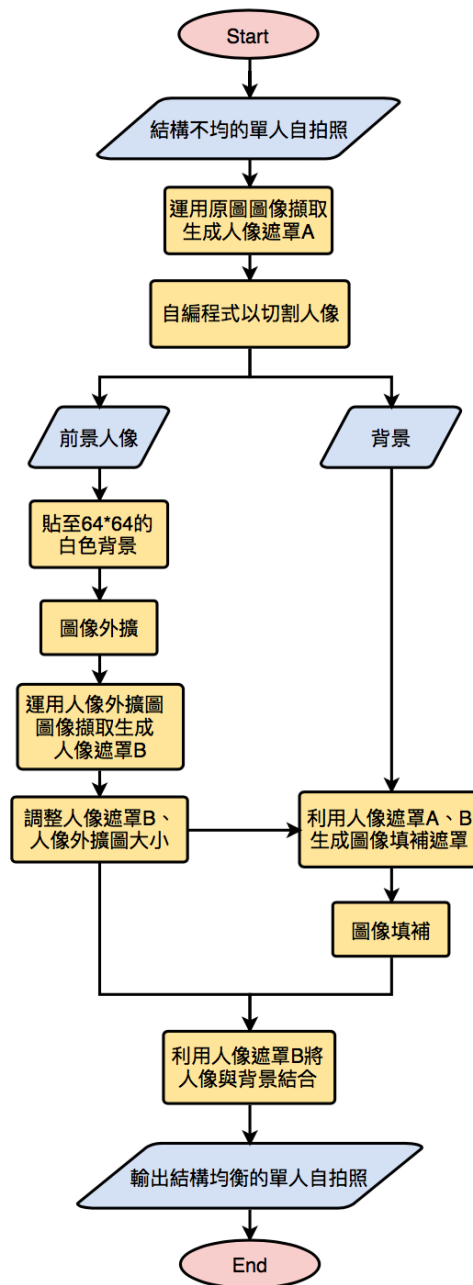
圖十六、模擬自拍照縮放示意圖



圖十七、圖像結合程式流程圖

四、步驟與流程

本研究透過自拍圖片中人像的縮小達成改善自拍人像構圖的目的，為此我們設計了步驟流程圖：



圖十八、實驗步驟流程圖

五、成效分析設計

由於圖像外擴的個人主觀性強，因此將外擴評分分為主觀以及客觀，能更精準判斷結果優劣，而擷取以及填補則以客觀測試為標準進行評分。而整體模型的輸出亦沒有客觀標準，因此運用主觀標準來評分。

(一) 主觀測試

1. 圖像外擴的評分

評分方式：根據多次模型訓練結果，判斷需要評分的項目分別為人像整體性、服

飾清晰度、姿勢合理性以及紋理風格，定義項目的評分標準，請 30 位受試者分別對生成的 8 張照片（附上原圖像）打分數，計算生成照片的效果總分。

表一、圖像外擴模型效果評分標準

	人像	服飾	姿勢	紋理
0	完全無法辨識輪廓	衣服零碎、不可辨識	人像比例不協調，姿勢怪異	紋理差異大且不和諧
+1	模糊且生成部位不完整	衣服輪廓大致清楚，但花樣不明顯	人像四肢些微不協調，大致和理	紋理大致相同、不突兀
+2	尚可辨識輪廓，但生成部位周邊模糊	衣服輪廓、花樣清楚，且分布在人像上	人像姿勢合理自然	紋理在視覺上和諧一致
+3	人像輪廓完整清晰			

2. 整體模型的評分

評分方式：請 30 位受試者分別對生成的 8 張照片（附上原圖像）打分數，計算生成照片的效果總分。

表二、整體效果模型效果評分標準

	清晰度	人像邊界銳利度	整體合理性
+1	模糊不可辨識人像部位	雜訊過多，邊界完全不可辨識	人像、背景大小完全不符合現實情況
+2	有大約部位輪廓但無法看出細節處	可見擷取痕跡但不影響整體完整性	人像、背景合理，但相比原圖並沒有使圖像更均衡
+3	人像部位輪廓清晰可辨，細緻度高	肉眼幾乎沒有看到加工痕跡	人像、背景合理，相比原圖更均衡

(二) 客觀測試

分別對圖像擷取、圖像外擴與圖像填補模型尋找最合適的客觀評分標準，分別為以下三項：

1. 圖像擷取—像素精度 (Pixel Accuracy)

$$PixelAccuracy = \frac{\sum_i n_{ii}}{\sum_i t_i}$$

計算標記正確的像素佔總像素的比例。類別 i 中， n_{ii} 為判斷正確像素和，

t_i 為總像素和。

2. 圖像外擴與填補—峰值信噪比（Peak Signal To Noise Ratio，PSNR）

$$PSNR = 10 \log_{10} \left(\frac{R^2}{MSE} \right), \quad MSE = \frac{\sum_{M,N} [I_1(m,n) - I_2(m,n)]^2}{M * N}$$

比較每像素的顏色相似度，值越大越相似。R 為圖中像素 RGB 的最大值（RGB 最大值為 255）。MSE：為均方誤差距離（Mean Square Error），其中 m, n 為像素的行列值；M, N 為圖像長與寬大小； I_1 與 I_2 為比較的兩張圖。單位為分貝（dB），數值的級距意義如下：

- (1) ≥ 40 dB：非常接近原始圖像。
- (2) 30~40dB：失真可以察覺但尚可接受。
- (3) 20~30dB：圖像品質差至不可接受。

3. 圖像外擴與填補—結構相似性（Structural Similarity Index，SSIM Index）

$$SSIM = L(X, Y)^\alpha * C(X, Y)^\beta * S(X, Y)^\gamma, \quad \begin{aligned} L(X, Y) &= \frac{2u_X u_Y + C_1}{u_X^2 + u_Y^2 + C_1} \\ C(X, Y) &= \frac{2\sigma_X \sigma_Y + C_2}{\sigma_X^2 + \sigma_Y^2 + C_2} \\ S(X, Y) &= \frac{\sigma_{XY} + C_3}{\sigma_X \sigma_Y + C_3} \end{aligned}$$

比較兩張圖片物件結構的相似度，X 為模型輸出的結果；Y 為原圖。一般設定 $\alpha = \beta = \gamma = 1$ 。L 為亮度對比函數；C 為對比度對比函數；S 為結構對比函數。 μ_x 、 μ_y 為 x、y 的像素平均值； σ_x 、 σ_y 為 x、y 的標準差； σ_{xy} 為 x 和 y 的共變異數；常數 C 為避免分母為 0 帶來的錯誤。SSIM 的範圍為 -1~1 之間，越接近 1 表示兩張圖片的相似度越高。

伍、研究結果

一、圖像擷取結果

測試集為 Deep Automatic Portrait Matting Dataset，總共 2,002 張，皆為單一人像在各地點的照片，人像包含腹部以上的自拍照片和其 Alpha Map，年齡、人種不拘。













我們測試 Mask_RCNN 輸出結果，並運用其切割遮罩製作 Trimap 測試 Closed Form Matting 效果，最終將二者結合得到優化的擷取效果。

(一) 展示部分測試及結果

1. 以下為運用 Deep Automatic Portrait Matting Dataset 為測試集的人像擷取效果。分別為 Mask_RCNN、Closed Form Matting 的輸出結果以及本研究結合兩模型輸出效果後的輸出結果。

表三、圖像擷取輸出結果

	圖片一	圖片二	圖片三	圖片四
測試集原圖				
Mask_RCNN 輸出結果				
Pixel Accuracy	0.863	0.845	0.836	0.881
Closed Form Matting 輸出結果				
Pixel Accuracy	0.733	0.788	0.789	0.851
本研究 (二者結合) 輸出結果				
Pixel Accuracy	0.868	0.840	0.890	0.889
	圖片五	圖片六	圖片七	圖片八
測試集原圖				

Mask_RCNN 輸出結果				
Pixel Accuracy	0.876	0.896	0.892	0.906
Closed Form Matting 輸出結果				
Pixel Accuracy	0.811	0.837	0.856	0.895
本研究 (二者結合) 輸出結果				
Pixel Accuracy	0.875	0.844	0.893	0.900

二、圖像外擴結果

(一) 模型一—外擴上半身至臀部

測試集為 Deep_fashion 訓練集取隨機 852 張，經過前置處理截至上半身至臀部。此測試集所需生成的人像部位較多，可達到大範圍的生成人像部位可供使用者自由選取之目的，總共需 8 天完整訓練此模型。下表由上至下分別為 Deep_fashion 中的測試集原圖、模型的輸入與輸出圖像。

1. 模型一的訓練結果展示與客觀成績

表四、圖像外擴模型一的訓練成果

	圖片一	圖片二	圖片三	圖片四
測試集原圖				
輸入				

輸出結果				
主觀評分 (分)	7.8	8.1	7.3	8.5
PSNR (dB)	34.316	35.921	34.504	34.911
SSIM Index	0.884	0.925	0.850	0.922
	圖片五	圖片六	圖片七	圖片八
測試集原圖				
輸入				
輸出結果				
主觀評分 (分)	7.6	8.5	7.2	7.8
PSNR (dB)	35.384	36.257	34.284	33.688
SSIM Index	0.907	0.942	0.841	0.793

2. 模型一的主觀評分

	總分	人像	服飾	姿勢	紋理
平均	7.4分 (9分)	2.2分 (3分)	1.6分 (2分)	1.7分 (2分)	1.9分 (2分)

(二) 模型二—外擴上半身至腹部

測試集為 Deep_fashion 訓練集與模型一測試集同樣的 852 張，再經過前置處理截至腹部做為測試集，因為此測試集較接近真實自拍照的人像部位分佈，可模擬外擴自拍照時的效果，完整訓練此模型總共需 6 天。下表由上至下分別為 Deep_fashion 中的測試集原圖、模型的輸入與輸出圖像以及其主觀和客觀成績。

1. 模型二的訓練結果展示與客觀成績

表五、圖像外擴模型二的訓練結果

	圖片一	圖片二	圖片三	圖片四
測試集原圖				
輸入				
輸出結果				
主觀評分 (分)	7.3	7.6	6.6	7.9
PSNR (dB)	33.075	34.296	31.484	34.110
SSIM Index	0.794	0.938	0.755	0.888
	圖片五	圖片六	圖片七	圖片八
測試集原圖				
輸入				
輸出結果				
主觀評分 (分)	7.5	7.3	8.4	7.0
PSNR (dB)	34.467	33.849	35.545	32.511
SSIM Index	0.909	0.893	0.941	0.725

2. 模型二的主觀評分

	總分	人像	服飾	姿勢	紋理
平均	7.6 分 (9 分)	2.5 分 (3 分)	1.5 分 (2 分)	1.7 分 (2 分)	1.9 分 (2 分)

(三) 模型三—加入人物解析圖

測試集為 Deep_fashion 訓練集與模型二測試集同樣截至腹部的 852 張，加入人物解析圖一同作為輸入，測試此模型優化方式是否有效提升外擴模型成效。下表由上至下分別為 Deep_fashion 中的測試集原圖、輸入人像圖、輸入的人物解析圖、外擴後的人物解析圖與輸出圖像及其主觀和客觀成績。

1. 模型三的訓練結果展示與客觀成績

表六、圖像外擴模型三的訓練結果

	圖片一	圖片二	圖片三	圖片四
測試集原圖				
輸入人像圖				
輸入人物解析圖				
外擴後人物解析圖				
輸出結果				
主觀評分 (分)	8.1	7.9	7.6	8.0
PSNR (dB)	33.066	33.947	35.790	34.255
SSIM Index	0.814	0.809	0.934	0.824
	圖片五	圖片六	圖片七	圖片八
測試集原圖				
輸入				

輸入人物解析圖				
外擴後人物解析圖				
輸出結果				
主觀評分 (分)	7.6	7.8	7.7	8.0
PSNR (dB)	33.369	33.549	32.793	32.247
SSIM Index	0.890	0.761	0.831	0.656

2. 模型三的主觀評分

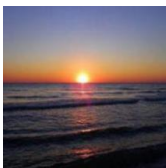







	總分	人像	服飾	姿勢	紋理
平均	7.8 分 (9 分)	2.6 分 (3 分)	1.6 分 (2 分)	1.7 分 (2 分)	1.9 分 (2 分)


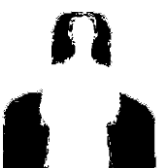

三、圖像填補結果





(一) 圖像填補的訓練結果展示與客觀成績

以下為圖像填補的結果，由上到下分別為：Places2 測試集原圖、Places2 圖像結合人像遮罩（模擬實際運用時模型的輸入圖像）、模擬實際運用時的差集填補遮罩、差集填補結果、無差集填補結果以及其客觀成績。

表七、圖像填補效果

	圖片一	圖片二	圖片三	圖片四
測試集原圖				
輸入圖像				

差集填補遮罩				
差集填補結果				
PSNR (dB)	35.623	34.605	32.092	36.857
SSIM Index	0.891	0.904	0.790	0.925
無差集填補結果				
PSNR (dB)	30.558	29.971	29.664	32.154
SSIM Index	0.647	0.653	0.506	0.746
	圖片五	圖片六	圖片七	圖片八
原圖				
輸入圖像				
差集填補遮罩				
差集填補結果				
PSNR (dB)	35.525	37.070	33.053	33.401
SSIM Index	0.871	0.921	0.796	0.876

無差集填補 結果				
PSNR (dB)	29.961	32.228	30.271	31.040
SSIM Index	0.556	0.780	0.534	0.758


(二) 圖像填補平均客觀成績

	PSNR(db)	SSIM
差集填補	34.778 (dB)	0.872
無差及填補	30.730 (dB)	0.648

四、整體模型效果

測試集為 Deep Automatic Portrait Matting Dataset，總共 540 張，皆為單一人像在各地點的照片，人像包含腹部以上的自拍照片和其 Alpha Map，年齡、人種不拘。結合修正與訓練過後的圖像擷取、圖像外擴、圖像填補三個模型，實作改善人像比例過大的自拍照片。因輸出並沒有原圖可以比對，所以並沒有客觀評分的數值可以呈現，只以主觀的方式評分。

表八、整體模型輸出結果

	圖片一	圖片二	圖片三	圖片四
原圖				
輸出結果				
主觀評分(滿分 9 分)	5.8	5.0	4.9	6.0
清晰度	1.7	1.4	1.6	1.9
人像邊界銳利度	1.9	1.5	1.5	1.6

整體合理性	2.2	2.1	1.8	2.5
	圖片五	圖片六	圖片七	圖片八
原圖				
輸出結果				
主觀評分(滿分 9 分)	5.1	4.6	4.5	4.0
清晰度	1.4	1.5	1.4	1.2
人像邊界銳利度	1.4	1.4	1.2	1.3
整體合理性	2.3	1.7	1.9	1.5




陸、討論

一、探討圖像擷取模型的效果

(一) 比較 Mask_RCNN 和 Closed Form Matting 模型輸出效果

本研究結合 Mask_RCNN 和 Closed Form Matting 的輸出結果以達到最佳的人像擷取效果，以下為三者的比較分析：

表九、Mask_RCNN 和 Closed_Form_Matting 模型輸出效果比較

	Mask_RCNN	Closed_Form_Matting	本研究(二者結合)
輸入	原圖	原圖、Trimap	兩模型輸出遮罩
輸出 (人像切割遮罩)			
特色	輸入簡單快速	人像切割細緻自然	切割結果最佳
Pixel Accuracy 平均	0.874	0.820	0.875

（二）分析 Closed Form Matting 的切割結果

從評分結果可見運用 Closed Form Matting 無法改善切割效果，以下幾點是我們認為無法改善的原因：

1. 在 Mask_RCNN 和 Closed Form Matting 的表現上，可以明顯看出 Mask_RCNN 是直接框出人像沒有漸層的效果。而 Closed Form Matting 利用不透明度遮罩會有漸層的效果。然而 Closed Form Matting 的漸層效果在圖片底部會造成誤判使不該有漸層處亦出現漸層。
2. 由於 Closed Form Matting 的 Trimap 是經由 Mask_RCNN 生成的，因此 Mask_RCNN 的誤差會間接影響到 Closed Form Matting 的結果。我們只能利用開運算擴大 Trimap 不確定區域減少其誤差。

（三）Mask_RCNN 和 Closed Form Matting 的結合結果

由於 Mask_RCNN 和 Closed Form Matting 各有優缺點，因此我們透過研究理論探討所述的方法將二者結合。在結合 Mask_RCNN 遮罩肩膀以下的明確邊界和 Closed Form Matting 遮罩頭部的細緻邊界後得到最佳的切割結果。從 Pixel Accuracy 來看二者結合的效果有些微進步，而相比分數相當的 Mask_RCNN，二者結合的切割細緻度大幅提升（此為 Pixel Accuracy 無法評斷的結果）。然而圖像左右兩側邊界如果有碰到人像，其框選結果仍然會出現些微誤差。但是在實驗中，第一次的切割人像會做外擴且此切割誤差微小，因此對實驗結果的影響不大。

二、分析圖像外擴模型的訓練效果

（一）探討不同前置處理對圖像外擴模型成效的影響—模型一與模型二

本研究總共訓練四次圖像外擴模型，分別以 ATR_dataset 的全身人像、Deep_fashion 的全身人像、Deep_fashion 的半身人像至臀部（以下稱作模型一）、Deep_fashion 的半身人像至腹部（以下稱作模型二）作為訓練集。其中以 Deep_fashion 的半身人像作為訓練集的兩個模型學習效果最佳，確實達到輸入肩膀以上人像，輸出延伸至腹部以上的部分圖像意義合理的人像的目標效果。

從受試者主觀評分結果可知，以 Deep_fashion 的半身人像作為訓練集的模型，其在姿勢、圖像紋理方面的表現較佳；在服飾方面表現尚可；在人像方面，其呈現效果在評分的四個項目中是較差的，大多受試者指出輸出結果的下端與人像身體部位的細

微處只有大致的輪廓，相當模糊。

由客觀評分可知，兩模型的輸出結果與原圖的結構相似性相當高，但每個像素的顏色差異還是存在的，也就是說輸出結果的人像與姿勢是意義合理的，不過其服裝顏色呈現與原圖還是有些微差異。

分別訓練以 Deep_fashion 的半身人像作為訓練集的模型一與模型二，在訓練上兩者皆有一定的困難度與必要性。從模型一的結果中可判斷模型生成至哪些人像部位會有較好的外擴效果，也能利用訓練集更接近真實自拍照的模型二模擬真實自拍照的外擴效果。以整體結果判斷，模型二有較好的外擴效果。

表十、圖像外擴模型一與模型二比較

	模型一	模型二
前置處理	擷取整個上半身	擷取上半身至腹部
訓練時間	8 天	6 天
優點	可生成較大範圍的人體部位，達到測試圖像外擴模型生成人像部位極限	訓練集較接近真實自拍照的人像部位分佈，有較好的外擴自拍照效果
缺點	較容易出現人像邊界模糊、服飾顏色不均等問題	無法生成過多的人像身體部位
PSNR(db)	34.93	34.23
SSIM Index	0.883	0.879
主觀測試結果	7.4	7.6

(二) 探討模型優化對圖像外擴模型成效的影響—模型二與模型三

由客觀評分成績可看出，加入人物解析圖一同作為輸入的模型三與未優化的模型二相比準確率與穩定度明顯提高；由主觀評分成績與輸出圖片可看出，在人像、姿勢兩評分項目有顯著進步，外擴人像輪廓、身體部位清晰且姿勢合理，而在服飾、紋理則沒有太大差別。由表五結果呈現可發現，圖片二和圖片四輸出人像的姿勢與原圖相比有很大的差異，圖片八的服飾與原圖差異也很大，然而兩者都不失其合理性，達到合理延伸的目的。整體而言，模型三的外擴效果優於模型二，大幅提升最終外擴人像的準確率。

表十一、圖像外擴模型一與模型二比較

	整體主觀評分	PSNR((db)	SSIM Index
模型二	7.6	34.23	0.879
模型三	7.8	33.62	0.815

(三) 與原論文圖像外擴的參考模型結果做比較

從觀察原論文訓練模型在全身人像上的效果，因為本研究只需利用到上半身人像即可，故我們判斷上半身效果即可。觀察圖片十九可發現其人像細微處如手的部分有輪廓不明顯的情況；而每張照片在服飾上都會有與我們模型訓練時衣服顏色不均、不對稱的情況；在姿勢和紋理方面，原模型也同樣有較好的表現。

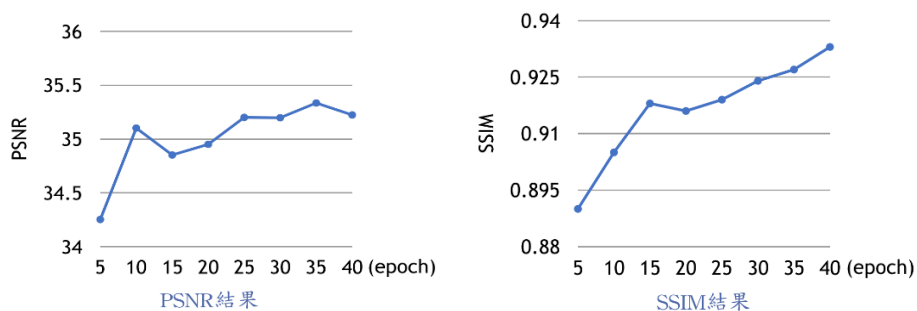
本研究的模型為訓練上半身人像，由於原模型沒有與本實驗相似的訓練模型，所以改為展示以下原訓練模型測試結果：



圖十九、原模型運用於全身人像的效果 [3]

三、探討時間對圖像外擴模型結果的影響

可從圖表中發現 SSIM 的評分隨著訓練時間增長確實有小幅度持續性遞增加，而 PSNR 的評分在訓練初期有些維提升，但後面則持平。推測導致此現象可能的原因為低維度的訓練資料如果在網路中擴展到過多的神經元，資料無法提供足夠多的細節，造成了許多無用、重複的數據，無法有效提升模型效果。



圖二十和二十一、訓練次數對圖像外擴模型效果的影響

四、探討圖像填補模型的訓練效果

將有無差異填補的結果相比較，差異填補所需填補的範圍與無差集填補相比大幅縮小，也因此可從客觀評分中發現差異填補有較好的填補效果，差集填補的評分平均分別為峰值信噪比 34.778，結構相似性 0.872；無差集填補的評分平均則為峰值信噪比平均達到 30.730，結構相似性則有 0.648，由次可見本研究提出的差異填補方法能有效提升背景填補的正確率。

五、探討整體輸出效果

(一) 探討整體模型輸出結果

由整體模型輸出結果（表七）可發現，若外擴的人像和填補後的背景色調太相近，會導致前景人像邊界模糊（如圖片六）。另外，由於圖像外擴生成的人像邊界模糊，會導致圖像擷取時不容易判斷人像邊界因而產生誤差（如圖八）。在訓練時就先將原圖轉為較低像素（128*128）的版本，是為了避免像素太高導致訓練時間過長，因此在貼回填補過後的背景時，縮放後的人像會有失真、模糊的問題，但實際應用上，照片的解析度夠高，縮放的圖像失真問題會被解決。

由主觀評分中可知，人像邊界銳利度偏低，部分評分者表示輪廓有些許粗糙、崎嶇不平；在人像部位清晰度評分中，衣服著色上並不均勻，細看同伴衣服的颜色差異有點大，無法顯示出衣服本身的花紋，不過生成的人像部位清楚，手臂、肩膀和衣服界線明顯；在圖像整體合理性中，多數圖片都有效達到縮小人像、照片結構更均勻的目的，說明本研究制定的一系列步驟確實達到使人像構圖更均勻的目的。

從和圖像外擴原論文相比較中，可以了解圖像外擴並非簡單的深度學習模型，除了訓練集筆數量大外，還有訓練集必須經過明確的前置處理擷取目標物件以及在訓練時的多種參數限制，儘管在圖像外擴效果呈現上還尚有改善空間，本實驗的整體輸出結果達到使自拍照人像構圖合理的目的。

(二) 探討整體輸出結果較不理想的可能原因

1. 以 Deep_fashion 訓練的模型卻是用 Deep Automatic Portrait Matting Dataset 做為測試集，兩者在拍攝角度和人像部位上有一定的差距。
2. 無法以 Deep Automatic Portrait Matting Dataset 作為訓練集，因為訓練集除了要同時

包含遮罩外，也須包含輸出後的完整圖像。

3. 訓練集的前置處理已盡量減少兩訓練集的差距，然而圖像外擴對於遮罩大小有必須為圖像大小因數的限制，侷限了訓練集的身體部位裁切，像是肩膀。

柒、結論

一、總結與運用

- (一) 訓練集與目標物件越相近，模型的學習效果越好，且背景單一、訓練物件清楚明瞭、無多餘物件等條件都有助於學習效果的提升。
- (二) 圖像擷取模型中，Mask_RCNN 可切割出較精確的人像邊界，而 Closed Form Matting 可切割較細微的部位，兩者結合後能達到最佳效果，擷取正確率可達 88%。
- (三) 以 Deep_fashion 半身人像至腹部為訓練集的模型外擴效果優於以 Deep_fashion 半身人像至臀部為訓練集的模型；而加入人物解析圖作為輸入的模型外擴成效優於原模型，故以 Deep_fashion 半身人像至腹部為訓練集並加入人物解析圖作為輸入的模型有最佳外擴效果。
- (四) 差集圖像填補模型的峰值信噪比平均達到 34.778，結構相似性則有 0.872，模型有效地達到目標效果，以利研究目的的達成。
- (五) 將訓練與修正後的圖像擷取、圖像外擴、圖像填補模型結合，將人像比例過大的自拍照經過實際操作後，確實達到使人像在原圖上顯得更均勻的目的。

二、未來展望

- (一) 研究中因為考慮到訓練模型的時間過長所以訓練集約為 4000 張，但訓練比數的資料增加能提升模型學習效果，希望未來探討訓練集比數與模型學習效果的關係。
- (二) 將此技術運用於團體照片上並製作親近使用者的介面，使用者能藉由此軟體自由調整畫面中的人像大小，提升照片後製的靈活度。

捌、參考資料及其他

- [1] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017) Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969)
- [2] Levin, A., Lischinski, D., & Weiss, Y. (2007). A closed-form solution to natural image

- matting. IEEE transactions on pattern analysis and machine intelligence, 30(2), 228-242
- [3] Wang, Y., Tao, X., Shen, X., & Jia, J. (2019). Wide-Context Semantic Image Extrapolation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1399-1408)
- [4] Liu, G., Reda, F. A., Shih, K. J., Wang, T. C., Tao, A., & Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 85-100)
- [5] Stanford University CS231n (2019) Retrieved from <http://cs231n.github.io/convolutional-networks> , 檢索日期：2020/06。
- [6] Preetham Govind K. (2019) Semantic Segmentation Techniques for Computer Vision Tasks. Retrieved from <https://medium.com/@preetham.gs93/semantic-segmentation-techniques-for-computer-vision-tasks-89b2e24e80f9> , 檢索日期：2020/06。
- [7] Fu ShuWu (2019) Mask-RCNN 算法及其實現詳解. Retrieved from <https://www.twblogs.net/a/5cbfa466bd9eee3aed78574a> , 檢索日期：2020/06。
- [8] Brandon (2016) Data Science and Robots. Retrieved from <https://brohrer.mcknote.com/zh-Hant> , 檢索日期：2020/06。
- [9] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Antonio Torralba, Aude Oliva. (2017) , <http://places2.csail.mit.edu/> , 檢索日期：2020/06。
- [10] Li, P, Xu, Y, Wei, Y, & Yang, Y. (2019) Self-Correction for Human Parsing. arXiv preprint arXiv:1910.09777
- [11] 邱泓翔 (2019)。運用 GAN 實現字體風格轉換。2019 年臺灣國際科學展覽會。

【評語】 052501

此作品針對手機自拍像人臉比例過大的問題，提出深度學習的方法，產生結構均衡的照片。所開發的技術包含圖像擷取、圖像外擴、圖像填補、以及圖像結合的技術。此應用想法創新且具實用性，但在有些情況下自動生成的效果不佳且很不自然，未來可以繼續改進。建議在研究的過程中應將不同作法的優劣，加以實驗數據來佐證。另外，所用到的四項技術需整合成一個完整的工具，才具有實用價值。

摘要

手機自拍中常有人臉佔照片面積比例過大的問題，本研究運用深度學習提出改善方法，包含四大項目：圖像擷取、圖像外擴、圖像填補以及圖像結合。圖像擷取結合兩模型優點以生成擷取遮罩；圖像外擴藉由更改模型架構與對訓練集做不同的前置處理，進行人像外擴；圖像填補自行設計差集填補法來提升準確率；最後自動將人像與背景做圖像結合。實驗結果顯示，所設計的方法可將自拍圖像中過大的人像縮小至合理大小，形成結構均衡的照片。

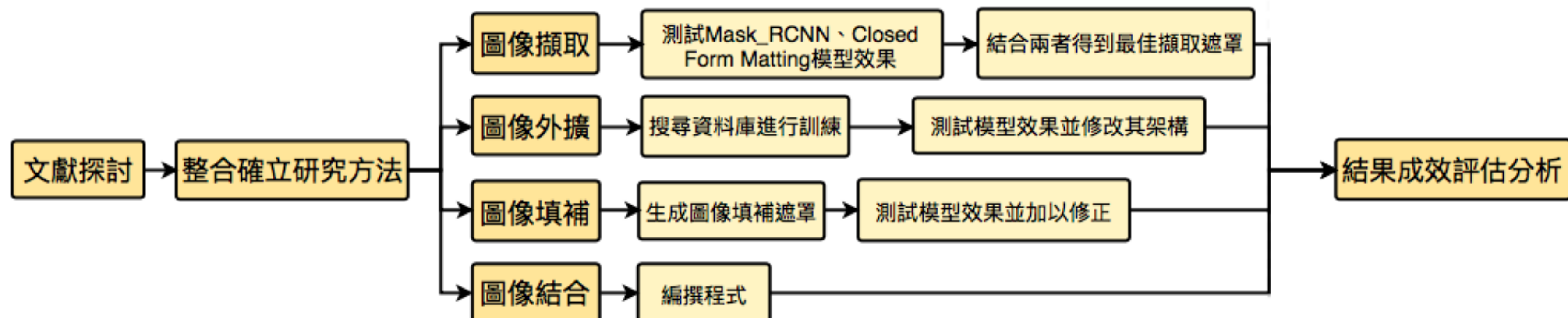
研究設備及器材

硬體：筆記型電腦、GPU：Nvidia Tesla k80、伺服器

軟體：Python（連結各模型的輸入輸出、以自編程式做圖片結合）、TensorFlow、PyTorch

資料集：Deep Automatic Portrait Matting Dataset（圖像擷取與整體模型測試集）、Deep_fashion（圖像外擴訓練與測試集）、Places2（圖像填補訓練與測試集）

研究流程及架構



研究過程及方法

一、研究方法設計

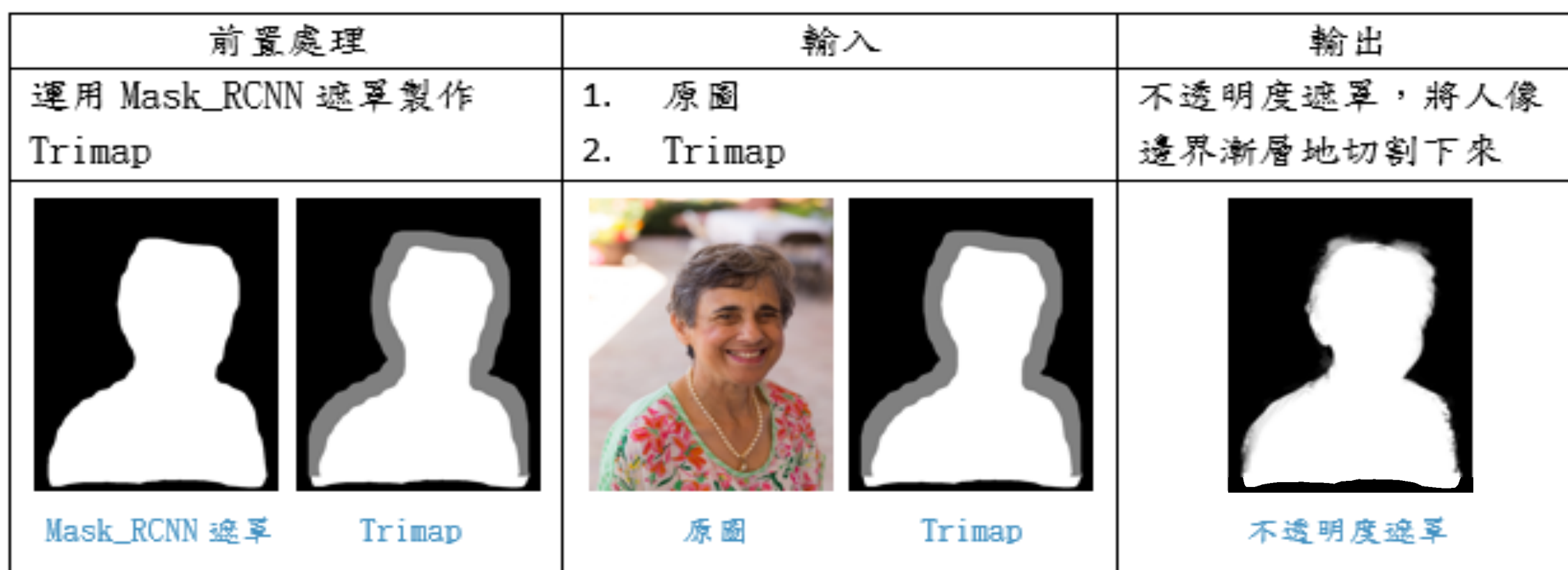
(一) 圖像擷取 Segmentation

1. Mask_RCNN

Mask_RCNN是一種深度學習影像分析技術，可以達成影像實體分割。

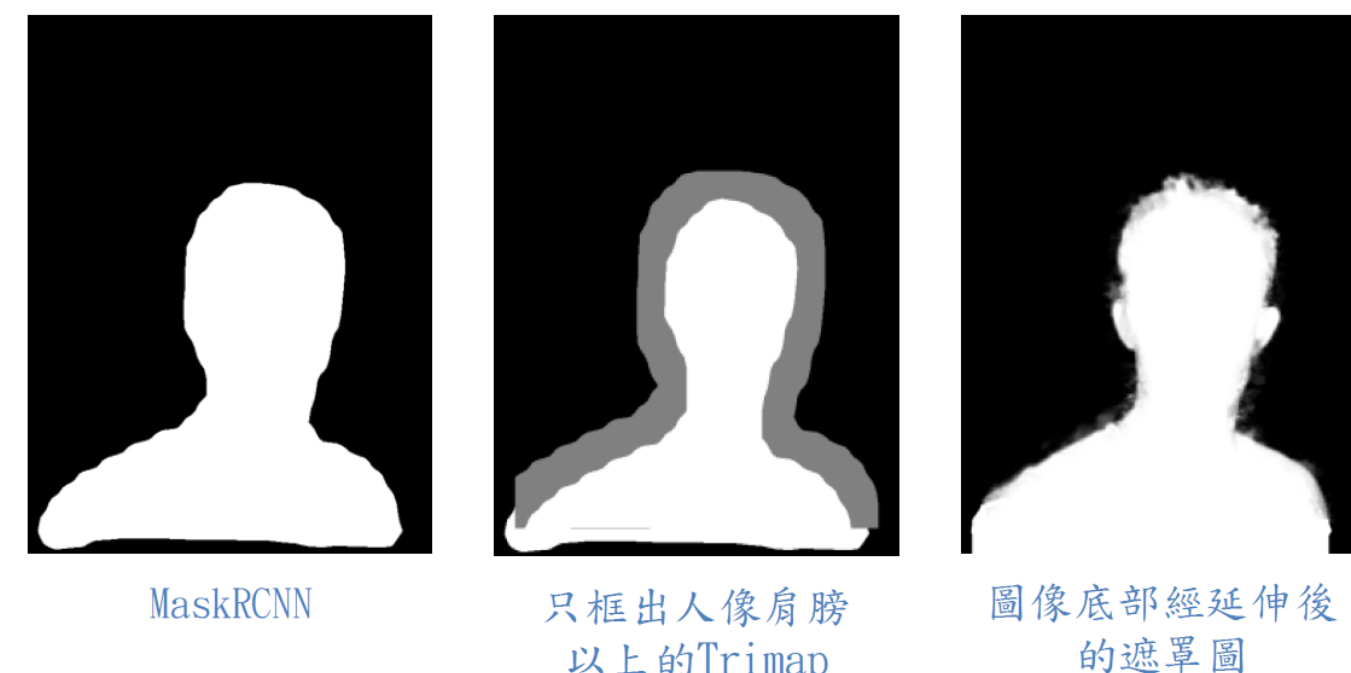


2. Closed Form Matting



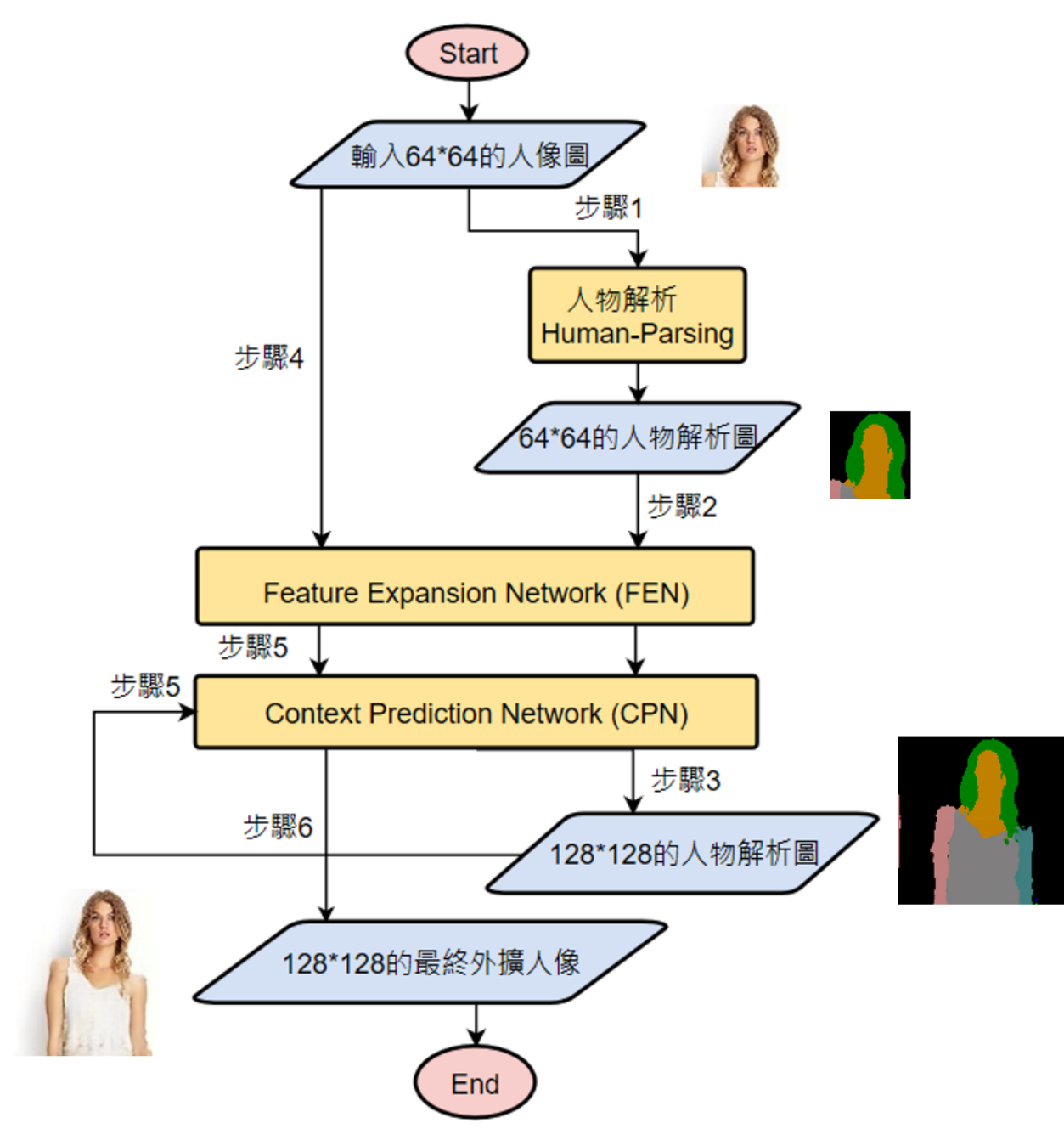
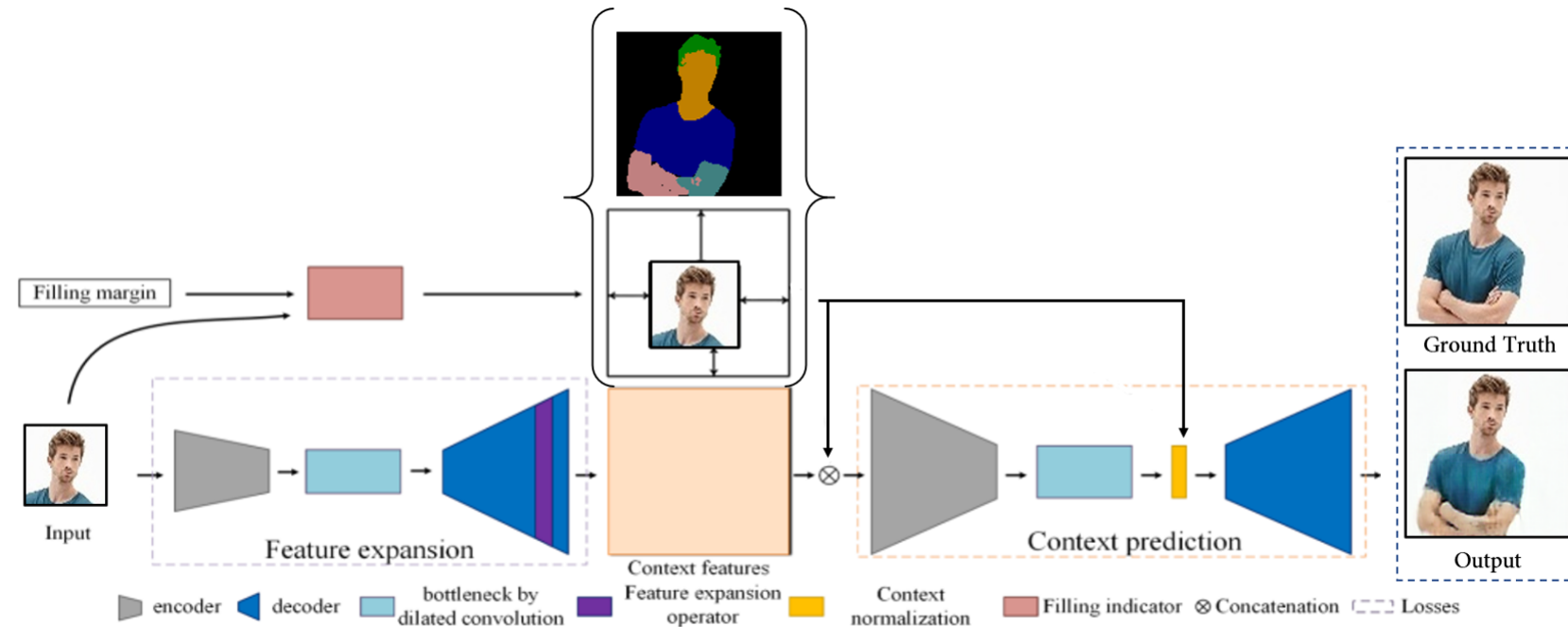
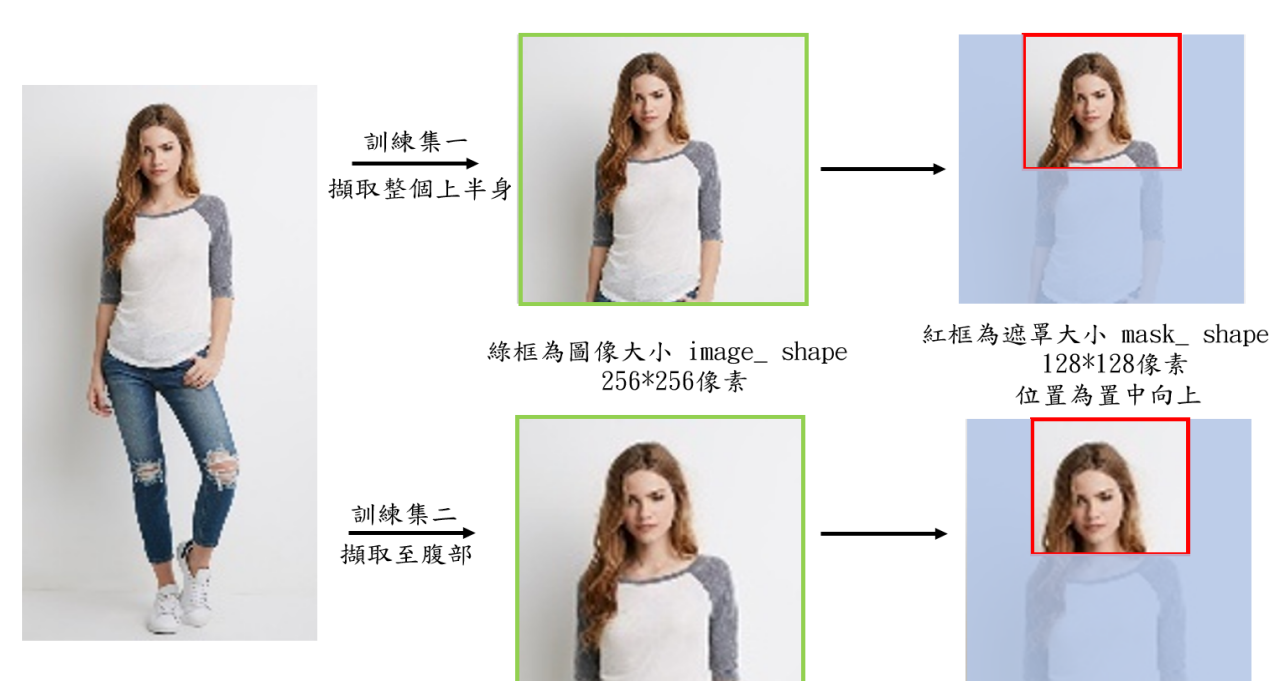
3. Mask_RCNN和Closed Form Matting的結合

- 為保留人像下半部邊界處不漸層，只框出人像肩膀以上的部分做Closed Form Matting。
- 由於Mask_RCNN在人像底部會有少框的現象，因此編撰程式偵測人像底部最左和最右處，將人像遮罩向下延伸至圖像底部。



(二) 圖像外擴 Outpainting

圖像外擴對模型輸入物件不完整的圖片，對圖片做意義合理的外擴，輸出物件完整的圖片。模型的訓練集為Deep_fashion中的全身人像圖，共約5,500張，再對其做前置處理。改動原模型架構以加入人物解析圖，提升外擴準確率。人物解析Human Parsing 為判斷圖片中人像身體部位或服飾並劃分區塊以歸類。



圖像外擴前置處理示意圖（模型一、二）

優化後圖像外擴模型之架構圖（模型三）

圖像外擴模型整體流程圖

(三) 圖像填補 Inpainting

圖像填補結合現實情況和圖像意義以重建圖像空缺部分。本文使用局部卷積神經網路(Partial Convolutions) 作為圖像填補模型。由於其局部卷積層的設定和突破未知區塊的形狀限制，因此有較穩定且良好的填補效果。測試集為Places2，共約15,000張照片，有400個以上的地點分類。

為縮減填補範圍使填補效果更加穩定，我們設計差集填補遮罩，將填補範圍縮小至「原圖背景空缺處減去外擴後人像遮蓋」的範圍，也就是差集填補遮罩的黑色區域。



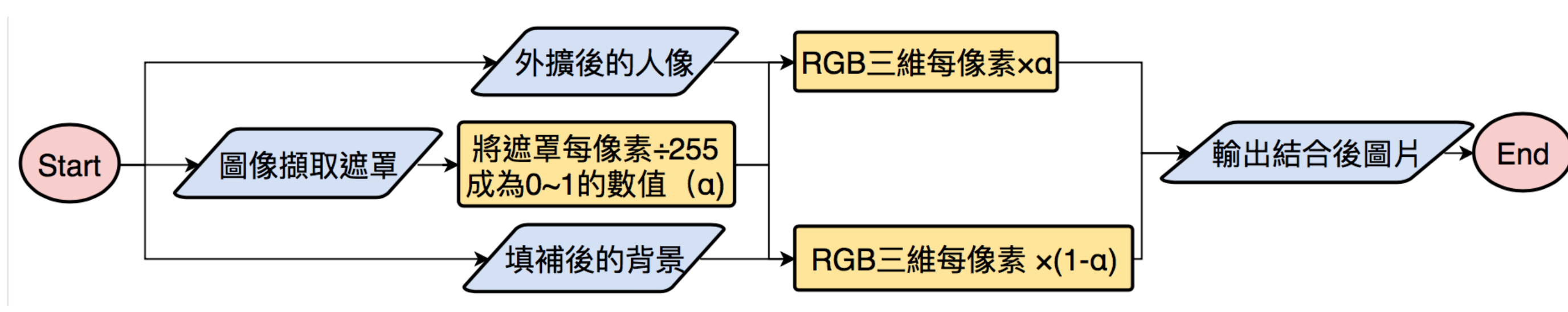
Deep_fashion訓練集展示



差集填補遮罩示意圖

(四) 圖像結合

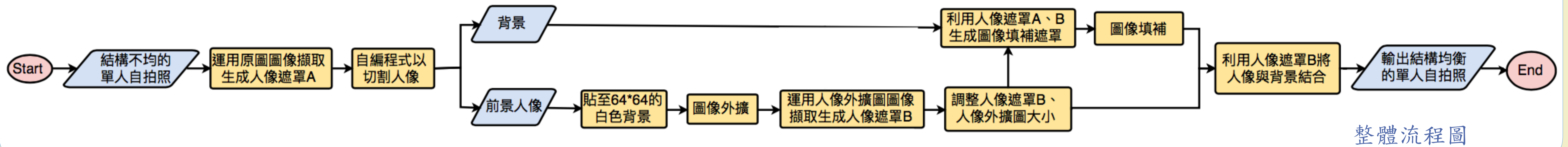
利用下式將人像和背景結合。以公式表示，其中F為前景，B為背景，走訪每個像素i，人像遮照(α)和前景、背景的乘積和為輸出(I)，其公式如右： $I_i = \alpha_i F_i + (1 - \alpha_i) B_i$



模擬自拍照縮放示意圖

二、整體步驟與流程

本研究透過自拍圖片中人像的選取、縮小達成改善自拍人像構圖的目的，以下是我們的流程：



三、成效分析設計

(一) 客觀測試

1. 像素精度(Pixel Accuracy)： $Pixel\ Accuracy = \frac{\sum_i n_{ii}}{\sum_i t_i}$
計算標記正確的像素佔總像素的比例。

2. 結構相似性(Structural Similarity Index, SSIM)：
比較圖片物件結構相似度，數值範圍為-1~1，數值越接近1表示兩者相似度越高。

3. 峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)：
比較每像素的顏色相似度。

$$PSNR = 10 \log_{10} \left(\frac{R^2}{MSE} \right)$$

40dB	非常接近原始圖像
30~40dB	失真可察覺但尚可接受
20~30dB	圖像品質差至不可接受

(二) 主觀測試

由於圖像外擴以及整體模型的效果較難用客觀方式評分，因此我們請30位受試者對生成的8張照片打分。

1. 圖像外擴模型的訓練效果評分標準：

	人像	服飾	姿勢	紋理
0	完全無法辨識輪廓	衣服零碎、不可辨識	人像比例不協調，姿勢怪異	紋理差異大且不和諧
+1	模糊且生成部位不完整	衣服輪廓大致清楚，但花樣不明顯	人像四肢些微不協調，大致和理	紋理大致相同、不突兀
+2	尚可辨識輪廓，但生成部位周邊模糊	衣服輪廓、花樣清楚，且分布在人像上	人像姿勢合理自然	紋理在視覺上和諧一致
+3	人像輪廓完整清晰			

2. 整體模型的訓練效果評分標準：

	清晰度	人像邊界銳利度	整體合理性
+1	模糊不可辨識人像部位	雜訊過多，邊界完全不可辨識	人像、背景大小完全不符合現實情況
+2	有大約部位輪廓但無法看出細節處	可見擷取痕跡但不影響整體完整性	人像、背景合理，但相比原圖並沒有使圖像更均衡
+3	人像部位輪廓清晰可辨，細緻度高	肉眼幾乎沒有看到加工痕跡	人像、背景合理，相比原圖更均衡

研究結果與討論

一、圖像擷取模型成效分析

(一) 部分測試結果展示

測試集為Deep Automatic Portrait Matting Dataset，共2,002張，為單人胸部以上自拍照和人像遮罩。

	圖片一	圖片二	圖片三	圖片四	圖片五	圖片六	圖片七	圖片八
測試集原圖								
Mask_RCNN 輸出結果								
Pixel Accuracy	0.863	0.845	0.836	0.881	0.876	0.896	0.892	0.906
Closed Form Matting 輸出結果								
Pixel Accuracy	0.733	0.788	0.789	0.851	0.811	0.837	0.856	0.895
本研究 (二者結合) 輸出結果								
Pixel Accuracy	0.868	0.840	0.890	0.889	0.875	0.844	0.893	0.900

(二) 本研究所提之方法 (二者結合) 之優點

- 在結合Mask_RCNN的明確邊界和Closed Form Matting的細緻邊界後得到最佳的切割結果。
- 相比分數較接近的Mask_RCNN，二者結合後切割細緻度大幅提升，此為分數無法評斷的優點。

	Mask_RCNN	Closed Form Matting	本研究 (二者結合)
輸入	原圖	原圖、Trimap	二者遮罩
輸出			
特色	輸入簡單快速	切割細緻自然	切割結果最佳
Pixel Accuracy平均	0.874	0.820	0.875

二、圖像填補模型成效分析

(一) 部分測試結果展示

	圖片一	圖片二	圖片三	圖片四	圖片五	圖片六	圖片七	圖片八
測試集原圖								
輸入圖像								
差集填補遮罩								
差集填補結果								
PSNR (dB)	35.623	34.605	32.092	36.857	35.525	37.070	33.053	33.401
SSIM Index	0.891	0.904	0.790	0.925	0.871	0.921	0.796	0.876
無差集填補結果								
PSNR (dB)	30.558	29.971	29.664	32.154	29.961	32.228	30.271	31.040
SSIM Index	0.647	0.653	0.506	0.746	0.556	0.780	0.534	0.758

(二) 本研究差集填補法之成效

- 從SSIM和PSNR評分可知訓練模型的填補效果相當良好，可以有效地達成研究目的。然而在抹除背景部分物件後，要還原完整物件仍有一定的困難性。
- 運用差集填補遮罩後的SSIM和PSNR評分有顯著提升。

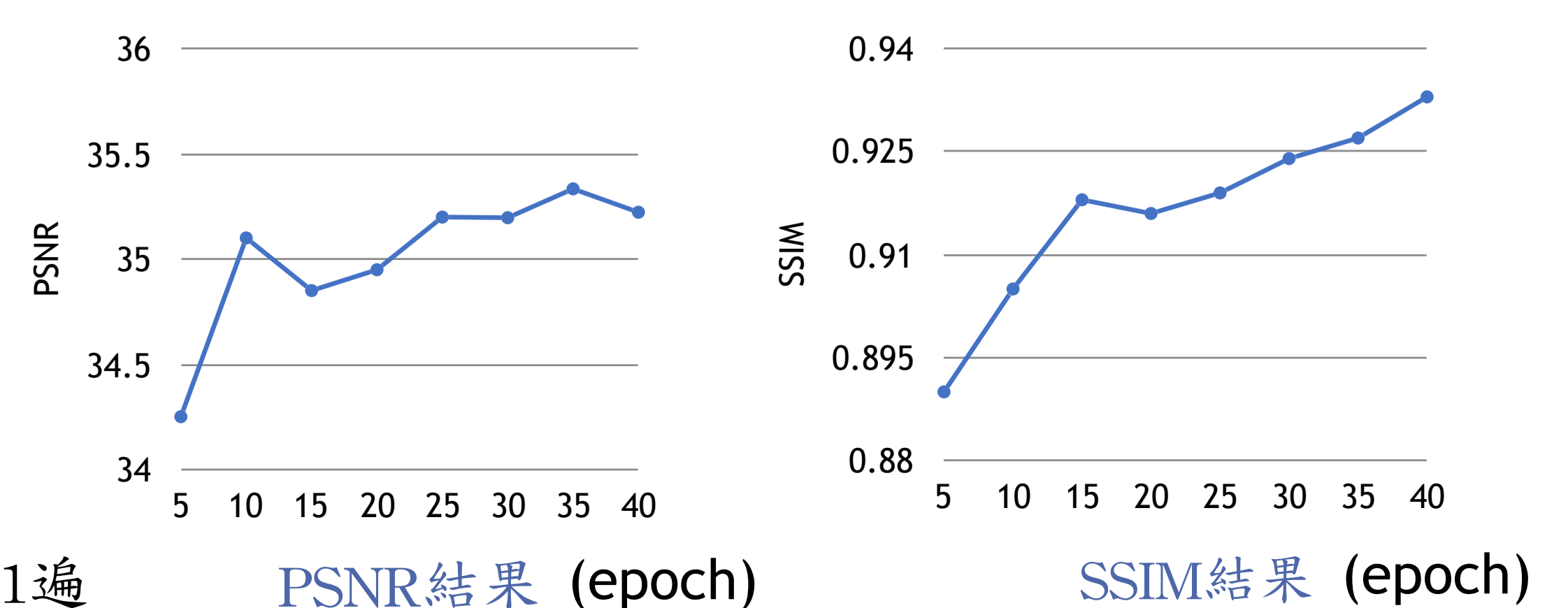
	PSNR(dB)	SSIM
差集填補	34.778	0.872
無差集填補	30.730	0.648

三、圖像外擴模型成效分析

(一) 分析訓練次數對模型影響效果

可從圖表中發現客觀評分成績隨著訓練次數增加有小幅持續性地增長，而在訓練次數到達25epoch時就能有良好的外擴效果。

epoch：完成訓練訓練集中的所有樣本1遍



(二) 模型一：訓練集為人像上半身至臀部

1. 部分測試結果展示

	圖片一	圖片二	圖片三	圖片四	圖片五	圖片六	圖片七
測試集原圖							
輸入人像圖							
輸出結果							
主觀評分(分)	7.8	8.1	7.2	7.3	7.6	7.8	8.5
PSNR (dB)	34.316	35.921	34.284	34.504	35.384	33.688	34.911
SSIM Index	0.884	0.925	0.841	0.850	0.907	0.793	0.922

2. 圖像外擴模型效果的主觀評分

	總分	人像	服飾	姿勢	紋理
平均	7.4分(9分)	2.2分(3分)	1.6分(2分)	1.7分(2分)	1.9分(2分)

(三) 模型二：訓練集為人像上半身至腹部

1. 部分測試結果展示

	圖片一	圖片二	圖片三	圖片四	圖片五	圖片六	圖片七
測試集原圖							
輸入人像圖							
輸出結果							
主觀評分(分)	8.2	7.6	7.3	8.4	7.0	7.9	7.3
PSNR (dB)	36.026	34.296	33.849	35.545	32.511	34.110	33.075
SSIM Index	0.945	0.938	0.893	0.941	0.725	0.888	0.794

2. 圖像外擴模型效果的主觀評分

	總分	人像	服飾	姿勢	紋理
平均	7.6分(9分)	2.5分(3分)	1.5分(2分)	1.7分(2分)	1.9分(2分)

(四) 分析模型一、模型二訓練效果

1. 從主觀評分可知，姿勢、圖像紋理表現較佳；服飾尚可；人像輪廓則待改進。
2. 從SSIM和PSNR評分可知，輸出結果的結構合理，但在服裝顏色呈現上會與原圖有些微差別。
3. 目前結果有人像邊界模糊、人像細微處生成粗糙、服裝顏色不均的問題。

	模型一	模型二
前置處理	擷取上半身至臀部	擷取上半身至腹部
訓練時間	8天	6天
優點	可生成較大範圍的人體部位，供使用者自由選取想要的的身體部位	訓練集較接近真實自拍照的人像部位分佈，有較好的外擴自拍照效果
缺點	由於輸入與自拍照形式不大相同，輸出結果並不穩定	無法生成過多的人像身體部位
PSNR(db)	34.930	33.657
SSIM Index	0.883	0.855
主觀測試	7.4	7.6

(五) 模型三（最終使用模型）：加入人物解析圖並延伸至人像腹部

1. 部分測試結果展示

	圖片一	圖片二	圖片三	圖片四	圖片五	圖片六	圖片七	圖片八
測試集原圖								
輸入人像圖								
輸入人物解析圖								
外擴後人物解析圖								
輸出結果								
主觀評分(分)	8.1	7.9	7.6	8.0	7.6	7.8	7.7	8.0
PSNR (dB)	33.066	33.947	35.790	34.255	33.369	33.549	32.793	32.247
SSIM Index	0.814	0.809	0.934	0.824	0.890	0.761	0.831	0.656

2. 圖像外擴模型效果的主觀評分

	總分	人像	服飾	姿勢	紋理
平均	7.8分(9分)	2.6分(3分)	1.6分(2分)	1.7分(2分)	1.9分(2分)

3. 分析模型三訓練效果

- (1) 由主觀評分成績可看出，在人像評分有顯著進步。
- (2) 優化後的外擴人像輪廓、身體細微處皆較清晰且姿勢合理，而在服飾上顏色更加均勻，整體而言模型三的效果比模型二佳。
- (3) 輸出結果與原圖有相當差異性，然而兩者都不失其合理性，已達到合理延伸的目的。

四、整體模型成效分析

(一) 部分測試結果展示

	圖片一	圖片二	圖片三	圖片四	圖片五	圖片六
原圖						
輸出結果						
主觀評分(滿分9分)	5.8	5.0	4.9	6.0	5.1	4.6
清晰度	1.7	1.4	1.6	1.9	1.4	1.5
人像邊界銳利度	1.9	1.5	1.5	1.6	1.4	1.4
整體合理性	2.2	2.1	1.8	2.5	2.3	1.7

(二) 整體模型輸出效果分析

1. 圖像外擴：

- (1) 圖像外擴的訓練集和測試集不同，兩者在拍攝角度上有一定的差距。
- (2) 為避免訓練時間過長，實驗中將原圖轉為較低像素(128*128)的版本訓練和測試，因此貼回背景後的人像會模糊。然而實際應用時，若訓練的照片解析度夠高便不會有此問題。
- (3) 由於外擴人像的邊界可能模糊，在後續圖像擷取時可能會產生誤差，造成圖像結合時人像邊界有白框。

2. 圖像整合：

從整體合理性評分可知，運用本研究所提之方法可達到使人像構圖更勻稱之目的。

結論

- 一、設計改善方法與流程，透過圖像擷取、圖像外擴、圖像填補以及圖像結合的技術，達到使單人人像自拍照構圖更加勻稱的目的。
- 二、圖像擷取結合Mask_RCNN和Closed Form Matting後正確率可達88%。相較於兩模型個別評分結果，結合後可以得到最佳效果。
- 三、圖像外擴模型以加入人物解析圖並延伸至腹部的效果最佳，證實人物解析圖可提升輪廓清晰度。
- 四、圖像填補模型運用差集填補遮罩的峰值信噪比平均可達34.778dB，結構相似性可達0.872，可以有效地達到目標效果。

未來展望

運用於團體照片上並製作親近使用者的介面，使用者能縮放畫面中的人像大小，提升靈活度。

參考文獻

- [1] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017) Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969)
- [2] Levin, A., Lischinski, D., & Weiss, Y. (2007). A closed-form solution to natural image matting. IEEE transactions on pattern analysis and machine intelligence, 30(2), 228-242
- [3] Wang, Y., Tao, X., Shen, X., & Jia, J. (2019). Wide-Context Semantic Image Extrapolation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1399-1408)
- [4] Liu, G., Reda, F. A., Shih, K. J., Wang, T. C., Tao, A., & Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV)
- [5] Li, P., Xu, Y., Wei, Y., & Yang, Y. (2019) Self-Correction for Human Parsing. arXiv preprint arXiv:1910.09777