**國立屏東科技大學資訊管理系**

**碩士學位論文**

**成對學習應用於偽造影像/視訊偵測**

**Detecting GAN Generated Image/Video based on Pairwise Learning**

**指導教授: 許志仲助理教授**

**研究生: 莊易修**

**中華民國108年12月20日**

成對學習應用於偽造影像/視訊偵測

Detecting Generated Image/Video based on Pairwise Learning

研究生：莊易修　Zhuang, Yi-Xiu

指導教授：許志仲　Hsu, Chih-Chung

**【摘要】**

隨著機器學習的發展，許多基於此領域的技術也逐漸被開發出來，而影像和視訊生成技術沒意外是這領域之中最受矚目的技術之一，例如網路上就有許多網站或 APP 可以生成逼真的偽造人臉。這些偽造人臉通常都是使用生成對抗網絡 (Generative Adversarial Nets, GAN) 的方法生成的，因為 GAN 可以很容易地從低維的隨機向量生成照片般逼真的影像。然而，這些所生成的影像不單純只是有趣，它們也可以是具威脅性的武器。例如偽造的人臉影像可以被用來生成對人不利的報導，而偽造的視訊更可以讓著名政治人物發表會對世界產生危害的演講內容。但是即使想訓練一個能分辨出偽造人臉的辨別器，你也很難收集到所有現今存在的 GAN 來生成訓練影像，因為現在有太多不同的 GAN 可以生成照片般逼真的人臉影像，以往基於學習的方法可能無法有效地檢測出沒收集到的 GAN 所生成的偽造影像。為了克服這個缺點，本論文提出了一種雙步成對學習 (Two-step pairwise learning) 方法來學習不同 GAN 生成的訓練影像上的共同偽特徵 (Common Fake Feature, CFF)。另外，我還提出了一種新的對偶網路 (Coupled Deep Neural Network, CDNN)，以確實學習出偽造或真實影像的局部和全局圖像特徵。最後根據實驗結果，我的方法優於一般常見的偽造人臉影像檢測方法。

*關鍵字*：偽造偵測，生成對抗網路，深度學習對偶網路，孿生神經網路

【Abstract】

Due to the growth of machine learning, many technologies have been gradually development, and the technology of image and video generate must be one of the most watched technology in this field. For example, there are lots of website or APP on the internet that can generate very realistic face images. Most of these images were generated by Generative Adversarial Networks (GANs), because GANs can easily generate the photo-realistic image from a low-dimension random vector. However, these images were not just interested, they could be threatening weapon too. For example, people with bad intention can use fake images to synthesize fake news which are bad for persons who may have a potent effect on society. Moreover, fake videos can make politician make a speech that will threaten the whole world. But its hard to train a discriminator which based on supervised learning to discriminate the fake images because its almost impossible to collect all the GANs on the internet to generate the training images, so the discriminator may be unable to effectively detect the fake image generated by an excluded generative model. To overcome this shortcoming, I propose a two-step pairwise learning approach to learn the common fake feature (CFF) over the training images generated by different generative models. Furthermore, I also propose a novel coupled network to adequately capture the local and global image features for the fake or real images. Experimental results demonstrate that the proposed method outperforms the baseline supervised learning methods for fake face image detection.

**keywords**: Forgery detection, GAN, deep learning, coupled network.

**目錄**

中文摘要…………………………………………………………………ii

目錄……………………………………………………………………vi

圖目錄…………………………………………………………………viii

表目錄……………………………………………………………………x

1. 緒論……………………………………………………………1
2. 研究背景與動機………………………………………1
3. 論文架構…………………………………………3
4. 文獻探討………………………………………………………5
5. 早期的影像偽造偵測技術…………………………....5
6. 外部特徵………………………….……………5
7. 內部特徵………………………………………6
8. 基於深度學習網路之影像偽造偵測技術…………....6
9. VGG16/VGG19...……………………………7
10. ResNet…….…………………………………8
11. DenseNet……………………………………10
12. 數據集之簡介…………………….………….………12
    * 1. GAN…………………………………………12
      2. DeepFake………………………………….…13
13. 研究方法……………………..………………………………15
14. 基於成對學習的偽造人臉偵測網路………………..15
15. 資料預處理…………………………………15
16. 網路架構…………………………….………16
17. 損失函數……….……………………………19
18. 結合對偶網路與三元損失函數之偽造人臉偵測網路..20
19. 網路架構……………………………………20
20. 損失函數……….……………………………22
21. 學習策略之探討……………………………25
22. 用於偵測偽造視訊之網路之探討…………………..26
    * + - 1. EfficientNet………………………………….28
          2. ResNeSt……………………………………...30
23. 評估指標……………………………………………35
24. Precision……………………………………….36
25. Recall…………………………………………..36
26. 實驗結果………………………………………………………37
27. 實驗環境………………………………………..…….37
28. 客觀品質評估……………………………………...…40
29. 主觀品質評估………………………………………...44
30. 偵測偽造視訊實驗結果&分析………………………
31. 結論……………………………………………………………46

參考文獻……………………………………………………………..…46

**圖目錄**

圖1-1、真實和偽造影像之比較…………………………………………3

圖2-1、浮現式數位浮水印…………………………….…………………5

圖2-2、影像雜訊示意圖....…………………………….…………………6

圖2-3、VGG16 網路架構...………………………….…………………8

圖2-4、ResNet20層與56層網路在 CIFAR-10 上的誤差…......………8

圖2-5、殘差學習單元………………………….…...….…………………9

圖2-6、Densely connected Network之示意圖…..……..………………10

圖2-7、由 DCGAN 生成之64x64家具影像……….…..………………13

圖2-8、由 PGGAN 生成之1024x1024偽造人臉影像……… .………14

圖2-9、由 DeepFake 生成之偽造和真實視訊對比圖…………...……14

圖2-10、由 FaceSwap 生成之偽造和真實視訊對比圖……….……14

圖3-1、基於成對學習的孿生神經網路架構應用於偽造人臉偵測…16

圖3-2、孿生神經網路之網路框架………………………………….…18

圖3-3、基於對偶深度神經網路與成對學習之偽造人臉偵測網路架構.21

圖3-4、三元損失函數示意圖………………………………………….23

圖3-5、偽造視訊偵測流程圖…………………………………...………28

圖3-6、各種 Compound Model Scaling 方法之示意圖…………..…...29

圖3-7、EfficientNetB0 之網路架構…………………………………....30

圖3-8、ResNeSt 之網路架構………………………..………………....31

圖3-9、ResNeSt 之 Split Attention block 架構…………..……...……33

圖3-10、ResNeSt 之 radix-major 網路架構………………...……..…35

圖4-1、偽造人臉影像的偽造特徵定位視覺化結果……………..……41

圖4-2、Data Augmentation 效果圖…………………………….………45

**表目錄**

表2-1、ResNet和其他網路之影像識別任務錯誤率…………………...10

表3-1、孿生神經網路架構與參數………………………….…………18

表3-2、EfficientNet 系列各網路在 ImageNet 上之表現……………30

表4-1、孿生網路與對偶網路之神經網路參數與學習參數….…..……38

表4-2、EfficientNetB5 和 ResNeSt101 網路之參數設置……………39

表4-3、偵測偽造影像實驗結果之比較……………………….………40

表4-4、偵測偽造視訊實驗結果之比較…………………...……………42

**第壹章** **緒論**

**第一節** **研究背景與動機**

最近基於深度學習的生成模型像是生成對抗網路 (Generative Adversarial Nets, GAN) 被廣泛用於生產部分內容用於和真實影像/視訊合成，或是製造出完整的影像/視訊。之所以 GAN 會廣泛流行的原因是因為他大幅減少訓練深度學習演算法所需的資料量，而且不只解決了這個問題，還提供一套訓練深度學習演算法的獨特方法，也就是從現有資料建立加上標籤的資料 (在多數情況下就是影像)。 GAN 裡面有兩個需要被訓練的模型 (Model)，一個是判別器網路 (Discriminator)，另一個是生成器網路 (Generator)。生成器網路會試著生產出像真實的偽造影像，而判別器網路則會檢視生成器網路生成的影像並嘗試判別真假。這兩個網路相互從中學習，一個努力找出偽造的影像，另一個則是不斷嘗試製造出無法被正確判斷真假的影像。一旦生成器網路成功製造出判別器網路無法分辨的影像，就代表成功訓練出了一個能夠模擬真正資料分布的影像生成器。

近年來，PGGAN [1] 和BigGAN [2]，甚至可以產生出極高解析度且如照片般逼真的影像或視訊，使得人類無法在短時間內識別出其真實性。雖然許多生成相關的應用像是場景合成或影像翻譯任務 [3] 依賴於這些先進的 GAN，但生成模型的不當使用卻會帶來嚴重的後果，特別是偽造人臉影像的合成。例如，它可以用於在 Facebook 或其他社群網站上創建虛假的個人帳戶並且使用合成的臉部影像 [4] 來混淆某事或欺騙某人，或是生成包含名人的臉的照片來發表對該人不利的新聞。因此，一個有效的偽造人臉影像檢測技術是現今社會迫切需要的。本論文分析了傳統與最新的影像偽造偵測方法的優缺點，發現截至目前為止，並沒有一個有效的方法或系統，可以有效率且正確地偵測出偽造的人臉影像。因此，本論文利用成對學習提出了一套有效的偽造人臉偵測系統。

而偽造視訊和偽造人臉影像則有些許不同，偽造視訊不像偽造影像單純只由雜訊生成一張完整的臉，一般偽造視訊的做法大致分為臉部屬性的操作，例如: 改變臉部的髮色、膚色、年齡…等等，和臉部表情的操作，例如將一個人的臉部表情轉換到另一張臉上。偽造視訊的危險度比起偽造影像有過之而無不及，舉例來說，有心人士可能會利用 Face2Face 將任何著名政治家的臉部轉換到自己製作的演講影片以發表能對社會，政治和商業活動造成嚴重影響的內容 [5] ([圖1-1](#圖1之1))。因此我除了使用本論文提出的人臉偵測系統外也使用了當下最新且在物件辨識領域表現最佳的兩個網路: EfficientNet 和 ResNeSt 來對偽造視訊進行了實驗，在後續章節會分別對兩個網路的架構進行敘述，並將實驗過程和結果呈現出來。

這篇論文的主要貢獻有以下四點:

* 本論文主要提出了一個有效率且易於推廣到未來其他 GAN 生成影像的檢測方法。
* 提出了一種基於新型對偶深度神經網路的偽造臉部影像檢測器，可以成功有效地提取全局和局部影像特徵，提高檢測性能。
* 提出的雙步成對學習 (Two-step pairwise learning) 方法可以偵測到不同的 GAN 所生成的偽造臉部影像，甚至一些在訓練

階段中沒有包括到的 GAN。

* 對偽造視訊的偵測結果進行分析，找到偽造視訊的本質並且歸納出數個能最大幅度改善網路效能的方法。

|  |  |
| --- | --- |
| 擷取 | C:\Users\L.S.S\AppData\Local\Microsoft\Windows\INetCache\Content.Word\1.png |
| 圖1-1、真實影像 (左) GAN 生成之偽造影像 (右) [5] | |

**第二節 論文架構**

本論文主要分為四個章節，第一章為緒論，第二章為文獻探討，第三章為研究方法，第四章為實驗結果。各章節簡單概述如下：

1. 緒論：主要介紹近年 GAN 給社會帶來的利益和威脅和說明本計畫之研究動機。
2. 文獻探討：敘述本論文對從古至今的偽造影像偵測技術的研究整理分析和簡介本論文用來產生數據集之技術。
3. 研究方法：介紹了本論文的重點:成對學習網路和更進階的對偶深度神經網路的架構，以及用於偵測偽造影像的EfficientNet和ResNeSt的網路架構和對這兩個網路進行的改動。
4. 實驗結果：展示本研究之實驗環境和個別網路之參數設定、實驗結果和效能分析最後章節為結論與建議，在這個章節會講述本研究未來預計實現之研究成果與目標。

**第貳章 文獻探討**

1. **早期的影像偽造偵測技術**

早期普遍的影像偽造偵測方法有兩種不同的類型:(1) 外部特徵 (Extrinsic feature) 和 (2) 基於內部特徵 (Intrinsic feature) 的影像偵測器 [6]。以下將分別討論兩種類型。

1. **外部特徵** **(Extrinsic feature)**

基於外部特徵的影像偵測器會採用一個外部標記，例如浮水印([圖2-1](#圖2之1)) [7]，並將其嵌入到原始影像裡。之後，在進行影像偵測任務時將會提取受測影像中的浮水印接著和原始版本的進行比較來確認受測影像是否已被篡改。

|  |
| --- |
| C:\Users\L.S.S\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Visible_digital_watermarking.jpg |
| 圖2-1、浮現式數位浮水印 [7] |

1. **內部特徵 (Intrinsic feature)**

基於內部特徵的影像偵測器會搜索影像中的統計資訊，例如雜訊殘留 ([圖2-2](#圖2之2))[8]。任何影像或視訊的篡改都會導致統計資訊發生變化，使影像可以用這種方式檢測篡改 [6][9]。在 [6] 中，被檢測出的相機感測器雜訊 (Pattern sensor noise) 是用於表示影像中的內部特徵，若照片經過竄改，此類雜訊的分布將會被破壞，利用此特點，論文便可採用貝氏分類器來偵測影像是否被篡改。

|  |
| --- |
|  |
| 圖2-2、由左到右分別為固定雜訊、隨機雜訊、條文雜訊 [8] |

然而這些傳統方法卻難以偵測到完全偽造的影像。例如影像中的內部特徵通常由相機在影像捕獲的過程產生。但是 GAN 可以直接生成偽造的臉部影像。換句話說，從這些影像中無法發現篡改操作，因為影像僅從隨機雜訊生成。所以，傳統的偽造檢測方法將無法檢測到這種生成的影像。因此，現在對於影像偵測的研究都是基於深度學習方法上進行的。

1. **基於深度學習網路的影像偽造偵測技術**

隨著深度學習時代的到來，許多團隊也使用深度學習網路來進行影像偽造偵測。在 [10] 和 [11] 中，他們採用了基本的深度卷積神經網路 (Convolution Neural Network, CNN) 以監督學習的方式檢測輸入影像是真實或是偽造的並且獲得了成果。之後為了更有效的進行影像偽造偵測，許多研究都運用了當時頂尖的深度神經網路，像是VGG16&VGG19 [12]、ResNet [13]、DenseNet [14]，為了在後續歸納這些深度學習網路的弱點以下先介紹這些網路。

1. **VGG16&VGG19**

當時關於 CNN 的改善主要分為兩種方向，一種是使用更小的Convolution filter 或是更小的 stride，另一種主要是利用不同的 Data Augmentation，例如: Multiple Scale training 等。而在 VGG 的論文中主要是將兩個合併在一起，提供一個更深且結果穩定的網路 ([圖2-3](#圖2之3)) [12]。

VGG 最重要的概念就是使用大量 的 Convolution filter。作者認為將較大的 Convolution filter 抽換成較小的 Convolution filter 可減少計算量同時維持獲得的資訊量。此外，使用許多較小的 Convolution filter 可以提高Non-Linearity，同時和大的 Convolution filter 比較下參數較少。另外在 Data Augmentation 方面，VGG 在每次訓練時，從一個固定的亂數範圍中，隨機選取一個數字，並將輸入縮放至那個數字，接著剪裁成所需大小。在測試的部分有使用多個 crop 進行預測，將輸入 rescale 成一個大小，並利用固定的 crop 大小預測左上、右上、左下、右下跟中間，並平均成最後預測結果。

|  |
| --- |
|  |
| C:\Users\L.S.S\AppData\Local\Microsoft\Windows\INetCache\Content.Word\vgg16-1-e1542731207177.png |
| 圖2-3、VGG16網路架構 [12] |
|  |

|  |
| --- |
| C:\Users\L.S.S\AppData\Local\Microsoft\Windows\INetCache\Content.Word\擷取.png |
| 圖2-4、左方為 ResNet56 層(紅線)和 ResNet20 層(黃線)在Cifar10上的訓練誤差曲線圖，右方為兩者在Cifar10上的測試誤差曲線圖 [13] |

1. **ResNet**

當時的深度學習網路普遍都遇到了一個問題: 照理來說深度越深，網絡可以進行更加複雜的特徵的提取，所以當模型更深時理論上可以取得更好的結果。但是當網路達到一個深度後，網路的準確度反而下降了，從[圖 2-4](#圖2之4) 可以看到訓練誤差增加了約 100%，而測試誤差則增加了約 50%。作者為了解決這個問題提出了殘差學習 (Residual learning) ([圖2-5](#圖2之5)) [13]。殘差學習的核心內容是建立層與層之間的Shortcuts connection，這有助於訓練過程中梯度的反向傳播 (Back propagation)，能讓網路深度在不斷增加時，準確度也不會因梯度消失的關係而下降；同時整個網路也更易於最佳化。除此之外 ResNet 的網路架構和 VGG 沒有太大差別，但因為加入了殘差學習單元這個特殊的 block，讓 ResNet 的表現勝過當時頂尖的其他網路 ([表 2-1](#表2之1)) [13]。

|  |  |  |
| --- | --- | --- |
| C:\Users\L.S.S\AppData\Local\Microsoft\Windows\INetCache\Content.Word\擷取.png | | |
| 圖2-5、殘差學習單元 [13] | | |
|  | | |
| 表2-1 ResNet和其他網路之影像識別任務錯誤率 (Dataset:ImageNet) | | |
| method | top-1 err. | top-5 err. |
| VGG (ILSVRC’14) | - | 8.43 |
| GoogLeNet (ILSVRC’14) | - | 7.89 |
| VGG (v5) | 24.4 | 7.1 |
| PReLU-net | 21.59 | 5.71 |
| BN-inception | 21.99 | 5.81 |
| ResNet-34 B | 21.84 | 5.71 |
| ResNet-34 C | 21.53 | 5.60 |
| ResNet-50 | 20.74 | 5.25 |
| ResNet-101 | 19.87 | 4.60 |
| ResNet-152 | 19.38 | 4.49 |

1. **DenseNet**

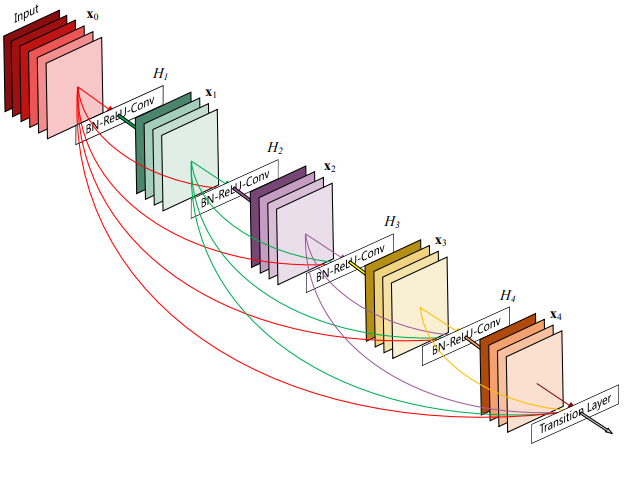


圖2-6、Densely connected Network示意圖 [14]

在 ResNet 發表一年後，基於相同概念的 DenseNet [14] 被提出來了；DenseNet 是一個前饋設計的網路，它的基本概念與 ResNet 一致，但是它不僅僅只是兩層之間的 Shortcuts connection，而是推廣到前後幾層之中，故稱之為這樣的連線為 Densely connected Network ([圖2-6](#圖2之6))[14]，這樣可以直接連接不同層的特徵圖來實現特徵重用 (Feature reuse) 以提升效率。這些特點讓 DenseNet 在參數和計算資源更少的情況下可以實現比 ResNet 更優秀的性能。

然而雖然近幾年的深度學習網路架構不斷改良，但是這些監督式學習網路有一個共同的問題，那就是無法預測沒有出現過的類別。例如，學習辨識貓狗的深度神經網路，無法辨識人車的差異。而在偽造人臉偵測任務裡，有非常多新型的 GAN 可以用來生成照片般真實的偽造人臉影像，像是 [15][16][17][18][19][20][21]。收集所有的 GAN 來學習通用的偽造人臉影像偵測器是非常耗時的，且科技日新月異，無法保證在現今將所有的 GAN 所生成的影像都拿進來學習之後，就可以確定辨識出未來新的 GAN 所生成的偽造人臉影像。因此與其去學習所有 GAN 產生的偽造影像之造假特徵，若我們能引導深度神經網路往另一個層面去學習，亦即學習已收集的 GAN 產生的偽造影像中，共同的偽造特徵是什麼，如此一來，便有可能找出目前 GAN 所生成影像是否都有共通的瑕疵，這樣即便未來出現新的 GAN，也較有可能會隱含這共通的偽造瑕疵存在，便可正確的判別出偽造影像。為了達到此目的，本論文引入了成對學習的概念，並藉由設計獨特的網路架構，來正確的偵測出偽造的人臉影像。

1. **數據集之簡介**
2. **GAN**

本論文採用五個著名的 GAN 來生成需要的影像數據集，分別為:

* DCGAN (Deep convolutional GAN) [18]
* WGAN (Wasserstein GAN) [19]
* WGAN -GP (WGAN with Gradient Penalty) [20]
* LSGAN (Least Squares GAN) [21]
* PGGAN (Progressive Growing of GAN) [1]

DCGAN 是 2015 年發表的 GAN，它是第一個把卷積層運用到 GAN 上的研究。這個運用讓它生成的影像的質量獲得了飛躍性的提升 ([圖2-7](#圖2之7)) [18]，直到現在大部分的 GAN 仍然使用 DCGAN 的網路架構。WGAN、WGAN-GP、LSGAN 就是其中幾個例子，這些 GAN 的網路架構都和 DCGAN 相似，差別只在它們各自採用了不同的損失函數來改進 GAN，並且都因為有效提升了 GAN 的效能而廣為人知。而 PGGAN 是上述這些 GAN 中最晚發表的，它不僅僅修改了損失函數，就連架構也有和上述的 GAN 有較多不同。它的網路架構是動態變化的，網路會先學習生成低解析度的影像，之後在加入新的卷積層讓網路學習生成較高階析度的影像，這樣子循序漸進的學習讓 PGGAN 最終可以生成解析度 且和照片一般真實的偽造影像 ([圖2-8](#圖2之8)) [1]。

1. **DeepFake**

在偽造視訊的部分本論文採用 Google 提供的 DeepFake 資料集做為訓練用的數據集使用。資料集內包含了 1363 部 Google 拍攝的真實視訊、四種偽造技術各以真實視訊為基底製作的共約 7000 部的偽造視訊。Google 使用的四種偽造技術分別為: DeepFake [22], NeuralTextures [23], Face2Face [24], FaceSwap [25]。FaceSwap 和Face2Face 是屬於電腦圖學領域的技術，分別可以做到換臉和換表情的功能。 DeepFake 和 NeuralTextures 則是屬於深度學習領域的技術，功能都是換臉只是使用的方法不同。因為現今網路上看到的偽造視訊大多是利用這四種技術製造而成的，如果辨識網路能在此資料集上得到有效的辨識率的話，相信之後用來對網路上流傳的視訊進行偵測也不會有問題，因此我選擇此資料集做為數據集使用，數據集的範例圖如[圖 2-9](#圖2之9) 和[圖 2-10](#圖2之10) 所示。

|  |  |
| --- | --- |
|  | |
|  | |
| 圖2-7、由 DCGAN 生成之 64x64 家具影像 [18] | |
|  |
| 圖2-8、由 PGGAN 生成之 1024x1024 偽造人臉影像 [1] |



圖2-9 由 DeepFake 生成之偽造和真實視訊對比圖，左圖為假，右圖為真 [22]



圖2-10、由 FaceSwap 生成之偽造和真實視訊對比圖，左圖為假，右圖為真 [25]

**第參章 研究方法**

在本章節中會分別介紹對用於偵測偽造人臉影像的偵測網路的網路架構和運作流程，以及用於偵測偽造視訊的 EfficientNet 和 ResNeSt 的網路架構和測試流程。

**第一節 基於成對學習的偽造人臉偵測網路**

為了學習出所謂的共同特徵，首要概念便是將成對學習的概念引入。成對學習的關鍵想法與傳統監督式學習，要學習出一個分類器可以正確區分影像的類別不同；相反地，成對學習更傾向先學習一個具有鑑別性的特徵，後續再透過其他的分類器來進一步分類。這類方法被廣泛利用在人臉身分的辨識上面，由於人臉影像身分可能非常多，然而每個身分的影像數量很有限，因此會導致學習上的困難。成對學習便是將不同身分人臉影像配對，引導機器學習的模型可以去比較特定身分人臉與其他身分的人臉，來學習出此特定身分人臉與其他身分人臉的特徵區別性。本論文基於此種概念，將偽造人臉偵測的學習任務中，引入成對學習的概念，並提出兩種不同的方法來提升整體偵測之效能。

1. **資料預處理**

每個 GAN 會生產 張影像，把 個 GANs 生成的偽造影像資料集命名為 = [,…,]。接著把真實影像集命名為 = [,…,]，共包含張影像。包含真實和偽造影像的訓練集大小為 *= +*。其影像對應的標註為 ，表示影像為真實 (=1) 或偽造 (=0)。之後為了讓網路能學習好共同偽特徵，我們採取透過成對資訊來引導深度神經網路學習。為達此目的，本論文透過排列組合，可以從訓練集 **X** 及其標籤 **Y** 生成非常大量的成對資訊，即真實-真實影像、真實-偽造影像、偽造-真實影像、偽造-偽造影像等組合。因此，所有的成對資訊 P = [] 將會有 C(2) 對的組合。若對所有的成對資訊都經過深度神經網路來學習，當然可以學習到良好的辨識效果，但需要花費較大量的訓練時間。在我們的實驗中，大約100萬對成對資訊 (i.e., )，就可以在共同偽特徵的學習任務上面取得良好的效能。

|  |
| --- |
| 1. **網路架構**   C:\Users\L.S.S\AppData\Local\Microsoft\Windows\INetCache\Content.Word\4.png |
| 圖3-1、基於成對學習的孿生神經網路架構應用於偽造人臉偵測 |

[圖3-1](#圖3之1)為本計畫提出之第一種偽造人臉影像偵測器的網路架構圖，為了有效的利用成對資訊來引導深度神經網路學習出共同特徵，本論文的共同偽特徵學習網路 採用孿生神經網路 (Siamese Neural Network) 的架構建構而成。孿生神經網路的基本架構如[圖3-2](#圖3之2)，是一個由兩個結構相同，且共享權重的子網路組合而成的網路框架。它的用途在於評估兩個輸入樣本的相似度。在此架構中，網路有兩個學習階段。首先，我們收集了由數個 GAN 生成的大量偽造影像稱為 Example-Generative Model (EG) 和真實影像組合成成對資訊後送進學習網路 中用對比損失 (Contrastive Loss) 進行學習。之後，鑑別器 將連接到 以進一步區分影像。在測試階段， 和 很輕易就能分辨出測試影像的真偽。 和 網路的結構分別呈現於[表3-1](#表3之1)。另外，鑑別器 是直接連接到共同偽特徵學習網路 的第四層。

|  |
| --- |
|  |
| 圖3-2、孿生神經網路之網路框架 |

表3-1、本論文提出之孿生神經網路架構與參數

|  |  |  |
| --- | --- | --- |
| Layers | D1 | D2 |
| 1 | Conv.layer,  Kernel=7\*7,  Stride=4,channel=96 | Conv.layer,  Kernel=3\*3  Channel=2 |
| 2 | Residual block\*2,  Channel=96 | Global average pooling |
| 3 | Residual block\*2,  Channel=128 | Fully connected layer, neurons=2  Softmax layer / Fully convolutional layer |
| 4 | Residual block\*2,  Channel=256 |  |
| 5 | Fully connected layer, neurons=128  Softmax layer |  |
|  |  |  |

1. **損失函數**

在學習過程中我們令特徵為=)，要學習到一對成對影像的共同偽特徵，相當於最小化彼此之間的特徵空間的歐式距離：

, (1)

但是直接最小化 會導致特徵 變成常值映射 (Constant Mapping)。舉例來說，上式最小化的方法之一，便是讓不論什麼影像 **x** 的映射都相同，便會得出 這種理想的情況。然而，這會導致特徵變的無效且無用，因此適當的導入一些正規化的資訊方可避免此問題。本論文在這裡導入了對比損失函數，如下式:

, (2)

一樣為 ，*m* 為事先定義的常數，根據上述式子，如果 (都為假圖或都為真圖)，則特徵 會傾向變得相似於 。如此一來，就可以有效的透過對比函數，來學習到已收集的各種 GAN 產生的偽造影像之共同偽造特徵，而不僅僅是純監督式學習只能學出好的分類效果，但卻無法知道這些特徵是否為共同偽造特徵。當我們得到共同偽特徵後有數種方法可以對接收到的特徵進行分類，例如支持向量積 (Support Vector Machine, SVM)，隨機森林分類器、單純貝式分類器等。在本研究中，我們直接連結卷積層和全連結層到 D1 網路末端，以設計出一個端到端 (End-to-End) 的深度神經網路架構，然後我們採用Cross-Entropy Loss 作為分類器的損失函數：

. (3)

Cross-Entropy之損失函數為許多不同監督式深度神經網路之常見分類損失函數，因此非常適合用來學習最後一層的分類器。藉由反向傳播就可以很容易的訓練好分類器。另外， 中所有層的參數都會因為共同偽特徵的改進而進一步更新，這樣我們相信可以有效偵測到偽造影像。

**第二節 結合對偶網路與三元損失函數之偽造人臉偵測網路**

第一節的方法雖已經結合了成對學習的概念，在後續研究成果的章節亦可達到良好的效能，然而，我們並未對網路架構和學習的本質進行較細緻的分析，因此效能應還沒到最佳化。本節我們進一步的探討深度神經網路的架構設計，並且修改了第一節的網路架構來提升效能。同時為了搭配修改後的網路架構，我們亦更改了損失函數。後續我們將探討損失函數對於網路學習所帶來的影響，並針對對比損失與三元損失函數的比較來了解這些損失函數對學習的效能影響。

1. **網路架構**

|  |
| --- |
|  |
| 圖3-3基於對偶深度神經網路與成對學習之偽造人臉偵測網路之架構 |

我們分析了上一節的網路架構發現了一件事: 以分辨一個二元問題的任務來看，這些網路過於複雜 (真實影像或偽造影像)。同時由於來自不同 GAN 生成的影像的共同偽特徵可能不同，因此傳統的單一 CNN (Convolution Neural Network) 可能無法很好地捕捉到偽造影像的所有偽造特徵。例如，DCGAN [18] 生成的偽造人臉影像可能出現在背景，也可能在臉部區域裡，也無法確定瑕疵的大小是否為大或小。傳統上用 GoogleNet 其 Inception 模組將一層的捲積轉換成同步有 5x5 與 3x3 的架構，多少程度上可以擷取出局部與較全域的特徵。然而，這些知名的網路都相當的深與複雜，在本問題上，不僅僅過於大材小用，更有可能有很多參數會是冗餘的存在。

為了克服這個問題，我們提出了一個較為淺層的神經網路，但分為兩路，分別由 3x3 與 5x5 的捲積核 (Convolutional Kernel) 來完成，稱之為對偶深度神經網路，藉此來學習局部和全域特徵，如[圖3-3](#圖3之3)所示。網路是由三個不同的殘差單元 (Residual Unit) 組成的，第一個殘差單元由兩個 96-Channel 之殘差區塊 (Residual Block) 組成，第二和第三個殘差單元分別由四個 128-Channel 與三個 256-Channel 的殘差區塊所組成。我們將資料輸入此對偶深度神經網路，分別由為 3x3 和 5x5 構成的子網路處理，並在最後合併這些特徵。捲積核大小為 3x3 的網路用於學習局部的共同偽造特徵，5x5 捲積核的網路則是用來學習較為全域的共同偽特徵。為了驗證對偶網路的性能，我們會在下一節與之前的網路做比較。

1. **損失函數**

成對學習的學習方法可藉由另一種三元損失函數來學習具有鑑別性的影像特徵 [26]。相似地，利用三元損失函數學到共同偽特徵後，便可以透過串接一個微型網路來當作是最後的分類器。假定分類器的訓練過程都一樣為 Cross-Entropy Loss 的損失函數，唯一有更改的就是成對學習的損失函數，比較易於探討不同成對學習中損失函數的優缺點。本節首先探討三元損失函數 (Triplet Loss) [26]。

|  |
| --- |
|  |
| 圖3-4、三元損失函數示意圖 |

由於傳統的對比損失函數會使正樣本盡量集中，容易產生過擬合 (Overfitting) 的現象。舉例來說，在對比損失中，有一個重大的影響因素就是其 Margin 值 *m*，此 Margin 決定了所有不相同的樣本的特徵空間之歐式距離至少要大於值，這多少程度假設了所有不同樣本的特徵空間是接近的分布。然而現實的情況這種情況是相當罕見的，大多數的樣本彼此之間特徵空間都會不同，因此要大家一視同仁，只要樣本的標記不同，其特徵的歐式距離就要大於 *m*，對於 *m* 較大的情況來說，會有較多的樣本可能無法滿足此情況；反之在*m* 較小的情況下，有許多標記不同的樣本，彼此之間的距離很容易就可以滿足大於此 *m* 的情況，換句話說，對學習沒有幫助。因此，對比函數很難找到一個最佳化的 *m* 使得多數樣本的情況都可以滿足，導致學習上的困難，而傾向於令正樣本之間距離越近越好就好，導致過擬合的情況發生。

反觀三元損失函數，為了解決上述對比損失函數的缺點，其只要求正樣本彼此的距離小於正樣本和負樣本的距離即可。為了達到這個目標，首先三元損失函數需要先定義一個 Anchor 樣本，一般又稱之為參考樣本。接著，再隨機取出一個與參考樣本相同類別的正樣本 (Positive Sample) 和一個與不同類別的負樣本 (Negative Sample)，因為有三個樣本構成的損失函數，故稱之為三元損失函數，如[圖 3-4](#圖3之4) 所示。有了正樣本與負樣本之後，定義損失函數的基本精神為：正樣本的距離，小於負樣本的距離，換句話說，只要參考樣本與其負樣本的特徵空間之歐式距離比正樣本來得較大，就可以了。這樣的作法，就不受限於不同標記類別的樣本之間，距離一定要大於 Margin 的值。

這裡將三元損失函數套用到網路上: 我們令特徵 = D，其特徵空間的距離可參考式子 (1) 獲得。根據三元損失的函數定義，我們分別定義參考樣本、正樣本、與負樣本之特徵分別為：1) 參考樣本特徵 ，2) 正樣本特徵 ，3) 負樣本特徵 。根據三元損失函數的目標，我們必須最小化參考樣本與其正樣本之特徵空間的歐式距離，同時令其距離小於參考樣本與負樣本特徵空間的歐式的距離。因此，上述的概念可解釋為不等式：

+ < , (4)

其中 為事先定義的常數，與對比損失函數中 Margin 所定義的內涵相似，即希望相同類別樣本之間的特徵距離，可以大過於不同類別樣本之特徵距離一個門檻，即是 。直接最小化不等式有困難，因此改寫成為下列的損失函數，使得不等式為：

(5)

其中+號代表只取上式中大於零的值，而小於零則設定此次損失值為零。根據上述式子，第 *i* 個參考樣本，將會與第 *j* 個正樣本之特徵距離 (即)，將會小於參考樣本與第 *k* 個負樣本之特徵距離 (即)，並至少保持 以上的差距。這樣子的做法，事實上並不要求相同類別的樣本之特徵距離之間要有多近，也不需要強迫不同類別樣本之特徵距離要大於 *m*，最大程度降低了學習上的困難，也有助於效能的提升。

另外，根據樣本的篩選，其實三元損失函數亦有三種狀況，即 Easy Triplet、Hard Triplet, 以及 Semi-Hard Triplet。由於這些並非本論文主要貢獻，因此不多做贅述。本論文選取常見的 Semi-Hard Triplet 的樣本選取策略來最佳化整個深度神經網路。

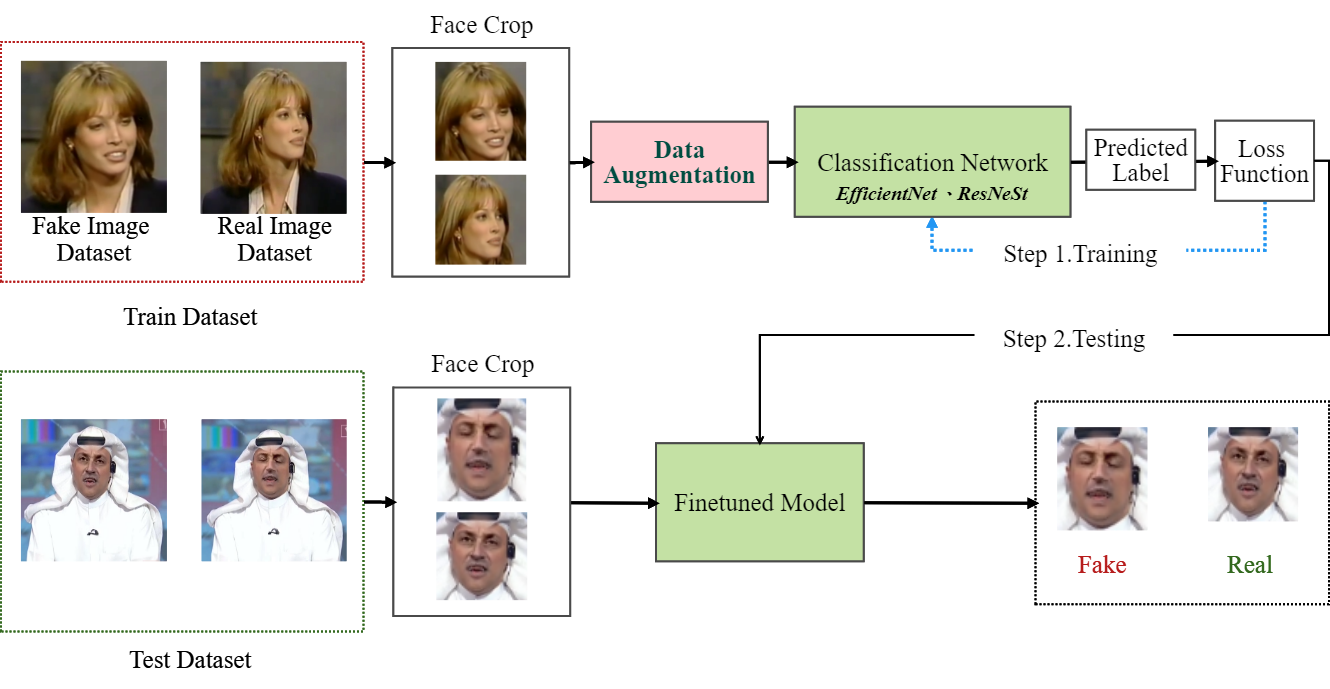
1. **學習策略之探討**

在整個偽造人臉偵測器的學習過程中，主要有兩種學習策略：1) 聯合學習 (Joint Learning) 以及 2) 雙步學習 (Two-Step Learning)。在聯合學習的策略中所有的損失函數會透過一個權重線性組合來得到一個整體損失函數，然而，我們很難在 Cross-Entropy 以及成對學習的損失函數之間找到合適的平衡因子，也比較難確定彼此之間的影響。另一方面，雙步學習方法可以使網路更快的學到偽造影像之間的共同偽特徵，因為我們可以先利用成對學習找出偽造影像之間的共同偽特徵，之後就將第一層網路的參數固定住，之後就只需要對分類器網路 (後面的分類子網路) 中的參數進行學習與更新即可。不過亦可更新第一層網路的部分參數以得到更好的效能，以本實驗為例，我們僅更新第一層網路的特徵擷取層的參數，即倒數第二個卷積層。實驗結果中亦有比較兩者之效能。

1. **用於偵測偽造視訊之網路之探討**

本論文使用 EfficientNet 和 ResNeSt 來嘗試對偽造視訊進行偵測，不使用第二節裡提到的人臉偵測網路進行偵測的原因是因為該網路是基於偵測共同偽特徵來進行判斷，但是偽造視訊並不完全和偽造影像一樣完全由 GAN 產生，因此不同偽造方法間不一定有共同偽特徵可以讓網路學習。就算同個領域的偽造方法之間有類似的特徵可以學習，也很難推廣到不同領域的方法，因此我放棄使用第二節裡提到的網路來嘗試做偽造視訊的偵測。另外不選擇 RNN 這類能學習到時間軸資訊的網路是因為現在的偽造視訊大部分都還是能在畫面中捕捉到原始視訊中沒有的雜訊部分，因此理論上我只要能訓練出一個可以捕捉到這些雜訊的 CNN 網路，那就不需要使用能學習到偽造視訊的時間軸資訊的 RNN 網路了，因為在這個狀況下 RNN 單純只是結構較複雜，效能卻沒比較好的 CNN 網路。最後採用 EfficentNet 和 ResNeSt 兩個網路是因為這兩個是當下 CNN 類型中表現最好的兩個網路，因此最終採用這兩個網路來進行偵測。偽造視訊偵測的流程如[圖 3-5](#圖3之10) 所示，首先我從每部視訊中擷取出固定數量的畫面，之後將圖中人臉部分的畫面擷取出來存成影像之後送進視訊偵測網路進行訓練和測試。網路會送出每一張圖的預測，我會將所有屬於同一部視訊中的人臉影像的預測結果集合起來並且計算網路判斷為真和為假的次數，較多的那方即做為網路對該視訊的判斷結果。以下將分別介紹兩個網路的網路架構，而實驗過程中的各種參數設定會在下一章節展示。

圖3-5、偽造視訊偵測流程圖



1. **EfficientNet**

EfficientNet 系列網路 [27] 是google使用新提出的 Compound Model Scaling 技術再加上為了能讓 Model Scaling 有效發揮而利用 AutoML 領域的 NAS 技術設計出的 baseline network 兩者合而為一所形成的網路。

Compound Model Scaling 是 Model Scaling的一種，傳統的Model Scaling 方法如[圖 3-6](#圖3之5) 所示：a) 為常見網路的架構； b) 為對網路進行 channels 之增加； c) 是對網路層數進行加厚； d) 是對輸入之影像大小進行放大； e) 則是 Google 提出的Compound Model Scaling 方法示意圖。

但是 Google 團隊認為不應該單獨對網路做加寬、加深或增加影像解析度，將複數作法因不同任務所需求綜合優化使用，能比對單一維度 (寬度、深度、解析度) 進行調整能更加有效的提高效能，同時避免過多的參數量。另外 baseline network 的好壞很大程度也會影響到 Model Scaling 的表現，因此Google 團隊也利用 NAS 技術去開發一系列不同尺度的 baseline network 來和 Compound Model Scaling 技術做搭配，最終得出一系列的網路命名為 EfficientNetB0-B7。B0 的網路架構圖如[圖3-7](#圖3之6)所示，圖中的 MBConv 和一般的 Conv 層不同，它指的是 MobileNetV2 中使用的 Inverted residual block [28]，和一般的CNN 層相比起來特徵學習更加有效。而 B1-B7 的網路都是基於 B0 的網路架構做擴張而得到的。各系列網路在 ImageNet 資料集上的表現則如[表3-2](#表3之2)所示。雖然其參數量依照網路編號由小至大逐漸增加，但在精確度上也有相對應的提升。本研究基於實驗環境的限制上使用 EfficientNetB5 來進行實驗。



圖3-6、各種 Compound Model Scaling 方法之示意圖

|  |
| --- |
|  |
| 圖3-7、EfficientNetB0 之網路架構 |
| 表3-2、EfficientNet 系列各網路在 ImageNet 上之表現 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Top-1 Acc | Top-5 Acc | Params | FLOPS |
| EfficientNet-B0 | 76.3% | 93.2% | 5.3M | 0.39 B |
| EfficientNet-B1 | 78.8% | 94.4% | 7.8 M | 0.70 B |
| EfficientNet-B2 | 79.8% | 94.9% | 9.2 M | 1.0 B |
| EfficientNet-B3 | 81.1% | 95.5% | 12 M | 1.8 B |
| EfficientNet-B4 | 82.6% | 96.3% | 19 M | 4.2 B |
| EfficientNet-B5 | 83.3% | 96.7% | 30 M | 9.9 B |
| EfficientNet-B6 | 84.0% | 96.9% | 43 M | 19 B |
| EfficientNet-B7 | 84.4% | 97.1% | 66 M | 37 B |

1. **ResNeSt**

ResNest [29]是 ResNet 的最新變種，是採用了 ResNext 的group convolution 和 SENet 的 split attention block 的概念融合而成的網路，其網路架構如[圖3-8](#圖3之7)所示。

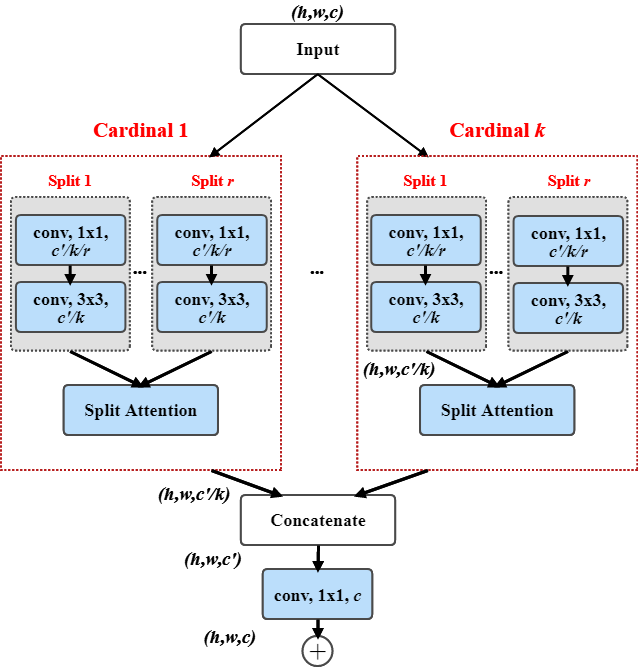


圖3-8、ResNeSt 之網路架構

在這個架構圖中 ResNeSt 首先將長度為 、寬度為維度為 *c* 的輸入分成 *k* 個，每一個命名為Cardinal 1-k，每個 Cardinal 會在拆分出 *r* 個 Split，Cardinal 中的每個 Split 會將輸出的特徵圖送進 Split Attention block 讓網路學到特徵圖之間的關係，最後再將每個 Cardinal 輸出的特徵圖疊加在一起後做一個 conv 調整 channel 後和輸入相加完成殘差學習。

ResNeSt 的 Split Attention block 的架構如[圖3-9](#圖3之8)所示，首先複數個輸入經過相加後會做一個 Global pooling 轉換為實數，之後經過兩次全連接層的非線性變換讓網路學習不同 channel 之間的關係，最後獲得的數字會做為權重和輸入的特徵圖相乘，這代表每個特徵圖有不同的重要度，理想情況下網路會學到每個 class 會在不同的 channel 表現出重要的特徵並且知道應該要注意的 channel。這個 Split Attention block 和 SENet 的 Split Attention block 相似，差別在 ResNest 的 Split Attention block 的輸入有複數個，SENet 的 Split Attention block 的輸入只有一個，同時因為 ResNeSt 的 block 的輸入有複數個，因此會得出複數個權重用來對每個 Input進行相乘。

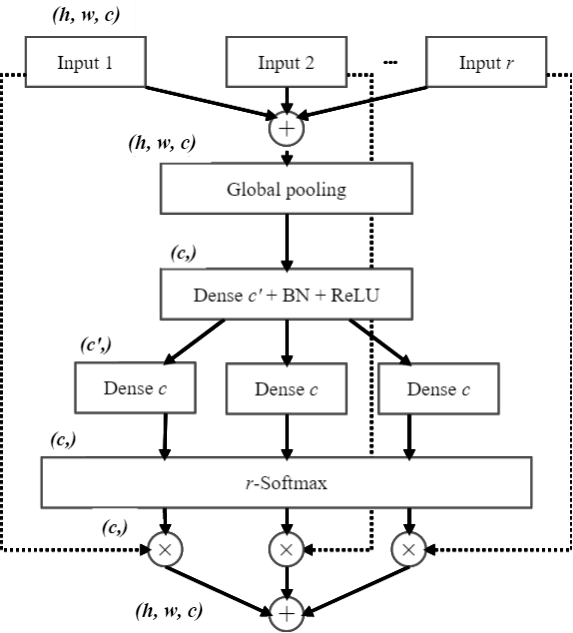


圖3-9、ResNeSt 之 Split Attention block 架構

[圖3-8](#圖3之7) 的 ResNeSt 為 Cardinality-major 的架構，這種架構比較直觀可以了解 ResNeSt 裡的 Split-Attention 部分的概念，但是並不是作者在實做程式時使用的架構。作者實際上是以 Radix-major 的架構 ([圖3-10](#圖3之9)) 的方式實做的，這種架構和 Cardinality-major 的架構出的結果完全一樣，但是比較容易模組化和加速。

[圖3-10](#圖3之9)的架構圖中一樣把輸入分成 *r* 個 Split 和 *k* 個 Cardinal，只是這裡是將每個 Split 在拆分成數個 Cardinal，接著把所有 Cardinal 裡的 , channel = 數量的 conv 統一成一個 , channel = 的層，這種作法比起一次做 個 conv 快上許多，並且輸出結果是一樣的。接著到了下一層 conv 時在每個 Split 裡將 channel分為 個進行卷積，只是這邊不是單純一次做 *k* 個 卷積，而是使用 group convolution 來做，這樣也比普通的卷積快上許多。之後在將 Split 內所有的 Cardinal 的輸出結果疊加在一起將 channel 還原回 送出。各個 Split 輸出的特徵圖會經過一次相加讓特徵圖的空間維度融合然後送進 Split Attention block，在 block 的全連接層部份會分成*k*個 group 來做，因為需要得出 *k* 組權重來對每個 Cardinal 中的特徵圖做相乘，最後乘上權重的特徵圖會再經過 conv 將 channel 數調整成和 Input 相同後和 Input 相加完成殘差學習。本研究基於實驗環境的限制上使用 ResNeSt101 進行實驗。

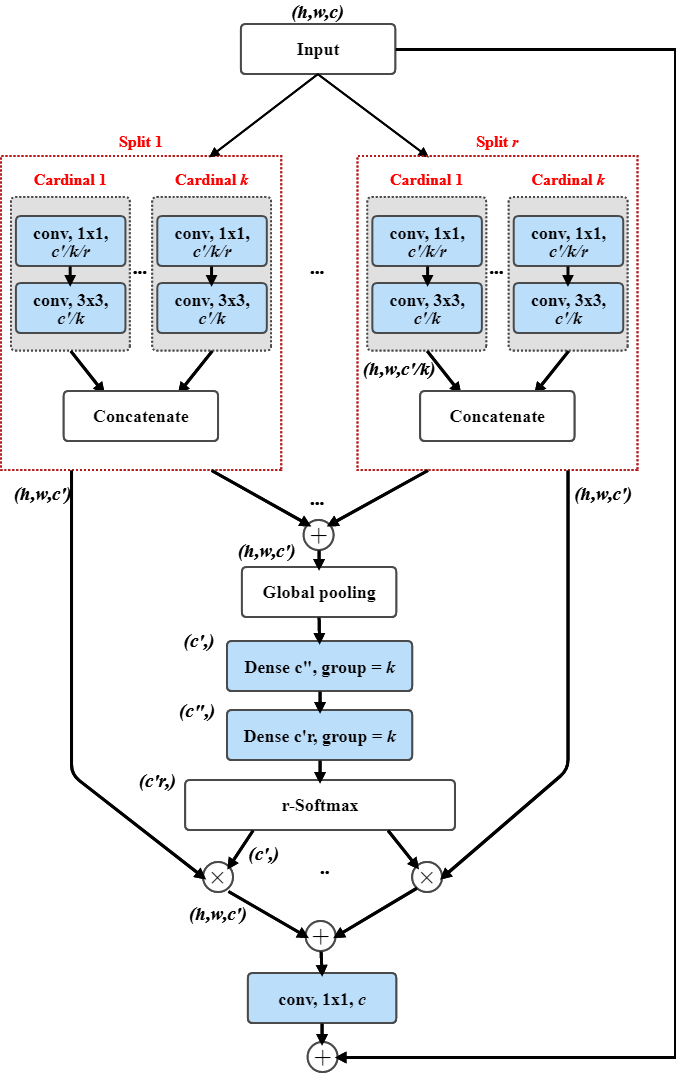


圖3-10、ResNeSt 之 radix-major 網路架構

1. **評估指標**

本論文採用 Precision 和 Recall 做為評估指標。TP 代表實際為真且判斷為真的數量，FP 代表實際為假但判斷為真的數量，FN 代表實際為真但模型判斷為假的數量。

1. **Precision**

Precision 是用來計算所有實際偽造影像中有被準確預測出的偽造影像之比例。

(6)

1. **Recall**

Recall 是用來計算所有被預測出為偽造影像占實際偽造影像之比例。

(7)

**第肆章 實驗結果**

1. **實驗環境**

在偽造影像的部分本研究結果之數據集使用 CelebA [30]。它包含了多樣表情與多元背景，總共有 10,177 個身分，總計有 202,599 張人臉影像。另外研究中我採用五個最知名的 GAN 來生成相對於 CelebA 數量的偽造影像，分別為:

* DCGAN (Deep convolutional GAN) [18]
* WGAP (Wasserstein GAN) [19]
* WGAN-GP (WGAN with Gradient Penalty) [20]
* LSGAN (Least Squares GAN) [21]
* PGGAN [1]

每個 GAN 會生成 200,000 張 的偽造影像。在 PGGAN中，我採用作者發佈的最佳模型然後將生成的影像壓縮到相同的解析度，以利進行公平的比較。之後從中隨機選取 192,599 張偽造影像加入訓練集。最後，我有 385,198 張訓練影像和 10,000 張測試影像。研究方法第一節與第二節之網路設定參數可參考如[表 4-1](#表4之1) 所示。

表4-1、孿生網路與對偶網路之神經網路參數與學習參數

|  |  |  |  |
| --- | --- | --- | --- |
|  | 孿生網路設計 | 對偶網路設計 | |
| Learning rate | 0.001 | | |
| epoch | 15 |  | |
| Contrastive loss的  margin value *m* | 0.5 |  | |
| Triplet loss的  margin value α |  | 0.5 | |
| Optimizer | Adam Optimizer | | |
| Batch size | 32 | | |
| 特徵學習網路的epoch |  | | 4 |
| 分類器的epoch |  | | 11 |

在偽造視訊的部分本研究採用 Google 發布的 DeepFake 做為訓練集，它包含了從 Youtube 取得的真實視訊 1000 部和用四種偽造方法對這1000 部視訊進行修改得出的視訊 4,000 部，另外還有拍攝演員在各種模擬情境下演出的視訊 363 部以及對這 363 部視訊進行偽造的視訊共 3068 部。視訊的長度從 11 秒到 40 秒不等，解析度也從 到 不等。我從每部視訊中以每 30 偵一張，最大 15 張的方式擷取出影像，之後再將影像中的人臉部分的區域以 的大小儲存下來做為訓練集使用。測試集的部分我使用 Deepfake Detection Challenge 比賽提供的資料集做為測試集 [31]，它是由 AWS、Facebook、Microsoft 等公司在 kaggle 平台上舉辦的比賽。這個數據集有 16154 部真實視訊和 96992 部對真實視訊做修改得到的偽造視訊。每部視訊的長度皆為 10 秒，解析度皆為 ，之後我對這些影片使用一樣的擷取方法來產生出測試集。另外因為訓練集的偽造視訊的數量遠大於真實視訊，因此我限制網路在訓練時只會讀取和真實視訊相同數量的偽造視訊進網路以增加訓練的效率。最後我對使用的兩個網路的修改部分如[表 4-2](#表4之2) 所示。

表4-2、EfficientNetB5 和 ResNeSt101 網路之參數設置

|  |  |  |  |
| --- | --- | --- | --- |
|  | EfficientNetB5 | ResNeSt101 | |
| Learning rate | Adam:0.001 / SGD:0.01 | | |
| Learning rate decay | PolyLR | | |
| Epoch | 30 | | |
| Label smoothing | 0.1 | | |
| Optimizer | Adam Optimizer / SGD Optimizer | | |
| Batch size | 16 | | |
| Batch Normalize | SyncBN | | |
| ResNest101的k |  | | 64 |
| ResNest101的r |  | | 2 |

1. **客觀品質評估**

表4-3、偵測偽造影像實驗結果之比較 (Recall / Precision Rate)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Method*/*Test target | LSGAN | | DCGAN | | WGAN | | WGAN-GP | | PGGAN | |
| precision | recall | precision | recall | precision | recall | precision | recall | precision | recall |
| Method in [6] | 0.205 | 0.580 | 0.253 | 0.774 | 0.235 | 0.673 | 0.242 | 0.604 | 0.222 | 0.862 |
| Method in [10] | 0.819 | 0.528 | 0.848 | 0.790 | 0.817 | 0.822 | 0.816 | 0.679 | 0.798 | 0.788 |
| Method in [11] | 0.833 | 0.725 | 0.812 | 0.833 | 0.840 | 0.809 | 0.826 | 0.733 | 0.824 | 0.838 |
| **Proposed-I (Contrastive Loss)** | **0.947** | **0.922** | **0.871** | **0.844** | **0.838** | **0.847** | **0.818** | **0.835** | **0.926** | **0.918** |
| Baseline-I | 0.921 | 0.915 | 0.887 | 0.831 | 0.860 | 0.855 | 0.822 | 0.837 | 0.919 | 0.898 |
| Baseline-II | 0.939 | 0.929 | 0.878 | 0.851 | 0.840 | 0.863 | 0.845 | 0.844 | 0.922 | 0.928 |
| Baseline-III | 0.845 | 0.785 | 0.796 | 0.816 | 0.833 | 0.799 | 0.819 | 0.805 | 0.835 | 0.854 |
| **Proposed-II**  **(Triplet Loss)** | **0.981** | **0.956** | **0.986** | **0.986** | **0.895** | **0.881** | **0.876** | **0.881** | **0.951** | **0.936** |

為了證明成對學習的效能，我輪流將其中一種 GAN 所產生的影像從訓練集中移至測試集，因此在訓練過程中，深度神經網路並沒有見過這個類型的 GAN 之偽造影像，更無法學習到此 GAN 所生成影像的偽造特徵。一個有用的偽造偵測網路，應該要能偵測到訓練集裡不包含的 GAN 產生的偽造影像，即推廣性。實驗過程中我也採用了基於雜訊偵測器的傳統偽造影像偵測方法 [6] 進行性能比較，以及利用深度神經網路來偵測偽造人臉影像的研究 [10][11]。此外，我亦將提出的方法做不同變化產生不同的基礎模型 (Baseline)。Baseline-I包含了成對學習的概念，利用三元損失函數來學習共同偽造特徵，但採取聯合學習的方法，其中不同的損失函數用了相同的權重。Baseline-II 則將對偶網路改為僅有包含其中 一個網路，即與第三章第一節方法相似，但成對學習的損失函數為三元損失。Baseline-III 則是移除三元損失，以純監督式學習的方法來得到偽造人臉偵測器。本論文第三章第一節提出的基礎方法標記為 Proposed-I，而第三章第二節之進階方法，含三元損失函數、對偶網路設計、雙步學習為最終完整的方法，標記為 Proposed-II。

實驗結果證明最終方法的效能在各個方面都優於基礎方法，進一步將Baseline-I、Baseline-II和基礎方法比較也能證明三元損失函數和雙步學習兩個更改的部分都能更加提升網路對於共同偽特徵的學習能力。另外Baseline-III說明了即使使用了對偶網路設計的網路架構並且採用雙步學習的學習策略，由於缺少了三元損失函數帶來的修正，網路完全無法從兩個輸入中學到彼此之間的共同偽特徵，導致效能和一般的監督式學習網路相同。

1. **主觀品質評估**

|  |
| --- |
| C:\Users\m1217\Google 雲端硬碟\Academic\NPUST\Research\DeepForensic\ICIP19\visresult.png |
| 圖4-1 偽造人臉影像的偽造特徵定位視覺化結果 (a) 與 (c) 分別為來自 PGGAN 與 DCGAN 之結果，而 (b) 與 (d) 分別是對應的瑕疵偵測結果 |

由文獻 [32] 可知，由於全卷積網路的特性可以從單純影像辨識任務，藉由設計最後一層的卷積層 Channel 數量相當於類別數量，來學習出影像上哪個區域的特徵響應 (Response) 較大導致於最終的分類器會將之辨識為該類別。藉由相同設計概念，我們在分類器網路的架構 將會設計成為全卷積的網路，藉此可以視覺化偽造影像的特徵，不僅可以提供使用者觀察偽造特徵產生的區域，亦可拿來當作辨識任務失敗時的核查機制 (例如偽造人臉辨識成為真實人臉)。本論文以 PGGAN 產生的偽造影像當作範例，眾所皆知，PGGAN 在影像生成任務的表現上是相當頂尖的，但是還是能從圖上定位出偽造的部分，如[圖 4-1](#圖4之1) 所示。很明顯可以發現，[圖4-1(a)](#圖4之1) 中，在頭髮髮線與脖子邊緣都呈現不自然的變化，且人臉比例也過於拉長，因此本論文之方法傾向於幾乎整張臉的有偽造的瑕疵可以發掘。而[圖4-1(c)](#表4之1) 中，在髮線與人臉與背景的邊緣中，也呈現著不規則變化，同樣的，都有機會被本論文之方法偵測出來。

1. **偵測偽造視訊實驗結果&分析**

表4-4、偵測偽造視訊實驗結果之比較 (Recall / Precision Rate)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Adjustment  Model | Original | | Data  Augmentation | | Model  Adjustment | | Mixed | |
| precision | recall | precision | recall | precision | recall | precision | recall |
| EfficientNetB5 | 0.6623 | 0.8572 | 0.5716 | 0.6908 | 0.7775 | 0.9027 | 0.5737 | 0.6922 |
| ResNeSt101 | 0.5826 | 0.7157 | 0.5525 | 0.5173 | 0.6829 | 0.604 | 0.5165 | 0.5421 |

在偵測偽造視訊的實驗中，我首先不對網路進行任何調整直接進行測試以得出基本的效能，接著我在對兩個網路進行一系列相同的調整來確認該調整是否能對網路進行最佳化，並且觀察哪類調整能對網路效能進行最大的提升以及嘗試分析背後原因。

我將所有在實驗中做的調整大致分為兩類:對網路做的調整和對圖像做的調整，對網路做的調整包含以下幾項:

* 優化器: 預設為 Adam 調整後為 SGD
* Learning rate: Adam 的learning rate為 0.001，SGD 的 learning rate 為 0.01
* Learning rate decay
* Batch Normalization
* Label Smoothing

對影像的調整則包含以下幾項:

* Shift&Scale&Rotate
* JpegCompression
* GaussNoise
* GaussianBlur
* IsotropicResize
* Color jittering

Data Augmentation 的效果圖如[圖4-2](#圖4之2)，[圖4-2(a)](#圖4之2) 為原圖，[圖4-2(b)](#圖4之2) 至[圖4-2(g)](#圖4之2) 分別為經過 Shift&Scale&Rotate、JpegCompression、GaussNoise、GaussianBlur、Color jittering、IsotropicResize 調整後的影像。Mixed 則為將兩類改動同時實做到網路上得到的結果。

從結果中我發現 EfficientNet 整體上的表現都優於 ResNeSt，我猜測因為EfficientNet 的網路架構是經過 NAS 建構和優化而成，因此架構和ResNeSt相比之下較完整，在訓練的 Epoch 數較少的限制下可以比ResNeSt 快學習出圖中的特徵。另外我發現兩個網路在使用經過 Data Augmentation 後的圖進行訓練後反而降低了在測試集上的表現，這應該是因為每張影像在經過 Data Augmentation 變化後一部份的特徵變得比較難學或是在調整過程中被破壞了，因此我判斷應該需要更多的 Epoch 數才能證明 Data Augmentation 實際對網路學習帶來的成效。相對於 Data Augmentation，Model Adjustment 對網路學習的幫助就明顯許多，兩個網路在 Precision Rate 上都增加了約 15% 的效能，在 Recall Rate上 EfficientNetB5 則增加了 5%。因為這些調整或多或少都加速了網路收斂的速度，除了 Label Smoothing 是為了解決網路的過擬合而增加的調整。從 Mixed 的結果中也能看到網路先做了 Model Adjustment 後在使用經過 Data Augmentation 的圖進行訓練比起直接使用 Data Augmentation 的圖進行訓練，兩個網路在兩種評估標準上的結果都有些微成長。因此我判斷如果要在偵測偽造視訊上得到最佳的成果的話，首先在 EfficientNetB5 上進行 Model Adjustment 類的調整後使用原始偽造影像進行訓練，直到網路出現過擬合的現象後將訓練停止並且使用前一個 Epoch訓練出的參數和經過 Data Augmentation 變化後的影像作為訓練集接著訓練以增加網路的穩健性 (Robustness)，降低網路出現過擬合的機率。

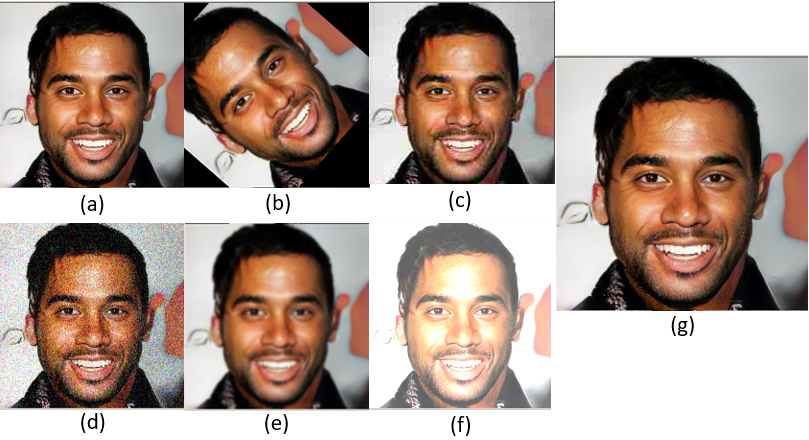


圖 4-2、Data Augmentation效果圖

**第五章 結論**

本論文中提出了一個以成對學習概念為基礎的共同偽特徵學習方法，可成功偵測最先進的 GAN 所生成的臉部影像，並引入了對比損失函數與三元損失函數於學習的策略中。共同偽特徵學習方法讓學到的特徵可以使得學習出之偽造人臉偵測器推廣到其他沒有見過的 GAN 之生成影像中。除此之外，本論文亦提出了進階的對偶深度神經網路，可以很好地捕捉到局部和全域特徵，讓偵測性能，進一步提升。最終，透過雙步學習的策略，避免了學習上的困擾，大幅提升本論文提出的偽造人臉偵測器學習效率與效能。實驗結果證明了我的方法在準確率和召回率上優於其他目前最頂尖的偵測方法，不過如果有人能蒐集到網路上所有的 GAN 來訓練偵測器，那該偵測器的效能就會比本論文提出的偽造人臉偵測器好，但是本論文提出的偵測器在通用性上一直都會是最好的。在偵測偽造視訊的部分雖然只是對現有的網路進行優化，但是藉由分析哪些調整能對網路效能帶來較大的進步可以對偽造視訊的本質有更好的了解，對於之後設計偵測偽造視訊的偵測器會有很大的幫助。

**參考文獻**

1. Karras, Tero, et al, “Progressive growing of gans for improved quality, stability, and variation,” *arXiv preprint arXiv:1710.10196*, 2017.
2. Brock, Andrew, Jeff Donahue, and Karen Simonyan, “Large scale gan training for high fidelity natural image synthesis,” *arXiv preprint arXiv:1809.11096*, 2018.
3. Zhu, Jun-Yan, et al, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” *Proceedings of the IEEE international conference on computer vision*, 2017.
4. “Ai can now create fake porn, making revenge porn even more complicated,” in http://theconversation.com/aican-now-create-fake-porn-making-revenge-porn-evenmore-complicated-92267, 2018.
5. Suwajanakorn, Supasorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman, “Synthesizing obama: learning lip sync from audio,” *Proceedings of the ACM Transactions on Graphics (TOG)*, Vol. 36, No. 4, pp. 1-13, 2017.
6. Farid, Hany, “Image forgery detection,” *IEEE Signal processing magazine*, Vol. 26, No. 2, pp. 16–25, 2009.
7. Bkell. “Digital vannmerking.” *Wikipedia* 2012. 18 December 2019 <https://commons.wikimedia.org/wiki/File:Visible\_digital\_watermarking.jpg>
8. 雪花台灣. “相機圖像質量之雜訊和視覺雜訊.” *雪花新聞* 14 March 2019. < https://tinyurl.com/tgsuwfx>
9. Chih-Chung Hsu, et al, “Video forgery detection using correlation of noise residue,” *Proceedings of the* *IEEE Workshop on Multimedia Signal Processing,* pp. 170–174, 2008.
10. Mo, Huaxiao, Bolin Chen, and Weiqi Luo, “Fake faces identification via convolutional neural network,” *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security*, 2018.
11. Marra, Francesco, et al, “Detection of gan-generated fake images over social networks, ” *Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2018.
12. Zhang, Xiangyu, et al, “Accelerating very deep convolutional networks for classification and detection,” *Proceedings of the IEEE transactions on pattern analysis and machine intelligence,* Vol. 38, No.10, pp. 1943-1955, 2015.
13. Kaiming He, et al, “Deep residual learning for image recognition,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
14. Gao Huang, et al, “Densely connected convolutional networks,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2017.
15. Andrew Brock, Jeff Donahue, and Karen Simonyan, “Large scale GAN training for high fidelity natural image synthesis,” *arXiv preprint,* arXiv:1809.11096, 2018.
16. Zhu, Jun-Yan, et al, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” *Proceedings of the IEEE international conference on computer vision*, 2017.
17. Ian Goodfellow, et al, “Generative adversarial nets,” *Proceedings of the* *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.
18. Alec Radford, Luke Metz, and Soumith Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint*, arXiv:1511.06434, 2015.
19. Martin Arjovsky, Soumith Chintala, and L´eon Bottou,“Wasserstein GAN,” *arXiv preprint*, arXiv:1701.07875, 2017.
20. Ishaan Gulrajani, et al, “Improved training of wasserstein GANs,” *Proceedings of the* *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.
21. Xudong Mao, et al, “Least squares generative adversarial networks,” *Proceedings of the* *IEEE International Conference on Computer Vision (ICCV)*, pp. 2813–2821, 2017.
22. Deepfakes github. https://github.com/ deepfakes/faceswap. Accessed: 2018-10-29.
23. Justus Thies, Michael Zollhofer, and Matthias Nießner, “Deferred neural rendering: Image synthesis using neural textures,” *Proceedings of the ACM Transactions on Graphics 2019 (TOG)*, 2019.
24. Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. “Face2Face: Real-Time Face Capture and Reenactment of RGB Videos,” *Proceedings of the* *IEEE Conference on Computer Vision and Pattern Recognition, pages 2387–2395,* 2016*.*
25. Faceswap. https://github.com/ MarekKowalski/FaceSwap/. Accessed: 2018-10-29.
26. Hoffer, Elad, and Nir Ailon, “Deep metric learning using triplet network,” *Proceedings of the International Workshop on Similarity-Based Pattern Recognition*, pp. 84–92, 2015.
27. Tan, Mingxing, and Quoc V. Le. “Efficientnet: Rethinking model scaling for convolutional neural networks,” *arXiv preprint arXiv:1905.11946*, 2019.
28. Sandler, Mark, et al. “Mobilenetv2: Inverted residuals and linear bottlenecks,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.
29. Zhang, Hang, et al. “ResNeSt: Split-Attention Networks,” *arXiv preprint arXiv:2004.08955,* 2020.
30. Liu, Ziwei, et al, “Deep learning face attributes in the wild,” *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.
31. Deepfake Detection Challenge *kaggle* 12 December 2019. < https://www.kaggle.com/c/deepfake-detection-challenge/data>
32. Oquab, Maxime, et al, “Is object localization for free?-weakly-supervised learning with convolutional neural networks,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 685–694, 2015.