基於深度學習資料隱藏之防止深度偽造之研究

聶聞華*
*國立台中科技大學資訊工程系
*nie991112@gmail.com

摘要

近年來，深度學習的迅速發展使得合成的影像愈發逼真，已經產生了可偽造人臉的深偽技術（Deepfake）又稱深度偽造，如果不能正確分辨偽造和原始人臉影像，則會對個人和社會產生嚴重威脅。現有研究著重於對偽造的影像進行鑒別，但在通用性和強韌性上仍然欠佳。因此，本文提出了一種對原始人臉影像進行隱寫加密的技術，將原始影像作爲秘密影像隱藏，使原有的人臉信息變爲兩層，對深度偽造進行預防。此方法使用深度偽造演算法，對模型生成的原始人臉影像掩護資料進行影像竄改。使用這些竄改之後的掩護影像作為資料，訓練資料還原模型將原始人臉影像還原，藉此提取出未被篡改之人臉影像。實驗結果表明根據本論文的方法能夠在掩護影像遭到竄改時，大致的還原原始人臉影像，以此證明本方法的可行性。希望未來對還原人臉影像的品質進行提升，且能將此方法沿用到深度偽造視頻還原之研究。

關鍵詞：資料隱藏、深度學習、Deepfake

1. 緒論

近些年來，深度學習在計算機視覺領域取得重大突破，生成對抗網絡合成影像越來越自然，肉眼也難以分辨。此類基於深度學習的多媒體合成與篡改技術被統稱爲深度偽造（Deepfake）。深度偽造采用“生成式對抗網絡”（GenerativeAdversarialNetworks，GAN）的機器學習模型，對原圖片進行修改，尤其是Deepfake技術生成的偽造視頻，通過篡改影像或交換視頻中的人臉區域從而改變載體本身所傳達的信息，是目前公认危害最大，研究也最广泛的一类篡改方式。

人臉深度篡改技術主要是基於生成對抗網絡，生成對抗網絡演算法包含兩部分神經網絡：生成器和判別器。生成器基於一個數據庫自動生成模擬該數據庫中數據的樣本；判別器則用來評估生成器生成數據的真偽。兩者在互相博弈學習中產生大規模且高精度的輸出。偽造影像產生的流程一般分為三步：先提取數據對人臉進行定位；然後通過訓練數據對人臉進行轉換；最後結合影像融合技術進行數據轉換，實現人臉影像拼接並合成視頻。

然而深度偽造技術對個人隱私數據、社會穩定和國家安全造成潛在的威脅。爲了有效的辨識深度偽造，需要開發一種有效的防禦機制。目前主流研究著重於對深度偽造后的影像進行辨識，然而不斷升級的GAN會抹除或破壞偽造的痕跡，使其難以分辨。

面對此問題，本文提出一種從影像上添加秘密影像的主動防禦方法來對抗深度偽造，該方法區別於針對偽造后影像的辨識，而是從源頭防範，通過解密隱藏在影像中的原始人臉信息，來判斷影像是否被惡意修改，還原原始人臉影像，嘗試追蹤深度偽造的源頭。

2. 文獻探討

在本節當中，將簡要介紹比較有代表性的Deepfake檢測、影像隱寫方法以及本文所使用之FC-DenseNets架構。

2.1 Deepfake 檢測

P Zhou 等學者[1]提出了一種雙流網絡結構，用來捕獲人臉偽造特徵和局部噪聲殘差線索，其中一個流是基於 GoogleNet[2]用作人臉分類網絡結構，通過訓練真假圖片數據得到一個二分類器，另一個基於補丁層次捕獲 CFA 模式和局部殘差的低層次相機特徵，使得模型既能發現高層次的篡改偽造特徵，又能捕獲到低層次的噪聲殘差特徵。通過融合雙流網絡的檢測分數，獲得一個比單流更好的效果。D Afchar 等學者[3]提出用 MesoNet 檢測深度偽造影像。由於在被壓縮的視頻內容中低層影像噪聲會被壓縮，而人眼很難分辨出高層語義水平的偽造影像。MesoNet 主要基於中層語義進行檢測，使用具有少量層的深層神經網絡。L Li 等學者[4]認為人臉影像被偽造都會存在一個換臉的邊界，所以提出 FaceX-Ray 用來確定影像是否由兩個來源不同的影像拼接而成，通過一種全新的方式檢測這個邊界，從而可以區別人臉是否被篡改。

2.2 深度資料隱藏技術

2017 年，Baluja[5]首次提出以圖藏圖的深度資料隱藏模型，實現了在彩色影像中隱藏彩色影像。所提出的網絡模型包括預處理網絡、編碼網絡和解碼網絡。預處理網絡將秘密影像變換為原始影像的大小，將基於色彩的像素轉化為利於編碼的特徵；秘密影像與原始影像經過編碼網絡得到掩護影像，在視覺上與原始影像沒有差異；解碼網絡將從掩護影像中提取秘

密影像。在訓練時，通過控制原始影像與掩護影像的距離，保證二者的相似性，減少秘密影像嵌入引起的影像變化；通過控制秘密影像與解密影像的距離，保證影像重構的準確性。

Baluja[5]提出的網絡模型中，掩護影像與重構的秘密影像視覺質量較差，影像的平滑區域存在許多噪聲，會引起數據統計異常且容易被人眼察覺。為瞭解決上述問題，Duan 等學者修改了編碼網絡的網絡結構，採用了 U-net[6]以及 FC-DenseNets[7]進行資料隱藏，兩種方法都有效改善了掩護影像的質量。但解碼器部分仍採用簡單的卷積堆疊方式，導致訓練速度較慢、效果較差。

2.3 DenseNet

DenseNet 不僅用於數據分類，還用於影像的超解析度和影像分割。Tong 等學者[8]提出了一種將 DenseNet 應用於影像超解析度的方法。影像通過 DenseNet 後，進行去卷積和重建，取得了良好的效果。鑒於傳統卷積網絡有一個L層，因此有L個連接。然而，DenseNet 網絡的每一層與後續的每一層之間都有一個連接，連接的總數為$L(L+1)/2$。

DenseNet 的優點是消除了梯度消失問題，增強了特徵傳播，大量特徵被復用，並能有效地減少模型參數。l^{th}層接受特徵圖X_0，X_1，…，X_{l-1}前面的所有層作為輸入：

$$X_l = Y_l([X_0, X_1, \cdots, X_{l-1}]) \quad (1)$$

$[X_0, X_1, \cdots, X_{l-1}]$表示層$0$，$1$，$2$，…，$l-1$中生成的特徵圖的串聯。將多個張量組合成一個張量，以便更好地應用。Y是一個包含三個操作的復合函數：即批量歸一化（BN）和校正線性單元（ReLU）、卷積（Conv）。

2.4 FC-DenseNet

FC-DenseNet 是一個密集的卷積神經網絡。它源於 DenseNet，FC-DenseNet 主要用於語義分割，根據標簽可以對圖片中的天空、樹木、車輛、行人、道路等進行分割，也適用於視頻中的分割。主要過程是在密集塊後面添加下採樣和上採樣功能。下採樣是一個2×2卷積塊，用於提取特徵映射的最大值，以減少影像並簡化計算。上採樣由3×3轉置卷積組成，卷積核的步長為 2，以補償池運算。最後，通過1×1卷積和類別數進行分類。更直接的是，輸出特徵圖的數量是不同類別的數量，分割的效果很好。

3. 研究方法

從文獻探討中能夠發現，目前資料隱藏技術的影像質量及藏密資料量有所提高，但秘密影像訓練考慮強韌性的部分較少，在掩護資料遭到攻擊破壞時進行解密會降低還原效率。一旦掩護資料遭到諸如高斯雜訊、影像壓縮等常見攻擊，還原效率便會降低，難以得到原始秘密影像。並且在面對資料隱藏分析工具時，有可能被辨識出來。使用 GAN 生成深度偽造的過程，本質上便是對影像攻擊的過程。所以本研究提出以深度偽造作為攻擊手段，以達到提升資料隱藏演演算法之強韌性和還原深度偽造影像的目的。下面會介紹本研究中所采用之實驗資料，隱藏網絡和還原網絡的網絡架構，如何設計損失函數以及訓練參數。

3.1 實驗資料

本文方法旨在應對常見的兩類基於 GAN 的深度偽造即全臉合成和局部編輯。為此，選擇了較好的 GAN 模型即 STGAN[9]，來合成深度偽造影像。對於全臉合成，使用 GAN 對影像進行重構。對於局部編輯，在 GAN 下分別選擇 3 種不同的屬性對影像進行對應的編輯獲得部分篡改的影像。例如修改人物的嘴巴開合程度、為其添加胡須、將性別統一設置為女性。3 種不同屬性在訓練集中比例為1:1:1。實驗中使用的所有影像來自人臉數據集 CelebA[10]，它包含 202599 張影像且每張影像對應個二進制屬性標註。實驗中將影像統一處理為128×128像素，共采用 5000 張影像，以3:1:1的比例劃分為訓練集、驗證集和測試集。驗證集和測試集則添加未修改之掩護影像，與 3 種不同屬性的比例為1:1:1:1。

3.2 網絡架構

本研究所提出之架構主要包括串聯、隱藏網絡(encoder)、GAN 模擬器以及還原網絡(decoder)，詳細內容如圖 1 所示。

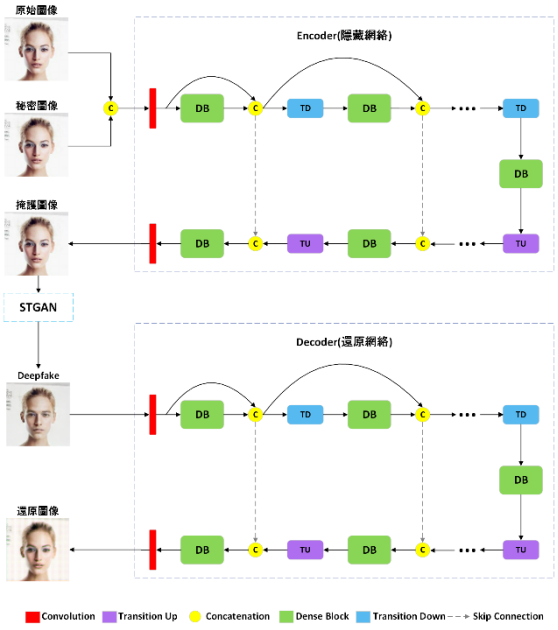

圖 1. 網絡架構

其中原始影像和秘密影像經過串聯後輸入。通過 TensorFlow 中的影像大小函數統一調整為 128×128。隱藏網絡(encoder)主要用於隱藏秘密信息，通過 GAN 網絡模擬深度偽造環節，達到提升強韌性之效果，最後還原網絡(decoder)用於提取秘密信息。本研究之目標是同時訓練隱藏網絡和還原網絡，以實現人臉信息隱藏的保密性。

3.3 隱藏網絡(Encoder)

隱藏網絡初始進行一次串聯操作。該函式將原始影像和秘密影像拼接，並輸出一個 6 通道張量，為接下來資料隱藏做準備。隱藏網絡的結構如表 1 所示。

表 1. 隱藏網絡架構

$Step$	$Feature-map(W \times H \times D)$
3×3Conv	$128 \times 128 \times 48$
4(BN+RELU+Conv) + TD	$64 \times 64 \times 96$
4(BN+RELU+Conv) + TD	$32 \times 32 \times 144$
4(BN+RELU+Conv) + TD	$16 \times 16 \times 192$
4(BN+RELU+Conv) + TD	$8 \times 8 \times 240$
4(BN+RELU+Conv) + TD	$4 \times 4 \times 288$
4(BN+RELU+Conv)	$4 \times 4 \times 336$
TU + 4(BN+RELU+Conv)	$8 \times 8 \times 372$
TU + 4(BN+RELU+Conv)	$16 \times 16 \times 336$
TU + 4(BN+RELU+Conv)	$32 \times 32 \times 288$
TU + 4(BN+RELU+Conv)	$64 \times 64 \times 240$
TU + 4(BN+RELU+Conv)	$128 \times 128 \times 192$
3×3Conv	$128 \times 128 \times 3$

在表 1 中，第一列表示隱藏網絡的架構，第二列表示輸出特徵圖的大小，D 表示通道數。隱藏網絡的輸入是 6 通道張量，輸出是 3 通道張量。隱藏網絡的首要目標是對通道數為 6 的張量進行編碼，輸出的影像應盡可能接近原始影像。隱藏網絡組件包括兩個3×3卷積、11 個 Dense Block（DB）、10 個 Concatenation、5 個 Transition Down（TD）和 5 個 Transition Up（TU）。第一個卷積的輸入是 6 通道張量，輸出是 48 通道張量。最後一個卷積的輸入是 192 通道張量，輸出是 3 通道張量。TD 由 BN、ReLU、1×1 Conv、2×2最大池組成。TU 是 3×3轉置卷積，$strides = 2$，$padding = 1$。TD 主要用於提取和融合秘密影像和原始影像的特徵。TU 的作用是逐漸恢復輸入的特徵。誤差 α 和 TU 的組合可以高精度地恢復原始影像。每個 DB 由 4 層組成，每層包括 BN、ReLU、3×3Conv、$Dropout = 0$，詳細內容如圖 2 所示。我們的目標是最大限度地減少掩護影像和原始影像之間的損失：

$$L_c = \|c - c'\| \qquad (2)$$

其中 c 和 c' 代表原始影像和掩護影像。

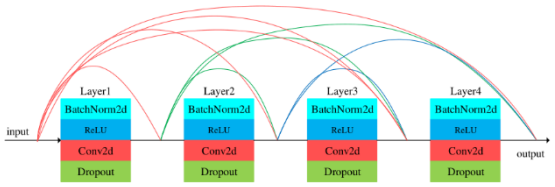

圖 2. Dense Block 架構圖

3.4 還原網絡(Decoder)

在圖 1 中，還原網絡的輸入是一個三通道張量，輸出也是一個三通道張量。為提高秘密影像信息的傳遞，嘗試還原網絡架構與隱藏網絡對稱相同的方式。還原網絡的主要目標是還原掩護影像中的秘密影像，還原後的影像應盡可能與原始秘密影像相似。我們在還原網絡中的目標是最大限度地減少已還原秘密影像和原始秘密影像的丟失，同時保證影像遭受 STGAN 攻擊后的損失能夠傳遞：

$$L_s = \|s - s'\| + \alpha \times \|c' - d\| \qquad (3)$$

其中 s、s'、d 分別代表秘密影像、還原影像、Deepfake 影像，α 是權重因子，固定為 0.5。

3.5 損失函數

與傳統的影像重建不同，資料隱藏過程需要兩幅輸入影像和兩幅輸出影像。因此，常規損失函數可能不適用於此目的。為了提高體系結構的性能，引入了定製的損耗函數。有三種損失需要計算：隱藏損失、還原損失、攻擊損失。隱藏損失由隱藏網絡在輸入原始影像和輸出掩護影像之間計算得出。還原損失通過還原網絡輸入秘密影像和還原的秘密影像之間計算得出。攻擊損失通過掩護影像和 Deepfake 影像計算得出。總損失是隱藏損失、還原損失、攻擊損失的總和：

$$L = L_c + \beta L_s \qquad (4)$$

其中 β 是權重因子，固定為 0.3。通過將 β 值從 0.3、0.6 和 0.9 變化來進行初始實驗。增加 β 值會增加損失，0.3 值產生最佳損失值。將隱藏網絡的損失函數反饋給隱藏網絡，並將整體損失反饋給還原網絡，以最小化還原的秘密影像的失真。

3.6 訓練參數

本研究使用 Adam 優化器聯合訓練嵌入網絡和恢復網絡，其中，學習率自定義為前 40 epoch 保持不變為 0.0003，從 40 到 100 epoch 降為 0.0001，最後 100 到 200 epoch 降為 0.00003。影像進入模型的 batch size 設置為 32。損失函數中的超參設置為 $\alpha = 0.5$、$\beta = 0.3$。GAN 模擬器接收掩護影像並在此基礎上合成虛假影像。在實驗中，選擇 STGAN 來構造模擬器。

4. 實驗結果

實驗硬體配置：1 個 GPU，GPU 型號為 NVIDIA TESLA P100。實驗所需的軟體環境：深度學習框架 TensorFlow v2.8.0，python 3.7。對於效果評估，我們採用了兩種評估方法：視覺評估和定量評估。影像失真由 SSIM 和 PSNR 評估。

4.1 視覺評估

為了查看模型的性能，本實驗從掩護影像中減去原始影像以獲得錯誤影像。在圖 3 的 6 和 7 列中，能發現掩護影像的偽影不太明顯，而秘密影像偽影較爲明顯。為了進一步驗證模型的泛化能力，從 CelebA[10]數據集中選擇了幾幅影像進行測試。試驗結果如圖 4 所示。可以發現還原結果在人的視覺下相對理想。

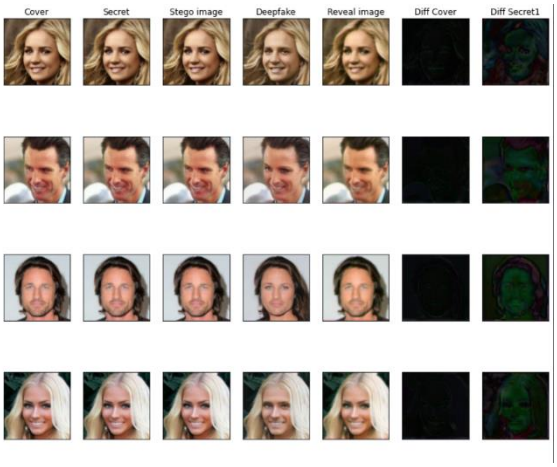

圖 3. 視覺評估

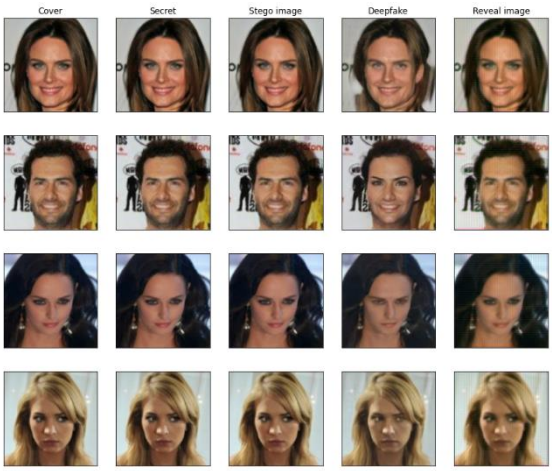

圖 4. 泛化能力測試

4.2 定量評估

本實驗還測量了隱藏影像的統計數學分佈質量。在影像處理中一個廣泛使用的衡量影像質量的指標是峰值訊號與雜訊比(簡稱為 PSNR)。PSNR 主要用於測量影像的失真率並將其顯示為分數。其定義如下：

$$\mathrm{PSNR}=10\times\log_{10}\frac{255^2}{\mathrm{MSE}} \qquad (5)$$

MSE 是統計裡面的均方誤差，定義如下：

$$\mathrm{MSE}=\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2 \qquad (6)$$

一般來說，計算出的 PSNR 越高，代表品質越好，還原的影像越接近原始影像。大於 30 的 PSNR，很難用肉眼察覺影像的差異。

為了更全面地評估掩護影像，我們計算了原始影像和掩護影像之間的 SSIM，定義如下：

$$\mathrm{SSIM}=\frac{(2\mu_X\mu_Y+k_1R)(2\sigma_{XY}+k_2R)}{(\mu_X^2+\mu_Y^2+k_1R)(\sigma_X^2+\sigma_Y^2+k_2R)} \qquad (7)$$

實驗中使用默認值$k_1=0.01$，$k_2=0.03$，返回值介於$[-1,1]$之間，其中 1.0 表示兩個影像相同。為了進一步說明深度資料隱藏技術的有效性，在表 2 中，對比了 Duan 等學者[7]以及本實驗原始影像和掩護影像、還原影像以及原始秘密影像的 PSNR 和 SSIM。在表 2 的第二行中，我們可以看到，原始影像和掩護影像的 PSNR 和 SSIM 的平均值達到（35.075，0.972），還原影像和秘密影像的 PSNR 和 SSIM 的平均值達到（26.583，0.756）。可以發現，在經過 GAN 模型的深度偽造攻擊後，還原影像和秘密影像的 PSNR 和 SSIM 明顯下降，說明在防止深度偽造方面仍有很大提升空間。但是，原始影像和掩護影像的 PSNR 和 SSIM 相對較好，所以可以說明此資料隱藏方法有一定可用性。

表 2. 隱藏網絡架構

schemes	原始影像和掩護影像的 PSNR 和 SSIM	還原影像和秘密影像的 PSNR 和 SSIM
Duan 等學者[7]	（39.556，0.985）	（37.092，0.975）
本實驗	（35.075，0.972）	（26.583，0.756）

5. 結論

本文不同於傳統深度偽造檢測手段，從一種新的角度，提出了基於深度資料隱藏法的深度偽造防禦研究。通過對掩護影像進行深度偽造攻擊，來提升深度學習模型還原原始人臉信息的成功率，同時提升模型之強韌性。實驗結果表明，目前雖能還原出大致人臉信息，但影像失真程度較高，影像之良率有待提升，模型

之泛化能力也有待提升。本研究之下一步將嘗試改善上述問題，同時嘗試采用不同的深度學習模型，提高人臉信息還原率和品質。最後希望此研究未來能夠沿用到 Deepfake 視頻僞造預防之研究。

參考文獻

[1] Zhou, P., Han, X., Morariu, V. I., & Davis, L. S, "Two-stream neural networks for tampered face detection," *In 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 1831-1839, July 2017.

[2] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A, "Going deeper with convolutions," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1-9, 2015.

[3] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I, "Mesonet: a compact facial video forgery detection network," *In 2018 IEEE international workshop on information forensics and security (WIFS)*, pp. 1-7, December 2018.

[4] Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B, "Face x-ray for more general face forgery detection," *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5001-5010, 2020.

[5] Baluja, S, "Hiding images in plain sight: Deep steganography," *Advances in neural information processing systems*, pp. 30, 2017.

[6] Duan, X., Jia, K., Li, B., Guo, D., Zhang, E., & Qin, C, "Reversible image steganography scheme based on a U-Net structure," *IEEE Access, 7*, pp. 9314-9323, 2019

[7] Duan, X., Nao, L., Mengxiao, G., Yue, D., Xie, Z., Ma, Y., & Qin, C, "High-capacity image steganography based on improved FC-DenseNet," *IEEE Access*, 8, pp. 170174-170182, 2020.

[8] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 4799-4807, October 2017.

[9] Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., & Wen, S, "Stgan: A unified selective transfer network for arbitrary image attribute editing," *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3673-3682, 2019.

[10] Liu, Ziwei and Luo, Ping and Wang, Xiaogang and Tang, Xiaoou, "Deep Learning Face Attributes in the Wild," *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.