**AI視頻換臉之deepfake技術綜述（深度偽造製作與檢測）**

https://zhuanlan.zhihu.com/p/389197155

1、背景

AI換臉 （face-swap）是指用另一個人臉來替換一張圖片或視頻中的一個人臉，合成新的媒體物，它是Deepfake（“深度偽造”）技術最廣為人知的一種應用形式。Deepfake是指基於深度學習等機器學習方法創建或合成視聽覺內容，如圖像、音視頻、文本等。

Deepfake的製作和檢測研究自2017年以來，相關論文數從3篇增加到150篇以上（2018-19年）。本報告以Deepfake技術為主線索，重點闡述AI換臉技術的發展、原理及其應用。



圖1-1 deepfake換臉示例（左邊是原人臉圖像，右邊是替換人臉後圖像）

**Deepfake技術發展過程（截止2020.01，涵蓋AI換臉）：**

2014.06：提出生成對抗網路（GAN），在圖片創建方面取得重大突破，之前的AI演算法可以較好的分類圖片但創建圖片困難；

2017.07：提出一種使用RNN 的LSTM（Long Short-Term Memory）學習口腔形狀和聲音之間關聯性的方法，僅通過音訊即可合成對應的口部特徵；

2017.10：提出一種基於 GAN 的自動化即時換臉技術；

2017.12：出現deepfake色情視頻，名人的面孔被換成色情演員的面孔，由Reddit使用者使用自動編碼器-解碼器配對結構開發（FakeApp前身）；

2018.04：出現美國前總統奧巴馬的deepfake演講，將聲優的口型（含模仿的配音）替換到奧巴馬演講視頻中，使用了FakeApp；

2018.08：提出一種將源視頻中的運動轉移到另一個視頻中目標人的方法，而不僅是換臉（效果imperfect）；

2019.03：提出一種控制圖片生成器並能編輯造假圖片各方面特性的方法，比如膚色、頭髮顏色和背景內容，不同於之前的假人圖片生成方法，這是一種重大突破；

2019.05：提出一種真實頭部說話神經模型的少樣本對抗學習，基於GAN元學習，該模型基於少量圖像（few-shot）訓練後，向其輸入一張人物頭像，可以生成人物頭像開口說話的動圖；

2019.06：出現“一鍵式”智慧脫衣軟體 Deepnude，迫於輿論壓力，開發者快速下架。

2020：主要進展在提高原生解析度、提升deepfake製作效果方面。

**2、視覺深度偽造技術分類**

從視覺角度，deepfake一般可劃分四類：重現（reenactment）、替換（replacement）、編輯（editing）、合成（synthesis）。

儘管人臉編輯和合成研究火熱，但重現和替換才是最大的隱患，它們可以讓攻擊者控制身份和欺騙。

**2.1、重現（reenactment）**

重現使用源身份Xs驅動目標身份Xt，使得Xt的**行為**和Xs一樣，包括表情、嘴部、眼部、頭部及軀幹。

l 表情重現

讓Xt模仿Xs的表情，可用於電影行業中後期調整演員的表情表演、教育行業中對歷史人物表情再現講解事物。

l 嘴部重現

也稱“配音”，Xt的嘴部由Xs驅動，或者是包含語音的音訊輸入，可用於將逼真的語音配音成另一種語言並進行編輯。

l 眼部重現

Xs視線驅動Xt眼睛的方向和眼瞼的位置，可用於改善照片品質或在視頻採訪中自動保持眼神交流。

l 頭部重現

Xt頭部位置由Xs驅動，可用于在安全錄影中對人臉進行朝向轉正，並用作改善人臉識別軟體的手段。

l 身體重現

也稱姿態遷移和人體姿勢合成，與面部重現類似，是身體軀幹的重現。

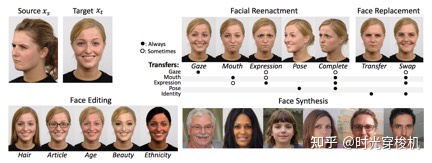


圖2-1 人臉重現、替換、編輯、合成分別生成的deepfake示例

攻擊模型：攻擊者能用其假冒身份，假冒他人說或做，比如誹謗、散佈錯誤資訊、篡改證據、騙取信任詐騙、生成虛假材料勒索等。

**2.2、替換（replacement）**

替換使用源身份Xs的內容替換目標身份 Xt，使得目標**身份**變成了源身份Xs。

l 轉移（transfer）

使用Xs的內容替換 Xt，常見是面部轉移（即用Xs的臉直接覆蓋 Xt的面部），在時尚行業中用於不同服裝中的個人虛擬試穿。

l 交換（swap）

由Xt驅動Xs的內容轉移到 Xt（從效果上看，Xt原有的面部表情等會被保留），最流行的是“換臉”，比如將某人的身份同名人的身份交換來產生諷刺或不良的內容效果、一種健康用途是在公共場合或平臺匿名化身份以代替模糊或圖元化。



圖2-2 人臉替換生成的deepfake示例

攻擊模型：替換技術因其有害應用而聞名，比如攻擊者將受害者的臉換到色情女演員的身體上，以侮辱、誹謗和勒索受害者，還可將一個人的臉轉移到一個看起來相似的身體上充分再現此人。

**2.3、編輯（editing）**

編輯是指添加、更改或刪除目標身份的**屬性**，比如，更換目標物件的髮型、衣服、鬍鬚、年齡、體重、顏值、眼鏡和種族等屬性。

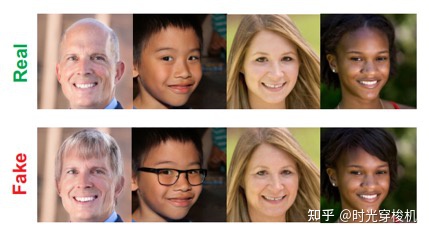


圖2-3 編輯屬性生成的deepfake示例

攻擊模型：攻擊者可以使用相同的過程來建立虛假的角色誤導他人，比如，可以使一個患病的領導者看起來健康的，震懾“敵方”，一個不道德的用途是去除受害者的著裝使其裸露，即所謂的一鍵脫衣。

**2.4、合成（synthesis）**

合成是指在**沒有目標身份作為基礎**的情況下創建deepfake角色，類似直接用GAN或者其它生成模型生成人臉，沒有明確的target。人臉和身體合成技術可以創建影視素材，生成電影和遊戲角色。



圖2-4 全臉合成生成的deepfake示例

攻擊模型：攻擊者可以線上創建虛假角色進行詐騙等違法活動。

**3、視覺深度偽造生成技術**

**3.1、深度偽造生成網路**

通常使用5種神經網路的變種或組合構建深度偽造生成網路：編碼－解碼網路（Encoder-Decoder, ED）、卷積神經網路（Convolutional neural network, CNN）、生成對抗網路（Generative adversarial networks, GAN）、圖像轉換網路（Pix2Pix, CycleGAN）、遞迴神經網路（Recurrent neural network, RNN）。

**3.1.1、編碼－解碼網路（ED）**

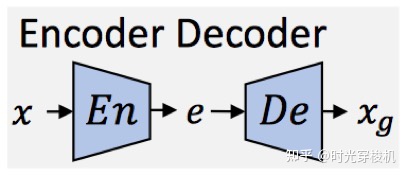


圖3-1 編碼－解碼網路架構

網路至少包含一個編碼器En和一個解碼器De，連接編碼和解碼的中間層較窄，因此當訓練De(En(x)) = xg時，網路將被迫學習、匯總訓練樣本的高層語義概念。

給定x的分佈X，En(x) = e，通常稱e為編碼或嵌入(embedding)，而E = En(X) 被稱為“潛在空間(latent space)”。

Deepfake技術通常使用多個編碼器或解碼器，並操縱編碼來影響輸出 xg。如果編碼器和解碼器是對稱的，並且以目標 De(En(x)) = x訓練網路，則該網路稱為自動編碼器（Autoencoders），輸出是x的重建。

ED的另一種特殊類型是變分自動編碼器（VAE），其中編碼器學習給定X的解碼器後驗分佈。VAE比自動編碼器在生成內容方面更好，因為潛在空間中的表徵可以被更好地解耦。

**3.1.1.1、自動編碼器（Autoencoders）**

自動編碼器提取面部圖像的潛在特徵，使用解碼器重建面部圖像，為了在源圖像和靶心圖表像之間交換面部，需要兩個編碼器/解碼器對。

兩對具有相同構造的編碼器網路，各自在圖像集上訓練，編碼器的參數在兩個網路對之間共用。（有點encoder“求同”、decoder“存異”的味道）

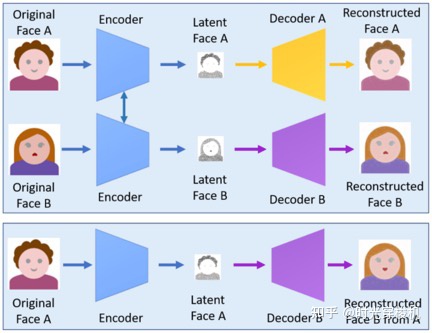


圖3-2 使用兩個編碼器／解碼器對的Deepfake創建模型

該策略使通用編碼器能夠找到並學習兩組面部圖像之間的相似性，面部通常具有相似的特徵（例如，眼睛，鼻子，嘴巴的位置），這是較為不具挑戰性的。

兩個網路使用相同的編碼器編碼，但使用不同的解碼器訓練解碼過程（圖3-2上部），人臉A的圖像使用通用編碼器編碼，並使用解碼器B解碼，以創建一個deepfake（圖3-2下部），人臉的角度、表情以及頭型的相似度影響人臉替換的效果。

**3.1.2、卷積神經網路（CNN）**

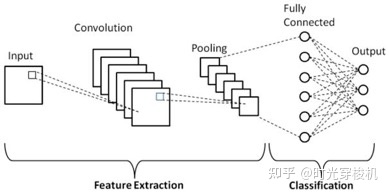


圖3-3 基本的CNN網路架構

與全連接網路相反，CNN擅長學習資料中（局部）結構模式並組合得到高層次表徵，因此在處理圖像方面效率更高。

CNN中的卷積層學習的是卷積核／濾波器參數，這些濾波器在輸入圖像上移動，提取抽象的特徵圖作為輸出。

隨著網路變得越來越深，池化層降低維數，上採樣層提高維數；它們可以靈活地構建用於圖像的ED CNN。

**3.1.3、生成對抗網路（GAN）**

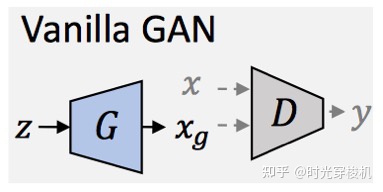


圖3-4 Vanilla GAN網路架構

GAN由Goodfellow等人於2014年首次提出，源於博弈論“零和博弈”思想，通過生成模型G (generative model) 和判別模型D (discriminative model) 互相博弈的方法來學習資料分佈的生成式網路。

G旨在欺騙D來創建偽樣本，它通過給定某種隱含資訊，隨機生成觀測資料樣本；D學會區分真實樣本（x∈X）和偽樣本（ xg = G(z) 其中z ∼ N），它預測資料樣本是否屬於真實訓練樣本。

具體來說，有一個分別用於訓練D和G的對抗損失：

Ladv (D) = maxlogD(x) + log(1−D(G(z)))

Ladv (G) = minlog(1−D(G(z)))

在對抗博弈下，兩者通過對抗式訓練提升其能力，G學習如何生成與原始分佈無法區分的樣本。訓練後，將D丟棄，並使用G生成內容。

最理想的狀態是生成模型能夠生成足以“以假亂真”的資料樣本, 而判別模型卻對其真偽性難以判別，即判斷正確的概率只有 50%。

GAN優勢是不依賴先驗知識，生成模型的參數更新來自判別模型的反向傳播，而非直接來自於資料樣本，故訓練不需要複雜的瑪律科夫鏈。當應用于圖像時，通常可以生成更高品質、逼真的圖像樣本。

**3.1.4、圖像轉換網路（Pix2Pix、CycleGAN）**

Pix2Pix和CycleGAN是兩種流行的使用GAN基本原理的圖像轉換網路。

（1）Pix2Pix可以完成從一個圖像域到另一個圖像域的轉換。

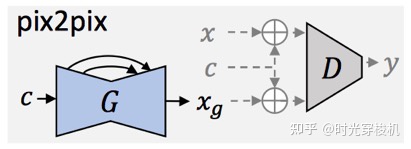


圖3-5 Pix2Pix網路架構

在Pix2Pix中，G以輸入圖像x c作為輸入，在給定目標標籤x，希望G學習xc -> x的映射，即希望G所生成的圖像xg和x無限接近，而D區分 (x, xc) 和 (xg, xc)。

Pix2Pix是一種監督式、成對式的訓練方式，對資料有嚴苛要求，提升版本的Pix2PixHD可用來生成具有更好保真度的高解析度圖像。

（2）CycleGAN可以通過不成對的訓練樣本來進行圖像轉換。

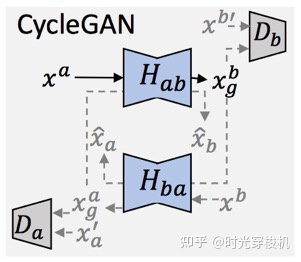


圖3-6 CycleGAN網路架構

該網路由兩組GAN構成，並形成一個迴圈約束：將圖像從一個域轉換為另一個域，然後再次返回時，確保一致性。

**3.1.5、遞迴神經網路（RNN）**

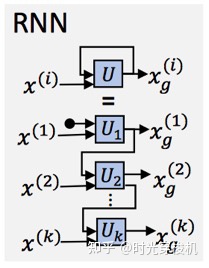


圖3-7 RNN網路架構

RNN可以處理序列和可變長度資料的神經網路，RNN的更高級版本包括長期短期記憶（LSTM）和門遞迴單元（GRU）。

在Deepfake製作中，RNN通常用於處理音訊、視頻。

**3.2、特徵表示**

大多數神經網路結構使用一些中間表示來捕獲、控制源身份s和目標身份t的面部結構、姿勢和表情等，常見的各種方法如下。

l 使用面部動作編碼系統（ facial action coding system，FACS）並測量面部的每個分類動作單元（AU）。

l 使用單目重建，從2D圖像獲得頭部的3D可變形模型（3DMM），其中，姿勢和表情使用向量和矩陣參數化、頭部使用參數或3D渲染。

l 使用頭部或身體的UV圖來使網路更好地瞭解形狀的方向。

l 使用圖像分割來説明網路分離不同的語義區域（面部、頭髮等）。

l 最常見的表示形式是特徵點（也稱為關鍵點），即面部或身體上一組定義的位置點（利用開源CV庫可有效跟蹤）。

l 有的按通道分隔特徵點，以使網路更容易標識和關聯它們。

l 類似地，面部邊界和身體骨骼點的表示等。

l 對於音訊，常見方法是分割音訊段，衡量音訊段的Mel-Cepstral Coefficients（MCC）捕獲主要語音頻率。

**3.3、泛化性**

一個deepfake網路一般被設計或訓練只針對特定的目標身份和源身份生效。身份不可知模型有時很難實現， 這是因為模型在訓練期間學習了s和t的相關性。

定義E：表徵x的特徵或從x提取特徵的模型或過程；M：一個訓練後的模型，用於實現重現或替換；據此定義模型的3種泛化類別。

l one-to-one

用特定身份驅動特定身份：xg = Mt (Es (xs))。

l many-to-one

用任意身份驅動特定身份：xg = Mt (E (xs))。

l many-to-many

用任意身份驅動任意身份：xg = M (E1(xs), E2 (xt))。

**3.4、技術挑戰**

l 泛化性

生成網路是資料驅動的，因此在所生成的結果中反映了訓練資料的特性；良好的效果受限於訓練資料集，特定身份的高品質圖像需要該身份的大量樣本；獲取驅動物件的訓練資料通常比受害者的容易的多；如何最小化訓練資料，並使訓練過的模型在新的目標身份和源身份（非訓練資料）上也生效。

l 成對式監督訓練

帶標籤成對具有GT（ground truth，正確打標記的資料）的資料實在難以獲取。對此，許多deepfake網路通過使用自監督的方式進行訓練，或者使用不成對的網路（例如CycleGAN），或者利用ED網路的編碼在潛在空間進行特徵編輯。

l 非期望特徵遷移

有時會把源身份的特徵或區域遷移到目標身份上，對此提出使用注意力機制、few-shot學習，解耦學習、跳躍連接等方法將更相關的資訊傳遞給生成器。

l 遮擋

遮擋可能是手、頭髮、眼鏡或任何其它物品，還有眼睛或嘴巴被隱藏或動態變化，這可能導致（裁剪的圖像、不一致的面部特徵）出現偽影、不合理的生成效果等，對此提出對障礙物進行分割和修補。

l 時間連貫性

Deepfake視頻通常會產生更明顯的偽影，例如閃爍和抖動，是因為大多數deepfake網路無先前幀的上下文、單獨處理每個幀，對此提出將上下文提供給G和D，或進行時間連貫損失約束，或使用RNN，或對它們進行組合使用。

**3.5、深度偽造製作**

**3.5.1、製作方法折衷**

每個製作方法有不同的成本和收益，最有效和最具威脅性的方法是：（1）最具實施性（訓練資料、執行速度、可訪問性）；（2）對受害者最可信（品質／可信度）。

l 資料&品質

用大量資料訓練的模型產出的效果通常更好，但需許多小時的訓練素材，僅適用于知名人士。

對任意個人的攻擊需使用具有many-to-many泛化能力的模型或少樣本學習方法，這類模型必須“想像”缺少的資訊（例如姿勢、遮擋），通過提供數量有限的參考樣本可達成資料和品質之間的折衷。

l 速度&品質

這兩者之間的取捨取決於攻擊是線上（互動式）或離線（存儲媒體），此類社會工程學攻擊通常是線上的，需要即時速度。

但高解析度的模型具有許多參數、可能會使用多個網路、可能會處理多個幀以提供時間連貫性；其它方法可能會因預處理／後處理步驟變慢。

已知的聲稱能即時產生偽造品的模型在主觀上趨於模糊或扭曲面部。當在遭受社會工程學攻擊的虛假誤導壓力之下，不完美的偽造品很可能對受害者是有效的。

攻擊者很可能在低解析度水準上實施複雜的方法以加快畫面播放速率。對於非即時攻擊，解析度和保真度重要，具有時間連貫性的高品質圖像和視頻是最佳選擇。

l 可訪問性&品質

可訪問性和可複現性是新技術擴散的關鍵因素，在網路公開發佈的代碼和資料更有可能被研究人員和罪犯使用，這是因為相比使用網路上有效可用的方法，複現論文的收益很小。

考慮以上因素，以下是目前最重要和最可用的深度偽造技術：

（1）面部重現：高效和實用性；

（2）嘴巴重現：品質高；

（3）人臉替換：高保真度和廣泛傳播使用。

**3.5.2、當前的局限性**

除了品質，當前的深度偽造製作還存在一些局限性技術。

l 重現總是以正面姿勢驅動和生成

這導致重現的表現非常靜態，可通過換臉把身份交換到相似的身體上避免，但不一定總能做到很好的匹配，實用性有限。

l 重現和替換依賴驅動器的表現來交付身份的特性

下一代的深度偽造將利用目標的視頻生成具有期望的表情和舉止的內容，使得創建令人信服的深度偽造過程更加自動化。

l 連貫渲染

頭髮、牙齒、舌頭、陰影的連貫渲染，以及渲染目標目標手的能力（尤其是在觸摸臉部時）。

l 原生解析度



圖3-8 深度偽造支持的原生解析度進展

用於訓練的資料原生解析度每增長P，則訓練時長至少增加P^2。

**3.5.3、基本製作流程**

總體分為三個核心步驟：資料收集、模型訓練、偽造內容生成。

l 資料收集

大量源和目標人物的人臉照片，盡可能覆蓋人臉的各種角度。

l 模型訓練

訓練耗時依賴硬體設定、訓練資料的品質。

l 偽造內容生成

對於生成目標xg，重現和人臉替換類型的網路一般遵循以下流程（將x傳遞到以下pipeline）：

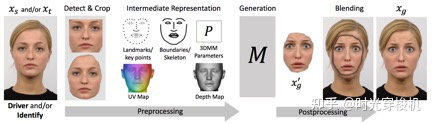


圖3-9 重現和人臉替換處理pipeline（通常僅執行部分步驟）

（1）檢測並裁剪面部；

（2）提取中間特徵表示；

（3）根據一些驅動資訊（如另一張臉）生成新的面部內容；

（4）將生成的臉融合到目標幀中。

l 通常使用的6種驅動圖像生成的方法

（1）讓網路直接在圖像上執行映射學習；

（2）使用ED解耦身份和表情並在傳遞給解碼器前修改／交換編碼特徵；

（3）在將其傳遞給解碼器之前添加其它編碼（如AU、embedding）；

（4）在生成之前將中間人臉／身體的特徵表示轉換為所需的身份／表情；

（5）通過源視頻幀序列的光流場驅動生成；

（6）使用3D渲染、扭曲的圖像或生成的內容進行組合，以創建原始內容（頭髮、場景等）的合成，然後將該合成（粗略結果）傳遞給另一個網路（ 例如pix2pix）以改善真實感。

**3.5.4、換臉工具－FakeApp**

Reddit社區原始的deepfake使用ED網路：1個編碼器En，2個解碼器Des和Det，作為2個autoencoders同時訓練：Des (En(xs)) = Xs，Det (En(xt)) = Xt，其中x是裁剪出的人臉圖像。En學會將s和t映射到共用的潛在空間：Des(En(xt)) = xg。

基於Reddit原始網路的更強大開源換臉工具有FaceSwap（34K星）、DeepFaceLab（23K星，95+% deepfakes的製作工具）、FaceSwap-GAN（3K星）等，支援了更多的模型，這些模型泛化能力均是one-to-one。

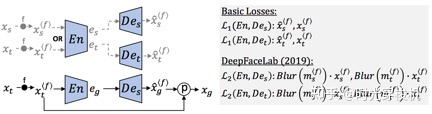


圖3-10 Reddit社區“deepfakes”模型及變種

**3.5.4.1、基本流程**

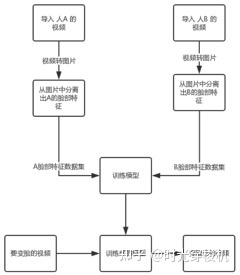


圖3-11 整個換臉過程

（1）資料收集

收集大量目標人臉的原始照片，提取目標人臉；

（2）模型訓練

對於64px input，64px output，訓練原始模型GPU耗時約12-48小時，比如（CUDA + GPU + TensorFlow），CPU耗時約數周；

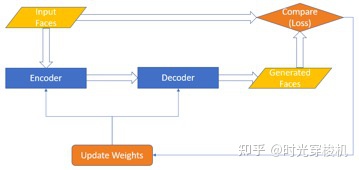


圖3-12 基本的編碼－解碼網路訓練過程



圖3-13 自動編碼解碼網路訓練過程

用2個資料集訓練共用的Encoder和 2個Decoder，Decoder A試圖重建Face A、Decoder B試圖重建Face B；

（3）偽造內容生成：向訓練過的模型輸入人臉圖像，用模型輸出的人臉，混合生成偽造內容。

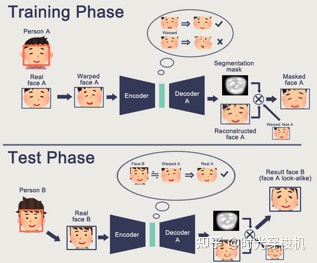


圖3-14 人臉轉換過程

**4、深度偽造技術應用**

**4.1、消極應用**

可被用於誤導輿論、擾亂社會秩序，甚至可能會威脅人臉識別系統、干預政府選舉和顛覆國家政權等，已成為當前最先進的新型網路攻擊形式。

（1）色情製作：2017年網路上顯著出現，Deeptrace公司調研估計換臉視頻96%是色情（2019.10）；

（2）虛假新聞：發佈或歪曲知名政客的言論，愚弄公眾等；

（3）語音詐騙：利用合成的語音進行金融詐騙；

（4）影像篡改：將個人面孔交換到電影明星身體插入影視剪輯中、移除CT或MRI醫療影像中證據進行保險欺詐。

**4.2、積極應用**

可用于推動娛樂與文化交流產業的新興發展。

（1）電影製作：電影製作中創建虛擬角色、視頻渲染、聲音類比；

（2）人物復活：“復活”歷史人物或已逝的親朋好友，實現“面對面”溝通，創造一種新型的交流方式。

**4.3、政策法規**

l 中國

2021年1月1日施行的《民法典》，第1019條首次創新性地規定，禁止任何組織或個人利用資訊技術手段偽造等方式侵害他人的肖像權，回應了技術發展背景下應對肖像權進行更為嚴密保護的需求。此處的“資訊技術手段”可能包括通過軟體修圖、偽造圖片、通過AI技術換臉、摳像或對他人肖像進行高精度偽造等方式。

l 美國

《2018 年惡意偽造禁令法案》中將“deep fake”定義為“以某種方式使合理的觀察者錯誤地將其視為個人真實言語或行為的真實記錄的方式創建或更改的視聽記錄”，其中“視聽記 錄”即指圖像、視頻和語音等數位內容。

l 歐盟

2019 年 4 月 8 日發 布《人工智慧道德準則》，將隱私和資料管理作為可信賴人工智慧需要滿足的七個要素之一。

**5、視覺深度偽造檢測技術**

Deepfake的快速發展和應用，給個人隱私資料、社會穩定和國家安全等造成了潛在威脅，針對深度偽造內容的檢測和防禦現已成為世界各國政府、企業乃至個人所關注的熱點問題之一。

**5.1、Detection**

檢測深度偽造（和人相關）的方法通常分為兩大類：偽影識別法和無定向法。

**5.1.1、偽影識別法（Artifact-Specific）**

基本原理是通過搜索特定類型的偽影，人眼對偽影或許不易察覺，但機器學習和取證分析法容易檢測，以下定義7種類型偽影。

l 融合（Blending (spatial)）

生成的內容重新融合到圖像幀時會產生一些偽影，檢測方法比如邊界檢測、品質度量、頻率分析。

l 環境（Environment (spatial)）

偽造的臉部內容和圖像幀的剩餘部分可能是不協調的，比如面部變形過程中的殘差、光照、保真度變化。

l 取證（Forensics (spatial)）

分析模型在偽造品中留下的細微特徵和樣式，比如GAN會留下獨特的指紋可能用於識別生成器、分析相機的獨特感測器雜訊（PRNU）識別粘貼的內容、尋找視頻中幀序列的殘差、尋找缺陷並預測和監測臉部特徵點（如頭部姿勢往往不一致）。

l 行為（Behavior (temporal)）

基於目標人物大量的資料，監測舉止和其它行為異常，比如對已錄製的目標人物素材庫建模、在沒有目標素材參考的情況下檢測音視頻片段中感知的情緒的差異。

l 生理（Physiology (temporal )）

基於生成的內容缺少生理信號的假設，比如監測心率識別偽造的面部、監測皮膚下血容量（脈搏）、監測不規則的眨眼模式，相反也有利用脈搏信號構建deepfake模型。

l 同步（Synchronization (temporal)）

不一致也是一個揭示因數，比如把把語音和嘴巴附近的特徵點相關聯檢測視頻配音攻擊、檢測嘴型和語音因素的不一致（如在嘴巴完全閉合的（B,P,M）音素上deepfake往往會失敗）。

l 連貫（Coherence (temporal)）

實際的時間連貫性很難偽造產生，由此檢測產生的偽影，比如使用RNN檢測閃爍和抖動、使用LSTM只檢測面部區域、使用成對的順序幀訓練分類器以及改進網路專注於監測幀的光流、（同一個作者）使用LSTM預測下一幀並監測重建誤差。

**5.1.2、無定向法（Undirected Approaches）**

不同于專注於一種特定的偽影，無定向法通過訓練通用的分類器，讓神經網路決定分析的特徵。通常，研究者採用兩種方法：分類器和異常檢測。

l 分類（Classification）

對於壓縮圖像，深度神經網路表現的比傳統的圖像取證工具更好。各種研究者展示了標準的CNN架構如何有效的檢測偽造視頻，比如使用真假圖片對訓練利用CNN構建的孿生網路、使用HMN網路架構同時考慮面部和以前見過的面部內容（出於擔憂CNN可能僅能檢測訓練過的攻擊）、使用全家桶方式將7種深度偽造CNN的預測結果輸入至元分類器。

使用分類器檢測偽造也存在問題，因為通過對抗性機器學習可以逃避檢測。

l 異常檢測（Anomaly Detection）

同分類相反，異常檢測模型使用正常資料進行訓練，部署期間檢測異常值，這些方法不假設攻擊的形式，對於未知的攻擊創建方法泛化性更好，比如在人臉識別網路中度量神經元啟動（相對於使用原始圖元可獲得更強的信號，能夠克服雜訊和其它失真）、訓練一類VAE網路用於重建真實圖像並對新圖像計算其和重建圖像兩者均值的MSE、利用ED隱空間度量輸入圖像和真實樣本的embedding距離、不直接使用神經網路而使用最新的基於置信度指標（ABC）的屬性（為了檢測偽造圖像，ABC用於確定圖像是否適合預訓練的人臉識別網路的訓練分佈）。

**5.2、Fake Video Detection**

大多數圖像檢測方法不能用於視頻，因為視訊壓縮後幀資料會嚴重退化。視頻具有在幀組之間變化的時間特性，對於設計為僅檢測靜態圖像的方法具有挑戰性。

使用跨視頻幀的時間模式的方法主要基於深度遞迴網路模型來檢測deepfake視頻，可分為兩類：採用時間特徵的方法和探索幀內視覺偽影的方法。

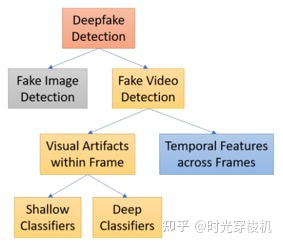


圖5-1 深度偽造視頻檢測

**5.2.1、Temporal Features across Video Frames**

利用視頻流的時空特徵來檢測深度偽造，視頻操作是在逐幀的基礎上執行的，可以認為由面部操作產生的低級偽影會進一步表現為跨幀不一致的時間偽影。

l 迴圈卷積模型（RCN）

基於卷積網路DenseNet和門控迴圈單位單元的集成，以利用幀之間的時間差異。

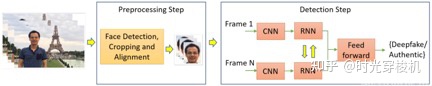


圖5-2 RCN檢測過程

檢測分兩步，預處理步驟包括檢測、裁剪和對齊一系列幀上的面部，第二步通過結合卷積神經網路（CNN）和迴圈神經網路（RNN）來區分真假面部圖像。

l 時間感知管線

強調深層視頻包含幀內不一致和幀之間的時間不一致，使用CNN和長期短期記憶（LSTM）來檢測Deepfake視頻。

CNN用於提取幀級特徵，並將其饋入LSTM以創建時間序列描述符。 之後，使用一個完全連接的網路根據序列描述符對真實視頻中的篡改視頻進行分類。

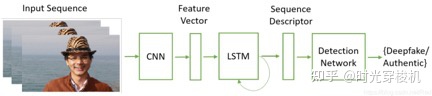


圖5-3 基於時間感知管線檢測過程

使用卷積神經網路（CNN）和長期短期記憶（LSTM）提取給定視頻序列的時間特徵（通過序列描述符表示）。 由全連接層組成的檢測網路被用來作為序列描述符的輸入，並計算屬於真假幀序列的概率。

**5.2.2、Visual Artifacts within Video Frame**

通常將視頻分解為幀並探索單個幀內的視覺偽影以獲得判別特徵， 再將這些功能分配到深層或淺層分類器中以區分真假視頻。根據分類器的類型（即深層或淺層）對方法分組。

l 深分類器（Deep classifiers）

Deepfake視頻通常以有限的解析度創建，需要仿射人臉變形方法（縮放、旋轉、剪切）以匹配原始視頻的配置（比如解析度）。由於扭曲的面部區域和周圍環境的解析度不一致，此過程留下了CNN模型可檢測到的偽影。

l 淺分類器（Shallow classifiers）

Deepfake檢測方法主要依賴於偽造圖像與真實圖像或視頻之間的固有特徵的偽影或不一致。

由於Deepfake在人臉生成管道中存在缺陷，有通過觀察3D頭部姿勢之間的差異（包括頭部方向和位置）來檢測，該方法基於中央面部區域的68個面部界標進行估算，檢查3D頭部姿勢，提取的特徵被饋送到SVM分類器中以獲得檢測結果。

有基於眼睛、牙齒和面部輪廓的視覺特徵來檢測偽影。視覺偽影是由於缺乏整體一致性，入射照明的錯誤或不精確估計或底層幾何結構的不精確估計而引起的。

還有各種基於特定類型偽影的檢測方法。

**6、深度偽造預防與緩解**

l 資料來源（Data Provenance）

為了預防深度偽造，有些建議使用分散式帳本和區塊鏈網路追蹤多媒體資料來源；相反，另一些建議利用乙太坊智慧合約像一個全域檔案系統那樣認證和管理內容。這些提議背後的假設是：一個內容只有在其來源可追溯時才被認為是真實的、可信的。

l 反擊（Counter Attacks）

為了反擊深度偽造，有些使用對抗性機器學習擾亂和破壞深度偽造網路的，比如添加嚴重的雜訊擾動以阻止深度偽造正確的定位面部、使用對抗性雜訊改變面部身份以使網路爬蟲找不到目標訓練圖像而無法訓練。

**7、展望**

**7.1、深度偽造製作趨勢**

l 趨向研究身份不可知模型和高解析度的深度偽造

（1）不成對的自監督訓練技術以減少初始訓練資料量；

（2）小樣本學習以實現單張照片盜竊身份；

預先用大量資料集進行元訓練（meta learning）獲得小樣本學習能力，通過元訓練學習到怎樣快速生成圖像的能力，而不是具體生成某一特定類別圖像的能力。

（3）通過AdaIN層、解耦和pix2pixHD網路元件改善面部品質和身份；

（4）通過時間判別器和光流預測提升視頻的流暢度和逼真度；

（5）使用輔助網路以減輕邊界偽影將合成物無縫融合到圖像。

l 其它方面取得的重大進展

（1）在預訓練的VGG人臉識別網路上使用感知損失

可顯著提升面部品質，已應用在流行的線上deepfake工具。

（2）使用網路管道流水線

不是在單個網路上強制實施一組全域損失，而是使用網路管道，其中每個網路都承擔著不同的責任（轉換、生成、遮擋、混合等）以更好的控制最終輸出，應對泛化性的挑戰。

（3）即時深度偽造是一種趨勢

一些研究以30fps實現即時偽造，但當前即時深度偽造的逼真度還很不夠。

**7.2、深度偽造攻防競技**

深度偽造攻擊方和防禦方相互博弈，大多數deepfake檢測演算法假定和對手是靜態博弈，它們或專注於識別特定的偽影、或在應對新的和沒見過的攻擊時泛化性不足，最近隨著偽造品質量的提升，表現很好的檢測器正迅速減少。

l 攻擊方避開基於偽影的檢測

減輕單個缺陷以逃避偽影檢測，比如添加判別器使得生成的生物信號受控、增加損失函數避免廣泛的神經元啟動使其覆蓋範圍最小化、重現整個頭部或從相同的資料來源中學習舉止以逃避檢測異常的姿勢和舉止、識別模糊內容的模型受雜訊和GAN銳化影響、一些搜索人臉融合邊界的模型不適用於經過圖像修復或輸出完整幀的改進的網路、將生成的偽造品傳遞給篩檢程式或執行物理複製或壓縮以逃避取證搜索法（或至少提高錯誤報警率）。

l 攻擊方避開深度學習分類器

使用對抗性學習添加擾動來逃避深度學習的檢測，這些攻擊會跨多個模型轉移而無論所使用的訓練資料，最近的進展表明這些攻擊對未知的分類器和訓練集也有效。

l 防禦方主被動防禦相結合

防禦方暫時提供了適度的被動防禦；深度偽造製作仍不完善，這些局限性提升了攻擊的難度閾值，有時可能因為太耗時間和資源使得普通攻擊者無法創建足夠的偽造而逃避檢測。

基於內容的防禦對策是被動的、不可持續的，需要提供基於非內容的主動防禦措施，比如建立線上視頻內容來源和真實性的框架、應用對抗性機器學習來保護內容免遭篡改。

**7.3、新威脅與應對措施**

l 即將出現的威脅

（1）越來越多的偽造品會變成攻擊個人或公司的貨幣化武器；

（2）深度偽造越來越實用和有效，即時深度偽造越來越逼真；

駭客組織將會利用深度偽造進行偵查作為APT（高級長期威脅）的一部分，國家級行為者通過重演官員或家庭成員進行間諜活動和破壞活動。

l 主動應對（保持領先）

（1）考慮攻擊對手的下一步，而不只是當前攻擊的弱點

評估這些攻擊的理論極限，比如，找出模型延遲的邊界以檢測即時攻擊、確定GAN的限制以設計適當的策略。

（2）探索當前deepfake檢測器的弱點和局限性

通過識別和瞭解了這些漏洞，制定出更強大的對策。

**8、給我們帶來的幾點啟示**

l 技術層面

（1）理解製作和防範深度偽造的技術原理、優缺點有必要性。

（2）可作為AI技術在具體場景工程化綜合運用的學習案例。

l 業務層面

（1）挑戰：視覺深度偽造製作正逐漸成為多媒體創作的一種新形式，其潛在危害性值得關注；多媒體隱私資料的保護越來越值得重視，在網路上傳播個人多媒體資訊的潛在危害性越來越大。

（2）機遇：視覺AI技術在多媒體AI工具方向上有一定的探索嘗試空間，比如表情包製作、虛擬人物製作等。

**9、參考文獻**

[1] Deepfake：[https://en.wikipedia.org/wiki/Deepfake](https://link.zhihu.com/?target=https%3A//en.wikipedia.org/wiki/Deepfake" \t "_blank)

[2]《The Creation and Detection of Deepfakes: A Survey》：

[https://arxiv.org/pdf/2004.11138.pdf](https://link.zhihu.com/?target=https%3A//arxiv.org/pdf/2004.11138.pdf)，2020.4.

<https://zhuanlan.zhihu.com/p/139489768>.

[3]《Deep Learning for Deepfakes Creation and Detection: A Survey》：

[https://arxiv.org/abs/1909.11573](https://link.zhihu.com/?target=https%3A//arxiv.org/abs/1909.11573)，2019.9,

[https://blog.csdn.net/Reddoge\_/article/details/103822699](https://link.zhihu.com/?target=https%3A//blog.csdn.net/Reddoge_/article/details/103822699).

[4]《DeepFaceLab: A simple, flexible and extensible face swapping framework》：[https://arxiv.org/pdf/2005.05535.pdf](https://link.zhihu.com/?target=https%3A//arxiv.org/pdf/2005.05535.pdf" \t "_blank)，2020.5.

[5]《DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection》: [https://arxiv.org/pdf/2001.00179.pdf](https://link.zhihu.com/?target=https%3A//arxiv.org/pdf/2001.00179.pdf)，2020.1,

<https://zhuanlan.zhihu.com/p/115070797>,

<https://zhuanlan.zhihu.com/p/92474937>.

[6]《視聽覺深度偽造檢測技術研究綜述》：[http://jcs.iie.ac.cn/xxaqxb/ch/reader/create\_pdf.aspx?file\_no=20200202](https://link.zhihu.com/?target=http%3A//jcs.iie.ac.cn/xxaqxb/ch/reader/create_pdf.aspx%3Ffile_no%3D20200202" \t "_blank)，2020.4.

[7]《從文本合成真實的唇語口型》：

[https://www.jiqizhixin.com/articles/2018-01-14-5](https://link.zhihu.com/?target=https%3A//www.jiqizhixin.com/articles/2018-01-14-5).

[8]《Buzzfeed Created a ‘Deepfake’ Obama PSA Video》：[https://www.extremetech.com/extreme/267771-buzzfeed-created-a-deepfake-obama-psa-video](https://link.zhihu.com/?target=https%3A//www.extremetech.com/extreme/267771-buzzfeed-created-a-deepfake-obama-psa-video" \t "_blank)，2018.4.

[9] 《Lyrebird claims it can recreate any voice using just one minute of sample audio》：[https://www.theverge.com/2017/4/24/15406882/ai-voice-synthesis-copy-human-speech-lyrebird](https://link.zhihu.com/?target=https%3A//www.theverge.com/2017/4/24/15406882/ai-voice-synthesis-copy-human-speech-lyrebird" \t "_blank)，2017.4.

[10] 《Realistic Deepfakes in 5 Minutes on Colab》：

[https://towardsdatascience.com/realistic-deepfakes-colab-e13ef7b2bba7](https://link.zhihu.com/?target=https%3A//towardsdatascience.com/realistic-deepfakes-colab-e13ef7b2bba7)，2020.3.

[11] 《Face2Face: Real-time Face Capture and Reenactment of RGB Videos》：[http://niessnerlab.org/projects/thies2016face.html](https://link.zhihu.com/?target=http%3A//niessnerlab.org/projects/thies2016face.html" \t "_blank)，2016.6.

[12] 《Adobe is working on an audio app that lets you add words someone never said》：[https://www.theverge.com/2016/11/3/13514088/adobe-photoshop-audio-project-voco](https://link.zhihu.com/?target=https%3A//www.theverge.com/2016/11/3/13514088/adobe-photoshop-audio-project-voco" \t "_blank)，2016.11.

[13]《Family fun with deepfakes. Or how I got my wife onto the Tonight Show》：[https://towardsdatascience.com/family-fun-with-deepfakes-or-how-i-got-my-wife-onto-the-tonight-show-a4454775c011](https://link.zhihu.com/?target=https%3A//towardsdatascience.com/family-fun-with-deepfakes-or-how-i-got-my-wife-onto-the-tonight-show-a4454775c011" \t "_blank)，2018.2.

[14]《Lip-syncing Obama: New tools turn audio clips into realistic video》：[https://www.washington.edu/news/2017/07/11/lip-syncing-obama-new-tools-turn-audio-clips-into-realistic-video/](https://link.zhihu.com/?target=https%3A//www.washington.edu/news/2017/07/11/lip-syncing-obama-new-tools-turn-audio-clips-into-realistic-video/" \t "_blank)，2017.6.

[15]《Synthesizing Obama: Learning Lip Sync from Audio》: [http://grail.cs.washington.edu/projects/AudioToObama/siggraph17\_obama.pdf?\_ga=2.261732173.1517641146.1606723485-1452069613.1606723485](https://link.zhihu.com/?target=http%3A//grail.cs.washington.edu/projects/AudioToObama/siggraph17_obama.pdf%3F_ga%3D2.261732173.1517641146.1606723485-1452069613.1606723485)，2017.7.

[16] 《Deepfake technology is changing fast - use these 5 resources to keep up》：[https://journalistsresource.org/studies/society/deepfake-technology-5-resources/](https://link.zhihu.com/?target=https%3A//journalistsresource.org/studies/society/deepfake-technology-5-resources/" \t "_blank)，2019.6.

[17] 《Generative Adversarial Nets》：[https://papers.nips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf](https://link.zhihu.com/?target=https%3A//papers.nips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf" \t "_blank)，2014.6.

[18]《Everybody Dance Now》：

[https://arxiv.org/pdf/1808.07371.pdf](https://link.zhihu.com/?target=https%3A//arxiv.org/pdf/1808.07371.pdf)，2018.8.

[19] 《A Style-Based Generator Architecture for Generative Adversarial Networks 1-2合集》：[https://nvlabs.github.io/stylegan2/versions.html](https://link.zhihu.com/?target=https%3A//nvlabs.github.io/stylegan2/versions.html)，2019.3.

[20] 《Few-Shot Adversarial Learning of Realistic Neural Talking Head Models》：[https://nvlabs.github.io/stylegan2/versions.html](https://link.zhihu.com/?target=https%3A//nvlabs.github.io/stylegan2/versions.html" \t "_blank)，2019.5.

[https://blog.csdn.net/a312863063/article/details/90728818](https://link.zhihu.com/?target=https%3A//blog.csdn.net/a312863063/article/details/90728818)，2019.6.

[https://cloud.tencent.com/developer/article/1477496](https://link.zhihu.com/?target=https%3A//cloud.tencent.com/developer/article/1477496)，2019.7.

[21] 《What Is a Deepfake?》：

[https://www.pcmag.com/news/what-is-a-deepfake](https://link.zhihu.com/?target=https%3A//www.pcmag.com/news/what-is-a-deepfake)，2020.3.

[22] 《Creating a dataset and a challenge for deepfakes》：

[https://ai.facebook.com/blog/deepfake-detection-challenge/](https://link.zhihu.com/?target=https%3A//ai.facebook.com/blog/deepfake-detection-challenge/)，2019.9.

[23] 《深度偽造與檢測技術綜述》：[http://www.jos.org.cn/jos/ch/reader/create\_pdf.aspx?file\_no=6140&journal\_id=jos](https://link.zhihu.com/?target=http%3A//www.jos.org.cn/jos/ch/reader/create_pdf.aspx%3Ffile_no%3D6140%26journal_id%3Djos)，2020.

[24] 《DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection》：[https://arxiv.org/pdf/2001.00179.pdf](https://link.zhihu.com/?target=https%3A//arxiv.org/pdf/2001.00179.pdf)，2020.

[25] Visual artifact: [https://en.wikipedia.org/wiki/Visual\_artifact](https://link.zhihu.com/?target=https%3A//en.wikipedia.org/wiki/Visual_artifact)

[28] 參見“楊冪朱茵‘換臉’了？變臉如此簡單到底有多可怕”：[https://www.sohu.com/a/299665489\_105446](https://link.zhihu.com/?target=https%3A//www.sohu.com/a/299665489_105446).

[29] 參見“私密空間偷拍，AI換臉？看看民法典人格權編如何回應”：[https://www.thepaper.cn/newsDetail\_forward\_7708586](https://link.zhihu.com/?target=https%3A//www.thepaper.cn/newsDetail_forward_7708586).

[30] 《Detect AI-generated Images & Deepfakes》：

[https://jonathan-hui.medium.com/detect-ai-generated-images-deepfakes-part-1-b518ed5075f4](https://link.zhihu.com/?target=https%3A//jonathan-hui.medium.com/detect-ai-generated-images-deepfakes-part-1-b518ed5075f4)，2020.

[31] 《First Order Motion Model for Image Animation》：

[https://proceedings.neurips.cc/paper/2019/hash/31c0b36aef265d9221af80872ceb62f9-Abstract.html](https://link.zhihu.com/?target=https%3A//proceedings.neurips.cc/paper/2019/hash/31c0b36aef265d9221af80872ceb62f9-Abstract.html)，2019，

[https://mp.weixin.qq.com/s?\_\_biz=Mzg4NDQwNTI0OQ==&mid=2247522761&idx=1&sn=60ad6b8154dc05dddfa36b9e2975982c&source=41#wechat\_redirect](https://link.zhihu.com/?target=https%3A//mp.weixin.qq.com/s%3F__biz%3DMzg4NDQwNTI0OQ%3D%3D%26mid%3D2247522761%26idx%3D1%26sn%3D60ad6b8154dc05dddfa36b9e2975982c%26source%3D41%23wechat_redirect)，2020.9.

[https://github.com/AliaksandrSiarohin/first-order-model](https://link.zhihu.com/?target=https%3A//github.com/AliaksandrSiarohin/first-order-model).