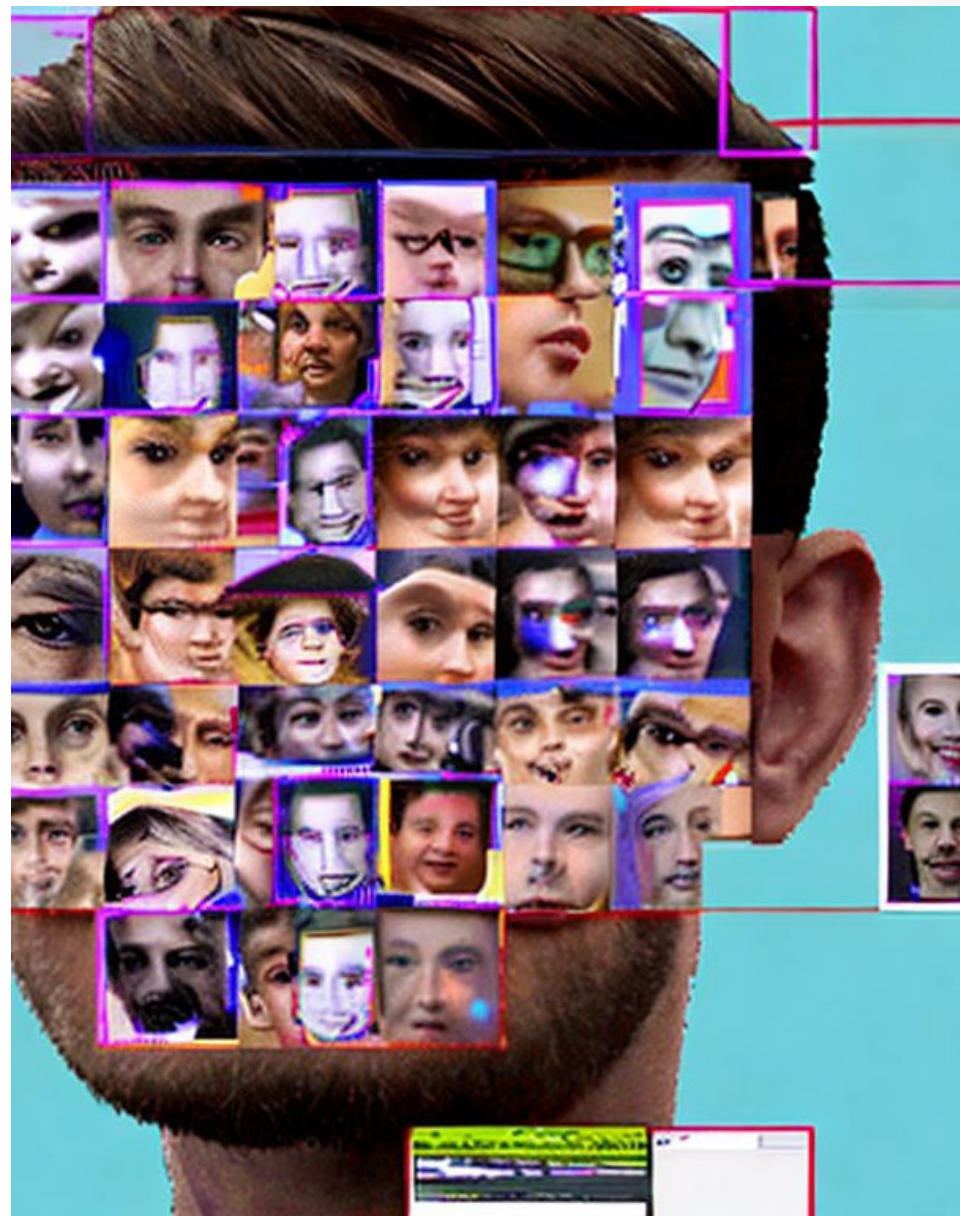




它在一本正經地胡說八道嗎？  
以子之矛，攻子之盾：仿製、  
實作、驗證

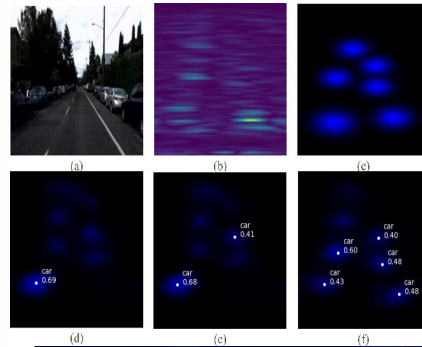
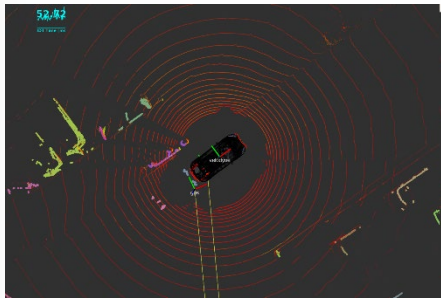
Chih-Chung Hsu (許志仲)

Associate Professor  
Leader of Advanced Computer Vision Lab  
Institute of Intelligent Systems  
College of Artificial Intelligence  
National Yang Ming Chiao Tung University

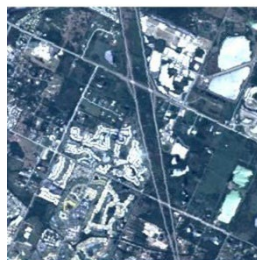
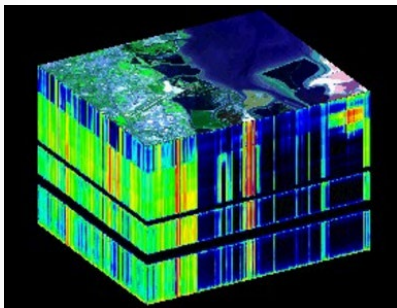
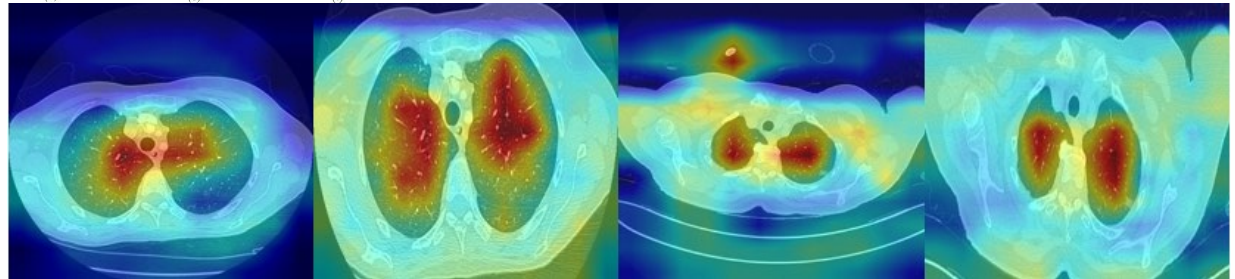
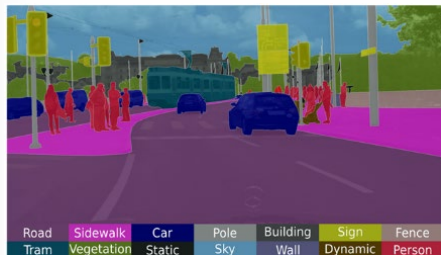


# Advanced Computer Vision Laboratory

- Deep learning-based image processing and computer vision
  - DeepFake detection, Super-resolution, Hyperspectral image processing, Perception of ADAS, Medical signal analysis.



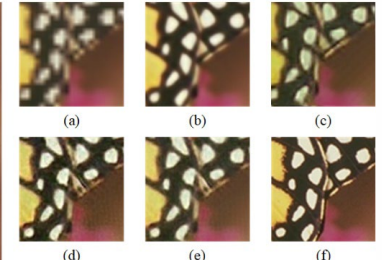
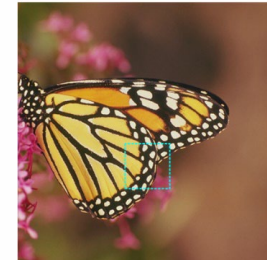
Real/Fake



C-Type Ground Truth



1% Super-resolved image



# What will we cover?

## ▪ Multimedia Forensics: DeepFake

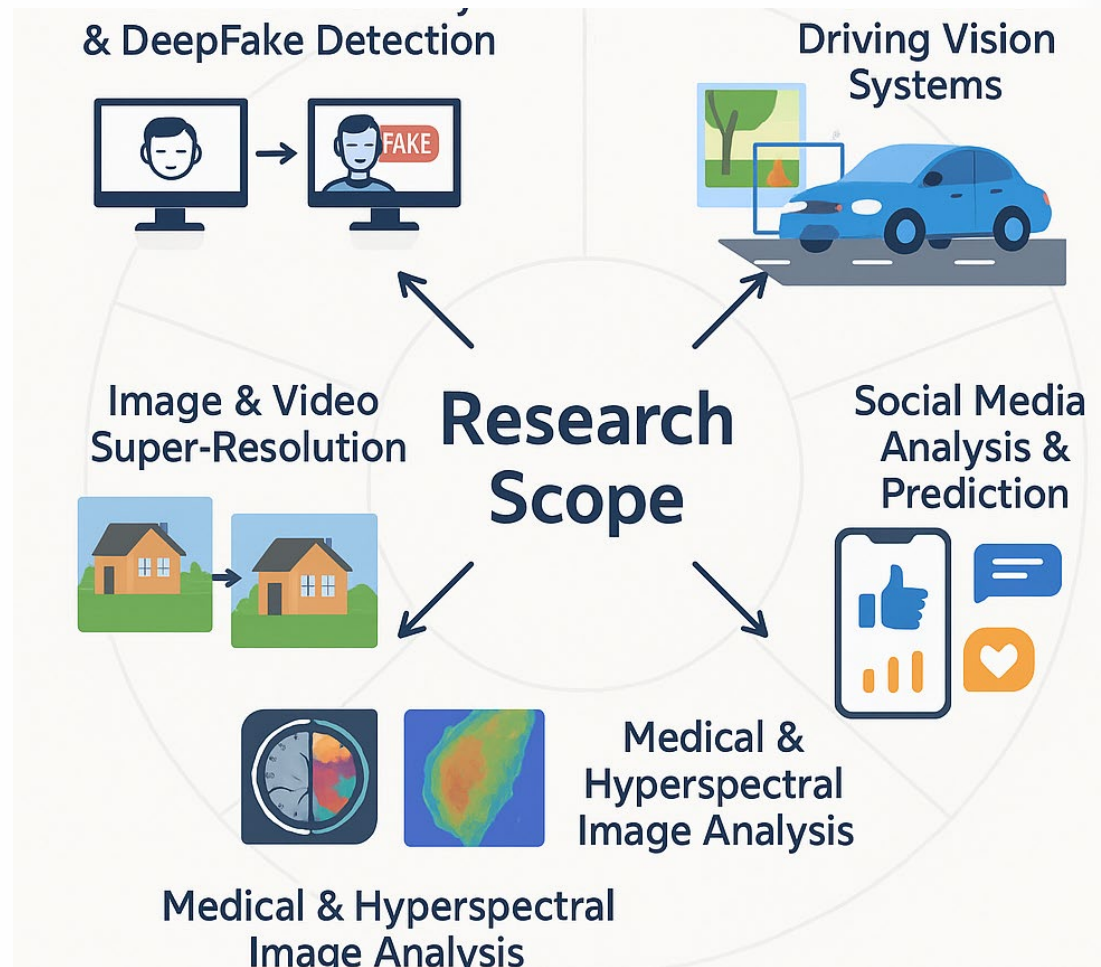
- Image Forensics
- Video Forensics
- Hype-rForensics

## ▪ Robust and Lightweight Restoration

- Satellite image processing
- Fusion under noised conditions
- Transmission of corrupted data

## ▪ Real-world Scenarios

- Toward resource-limited vision tasks
- E.g. NV Jetson series, TI TDA4VM



# 今天的主題

## 科普介紹 –

- 從AI到生成式模型、從生成式模型到偵測生成影視訊

## 微學術內容 –

- 這些技術到底能拿來做什麼關鍵應用？技術背景？

## 實務操作 –

- 人眼看的出來嗎？
- 自己能換臉嗎？
- 怎麼偵測換臉？

# Deepfake簡介

## ? 什麼是Deepfake ?

**Deepfake** 是結合「Deep Learning」( 深度學習 ) 和「Fake」( 偽造 ) 的技術，利用人工智慧演算法將一個人的臉部特徵合成到另一個人的影像中。

主要基於**生成對抗網路**( GAN ) 和**自編碼器**

( Autoencoder ) 等深度學習模型。

## 📈 市場規模與趨勢

Deepfake AI市場在2023年估值約5.5億美元，預計到2032年將以40.5%的年複合增長率成長。



## 🕒 Deepfake的發展歷史

**2017年**

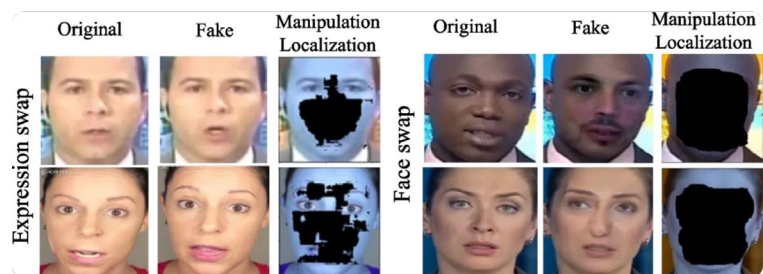
Reddit用戶「deepfakes」發布首個換臉應用

**2018年**

技術開始普及，出現多種開源工具

**2020年**

技術成熟，能夠即時生成高品質換臉影片



# Deepfake應用實例

## ✂ 常見的Deepfake應用工具



### DeepSwap

線上換臉工具，支援照片和影片，操作簡單易用



### Roop

開源換臉工具，只需一張照片即可進行影片換臉



### FaceSwap

功能強大的開源Deepfake工具，支援多種模型



### DeepFacer

手機應用程式，提供簡易的換臉功能

## 📺 換臉技術的操作方式

- 1 **選擇素材**: 上傳源人臉圖片和目標影片
- 2 **臉部偵測**: 系統自動識別並標記人臉位置
- 3 **參數調整**: 設定模型參數，如臉部匹配度、平滑度
- 4 **生成處理**: AI模型進行換臉處理，可能需要一定時間
- 5 **後期優化**: 調整色調、光影等細節，提高真實感



# Deepfake的社會影響

## + 正面應用



**娛樂產業**：電影特效、虛擬角色創作、已故演員復活演出



**教育應用**：歷史人物再現、多語言教學內容、虛擬講師

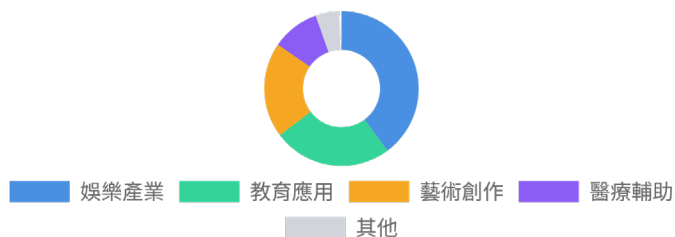


**藝術創作**：新型態數位藝術、互動裝置、創意表現媒介



**醫療輔助**：醫學影像模擬、臨床訓練、整形手術預覽

Deepfake正面應用分布



## - 負面影響



**假新聞與虛假資訊**：政治人物發表不實言論、製造社會混亂



**詐騙與身份盜用**：冒充親友詐騙、語音詐騙、商業欺詐

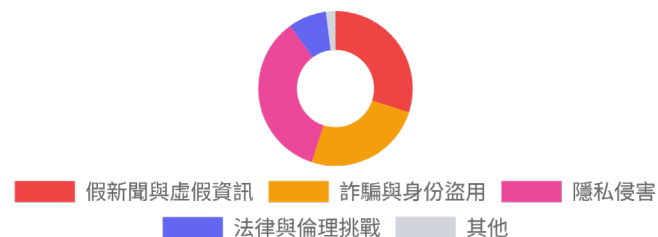


**隱私侵害**：未經同意的換臉色情內容、名譽損害



**法律與倫理挑戰**：著作權爭議、責任歸屬、監管困難

Deepfake負面影響分布



# 生成模型的好與壞

- 但就像世界上所有的事物一樣，有好有壞
  - 技術本身是中立的，但使用者卻**不一定**
- **案件激增**：深偽技術被不法之徒利用於詐騙呈爆炸式增長。全球深偽詐欺事件2022-23年間增加超過**10倍** [\[link\]](#)
  - 金融領域受衝擊尤其明顯，**2023年金融科技業的深偽事件增加了700%** [\[link\]](#)
  - 詐騙集團藉由AI偽造聲音影像行騙，傳統防詐機制備受挑戰。
- **損失擴大**：2023年因深偽冒充造成全球損失估計高達**27億美元** [\[link\]](#)
  - 個案金額從數百到數千美元不等，中位數約\$1000
  - 更有大型企業因此遭受巨額詐騙，如香港一公司被假冒主管的深偽視訊會議騙走**2500萬美元** [\[link\]](#)
  - 各界預測若無有力對策，未來損失將繼續攀升（估計美國2027年因AI詐欺損失可達\$400億美金，較2023年翻數倍 [\[link\]](#)）。

# 名人投資詐騙影片 (Celebrity Investment Scam Video)

## ■ 手法說明：2023年11月出現AI深偽投資詐騙廣告

- 騙徒將新聞片段後製，偽裝成知名人士（總統蔡英文、企業家張忠謀、馬斯克等）在推薦一個“政府核准”的加密貨幣投資平台。影片透過**換嘴型與變聲**，讓名人看似親口保證高額回報 [\[link\]](#)

## ■ 查證結果：查核團隊(MyGoPen等)確認該影片為假，名人從未代言此平台

- 這是假訊息與詐騙結合的典型案列，利用名人公信力誘使民眾受騙。警方已呼籲大眾提高警覺，勿信社群上由名人“親述”的投資廣告，必要時可撥165反詐騙專線查詢



# Stable Diffusion介紹

## 📌 什麼是Stable Diffusion？

Stable Diffusion是由**Stability AI**於2022年發布的開源擴散模型，專門用於生成高品質影像。

它是目前最受歡迎的開源AI影像生成模型之一，可以根據文字描述（提示詞）生成各種風格的影像。



## ★ Stable Diffusion的特點

- 🔗 **開源**：完全開源，可自由下載、修改和部署
- 💻 **本地運行**：可在個人電腦上運行，保護隱私
- 🧩 **可擴展**：支持LoRA、ControlNet等擴展模型
- 👥 **活躍社群**：龐大的用戶和開發者社群支持
- 🔄 **快速迭代**：從1.5版本到3.0版本持續改進

## ⚙️ Stable Diffusion的架構

Stable Diffusion採用**潛在擴散**架構，主要由三個關鍵組件組成：

- 🔗 **VAE編碼器/解碼器**：將影像壓縮到潛在空間並還原
- 📄 **文本編碼器**：將文本提示轉換為向量表示
- 🔗 **U-Net去噪器**：在潛在空間中執行擴散過程

## 🔄 文字轉影像的工作流程

- 1 使用CLIP模型將文字提示轉換為嵌入向量
- 2 從隨機噪聲開始，在潛在空間中進行擴散
- 3 U-Net在文字嵌入的條件下預測並去除噪聲
- 4 VAE解碼器將潛在表示轉換回像素空間
- 5 輸出最終生成的影像

# Stable Diffusion使用教學

## 基本提示詞技巧

提示詞 ( Prompt ) 是指導Stable Diffusion生成影像的文字描述，好的提示詞能顯著提升生成品質。

A beautiful landscape with mountains, lake, sunset, photorealistic, 8k, detailed

### 💡 提示詞撰寫技巧

- 使用具體、詳細的描述
- 加入風格關鍵詞 ( 如油畫、攝影風格 )
- 指定畫質和細節程度 ( 如8k、高細節 )
- 使用權重調整關鍵詞重要性 ( 如 (keyword:1.2) )

### ⚠️ 負面提示詞

使用負面提示詞 ( Negative Prompt ) 來避免不想要的元素：

blurry, low quality, deformed, ugly, bad anatomy

## 常用擴展功能

### ControlNet

參考影像控制生成結果的姿勢、線條、深度等

### Textual Inversion

學習特定概念，使用自定義詞彙生成

### LoRA

低秩適應模型，用於添加特定風格或主題

### img2img

以參考影像為基礎進行生成或修改

## 參數調整與控制

**CFG Scale** 提示詞遵循度，值越高越遵循提示詞 ( 通常7-12 )

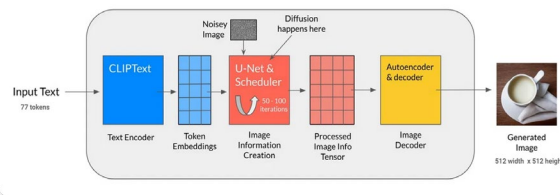
**Steps** 去噪步數，步數越多細節越豐富 ( 通常20-50步 )

**Sampler** 採樣器類型，影響生成風格和速度 ( 如DPM++, Euler a )

**Seed** 隨機種子，固定種子可重現相同結果

**Size** 影像尺寸，影響生成品質和顯存需求

Stable Diffusion Architecture



## 常見使用平台

### Stable Diffusion WebUI

最受歡迎的開源界面，功能豐富

### DreamStudio

官方網頁版，使用簡單但功能較少

### ComfyUI

節點式界面，靈活度高，適合進階用戶

### Google Colab

雲端運行，適合無高階顯卡的用戶

# Midjourney介紹

## 📌 什麼是Midjourney？

Midjourney是一款基於AI的影像生成工具，由Midjourney公司開發，於2022年推出。它以**藝術性強**、**視覺效果震撼**而聞名

不同於Stable Diffusion，Midjourney是**封閉源碼**的商業服務，主要藉由Discord平台或官方網頁版使用。



## 📱 使用方式

Midjourney主要藉由**Discord機器**或**官方網頁版**使用：

/imagine prompt: [你的提示詞描述]

### U1-U4

放大指定影像

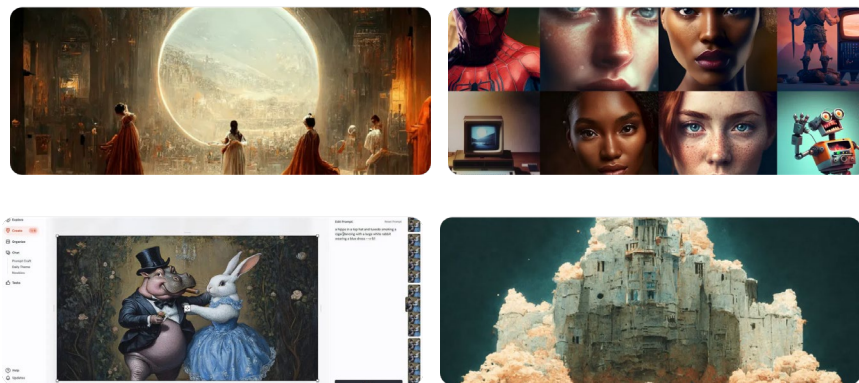
### V1-V4

變化指定影像

## ★ Midjourney的特點與優勢

- 🧠 **卓越的藝術性** 生成的影像具有高度藝術感和美學價值
- 🖋️ **直覺的提示詞** 對簡單提示詞也能產生優質結果
- ☁️ **雲端運算** 無需本地硬體，直接在雲端生成影像
- 🔄 **快速迭代** 從V1到V6版本持續提升生成品質

## 🖼️ 案例展示



廣泛應用於**概念藝術**、**插畫設計**、**品牌視覺**等領域

# DALL-E介紹

## 📌 什麼是DALL-E？

DALL-E是由 **OpenAI** 開發的AI影像生成模型，名稱結合了藝術家「Salvador Dalí」和機器人「WALL-E」。

它是一個**封閉源碼**的商業服務，能夠根據文字描述生成高度逼真和創意的影像，與ChatGPT整合緊密。



## 🔗 DALL-E的演進

### DALL-E (2021)

首個版本，基於GPT-3架構，能生成低解析度影像

### DALL-E 2 (2022)

大幅提升影像品質和解析度，加入編輯功能

### DALL-E 3 (2023)

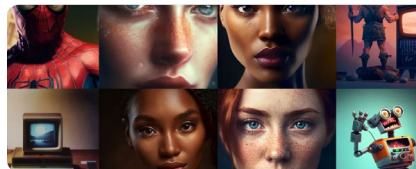
與ChatGPT整合，提升文字理解力，支援更複雜的場景

## ★ DALL-E的特點與優勢

- 🧠 **強大的理解力** 能精確理解複雜的文字描述
- 🛡️ **內容安全** 內建內容過濾和安全措施
- ✂️ **編輯功能** 支援局部編輯和擴展畫布
- 🔌 **API整合** 提供開發者API，易於整合到應用中

## 🔧 使用方式與案例

DALL-E可藉由**OpenAI官網**、**ChatGPT**或**API**使用：



# Diffusion Model影片生成

## 影片生成技術原理

擴散模型的影片生成技術主要基於**時空一致性**的擴展：

- 時間維度擴展**  
將影片視為時間序列上的影像集合，保持幀間連貫性
- 運動一致性**  
使用運動模型或光流預測確保物體移動自然流暢



## 常見應用場景

- |                             |                            |
|-----------------------------|----------------------------|
| <b>短影片創作</b><br>社交媒體短片、創意表達 | <b>視覺效果</b><br>電影特效、視覺藝術創作 |
| <b>教育內容</b><br>教學動畫、概念視覺化   | <b>廣告製作</b><br>產品展示、品牌宣傳   |
- 目前生成時長通常在 **數秒至數十秒** 之間，較長影片需要分段生成後合併。

## 主要影片生成工具

- Runway Gen-2**  
支援文字/影像轉影片，生成品質高，商業服務
- Pika Labs**  
基於Discord的影片生成服務，操作簡單
- AnimateDiff / Stable Video Diffusion**  
開源影片生成模型，可與Stable Diffusion整合

## 技術挑戰與發展趨勢

- |                                                                                                                            |
|----------------------------------------------------------------------------------------------------------------------------|
| <b>當前挑戰</b> <ul style="list-style-type: none"><li>時間一致性維持困難</li><li>生成時長有限</li><li>計算資源需求高</li><li>細節控制不夠精確</li></ul>      |
| <b>未來發展趨勢</b> <ul style="list-style-type: none"><li>更長時間的連貫影片</li><li>更精確的場景控制</li><li>與3D技術的結合</li><li>即時生成能力提升</li></ul> |

# Deepfake深度學習偵測

## 深度學習偵測模型

### 卷積神經網路 (CNN)

分析影像中的空間特徵，識別Deepfake生成過程中留下的視覺痕跡

### 循環神經網路 (RNN)

分析視頻序列中的時間特徵，檢測幀間不一致性和異常變化

### 生成對抗網路 (GAN)

使用對抗性訓練來提高偵測模型的魯棒性，應對新型Deepfake

### 變換器模型 (Transformer)

利用自注意力機制捕捉長距離依賴關係，提高偵測準確率

## 深度學習偵測效能

### 深度學習偵測模型效能比較



深度學習模型在Deepfake偵測中表現優異，特別是針對特定類型的Deepfake進行專門訓練時。然而，對於 **未見過的新型Deepfake**，準確率可能會顯著下降。

## 主流偵測工具

### Microsoft Video Authenticator

提供實時分析和信心分數，檢測操縱痕跡

### Sensity AI

專業Deepfake偵測平台，提供API和企業解決方案

### DeepDetect

開源深度學習偵測框架，支持多種模型

## 未來發展方向

### 多模態融合

結合視覺、音頻和元數據特徵，提高偵測準確率

### 持續學習

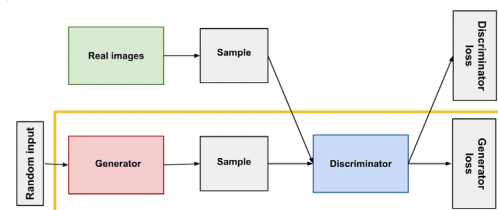
開發能夠不斷適應新型Deepfake技術的自更新模型

### 輕量級模型

開發適用於移動設備和瀏覽器的高效偵測模型

### 可解釋AI

提高模型透明度，解釋為何判定內容為Deepfake



# Diffusion Model生成內容偵測

## AI生成影像的特徵

### 像素級特徵

噪點分布不自然、顏色直方圖異常、壓縮痕跡不一致

### 物體結構特徵

不合理的物體結構、解剖學錯誤、物理規則違反（如手指數量異常）

### 紋理與細節特徵

過度平滑或過度細節化的區域、重複紋理、不自然的邊緣過渡

### 文字與標誌特徵

文字扭曲、不連貫或無意義的文字、標誌細節錯誤

## 主要偵測工具

### HmIA (How much is AI)

線上工具，提供AI生成可能性百分比評分

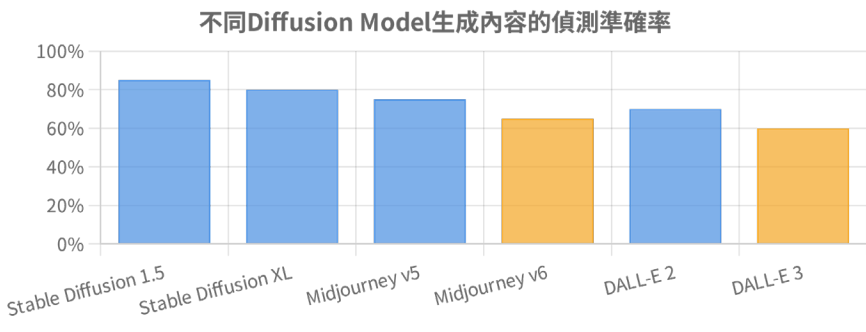
### AI or Not

簡單易用的在線檢測工具，適合普通用戶

### Illuminarty

開源偵測工具，專注於Diffusion Model生成內容

## 偵測準確率比較



Diffusion Model生成的影像比傳統GAN更難偵測，特別是高解析度和經過後期處理的影像。目前最佳偵測模型對Stable Diffusion生成內容的準確率約為 **75-85%**，但對於最新模型如DALL-E 3和Midjourney v6，準確率可能降至**60-70%**。

## 浮水印與內容認證

### 內容憑證 (C2PA)

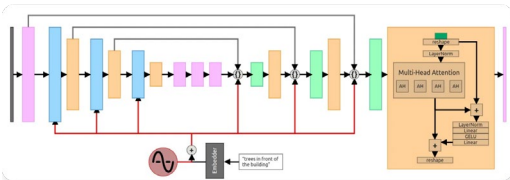
由Adobe、微軟等推動的內容來源認證標準

### 隱形浮水印

Stable Diffusion、DALL-E等模型可添加肉眼不可見的浮水印

### SynthID

Google開發的AI生成內容標識技術，可抵抗編輯和壓縮



# 台灣Deepfake案例：小玉事件

## 🕒 事件始末

**2021年2月**

YouTuber「小玉」(本名蕭宇辰)使用Deepfake技術將多位女性藝人臉部換到色情影片中

**2021年2月9日**

警方接獲檢舉，搜索小玉住處，查扣電腦設備與影片檔案

**2021年2月18日**

小玉被依妨害名譽罪嫌移送台北地檢署偵辦

**2022年1月**

台北地檢署依妨害名譽罪起訴小玉，求處有期徒刑3年

**2022年9月**

台北地方法院一審判決小玉有期徒刑3年10個月

## 🔪 法律處理結果

小玉案是台灣首宗以 **刑法妨害名譽罪** 起訴並判刑的Deepfake案件，具有指標性意義。

「被告明知使用AI技術製作之影片非真，卻仍以此方式製作多部影片，並上傳至色情網站，使不特定多數人均可觀看，嚴重侵害被害人之名譽權...」

— 台北地方法院判決書

法院認定小玉的行為已構成 **加重誹謗罪**，因為：

**散布於眾**：將影片上傳至公開網站

**意圖營利**：透過點閱率獲取廣告收益

**侵害多人**：影片涉及多位女性藝人

## ⚠️ 社會影響

### 👥 公眾意識提升

引發台灣社會對Deepfake技術潛在危害的廣泛討論，提高公眾警覺性

### ⚖️ 法律適用探討

促使法律界討論現有法律對新興科技濫用的適用性與不足之處

### ♀️ 性別議題關注

引發對數位性暴力、女性形象被惡意利用等議題的關注

## 💡 案例啟示

小玉事件為台灣社會帶來多方面的啟示：

### 法律保護的重要性

現有法律雖可適用於Deepfake案件，但仍需針對新技術制定更明確的規範

### 技術倫理的必要性

AI技術發展需要同步建立倫理準則與自律機制

### 媒體素養的重要性

公眾需具備辨識Deepfake內容的能力，不輕信網路影片

### 平台責任的重要性

社交媒體與內容平台需建立有效機制防止Deepfake內容傳播

# 台灣Deepfake詐騙案例

## 常見Deepfake詐騙手法

### 視訊詐騙

冒充親友或主管進行視訊通話，要求緊急轉帳

### 語音詐騙

模仿親友或權威人士聲音進行電話詐騙

### 名人投資詐騙

製作名人推薦投資產品的假影片

### 感情詐騙

使用虛假身份進行網路感情詐騙

## 台灣實際案例

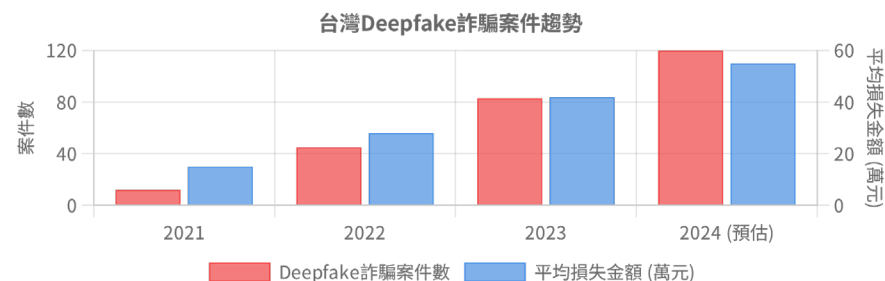
### 台北企業主管詐騙案 (2023)

詐騙集團冒充公司高層，透過視訊會議指示財務人員緊急轉帳

### 名人投資推薦詐騙 (2023)

製作台灣知名企業家推薦投資的Deepfake影片

## 詐騙趨勢分析



根據刑事警察局統計，2023年台灣AI相關詐騙案件較2022年增加約85%，其中Deepfake相關詐騙佔比從5%上升至18%。

## 防範措施

### ✓ 多重驗證

透過其他管道確認對方身份

### 🔒 設置安全問題

預先設置只有彼此知道的問題

### 🔄 隨機動作測試

要求對方做出特定隨機動作

### 🚫 拒絕緊急轉帳

建立多重確認和授權機制

# 台灣相關法律規範

## 🏛️ 現行適用法律

### 🔗 刑法第310條：妨害名譽罪

散布不實言論足以損害他人名譽者，處一年以下有期徒刑、拘役或五百元以下罰金

### 🔗 刑法第235條：散布猥褻物品罪

散布、播送或販賣猥褻之圖畫、聲音、影像或其他物品者，處二年以下有期徒刑、拘役或科或併科三萬元以下罰金

### 🔗 民法第18條、第184條、第195條

人格權受侵害時可請求法院除去侵害，並得請求損害賠償

### 🔗 著作權法相關條款

未經同意使用他人肖像製作Deepfake可能構成著作權侵害

## 📋 法律適用挑戰



現行法律在應對Deepfake和AI生成內容時面臨多重挑戰，包括**取證困難**、**跨境執法問題**、**技術發展速度快於立法**等。台灣目前尚無專門針對Deepfake的法律規範，主要依靠現有法律框架處理相關案件。

## 🌐 國際法規參考

### 美國

多州已藉由Deepfake專法，如加州要求標示AI生成內容

### 歐盟

《人工智能法》要求明確標示AI生成內容

### 中國

《深度合成管理規定》要求標示AI生成內容

### 韓國

修訂《資訊通信網法》，加重對Deepfake色情內容的處罰

國際上的法規趨勢主要集中在**內容標示**、**平台責任**、**加重處罰**三個方向，台灣未來立法可參考這些國際經驗。

## 💡 未來法律發展趨勢

### 🔗 內容標示義務

要求AI生成內容必須明確標示，包括影像、音頻和視頻

### 📋 平台責任強化

要求社交媒體和內容平台建立Deepfake內容的審核和舉報機制

### 🔗 專門立法保護

制定專門針對Deepfake和AI生成內容的法律，明確法律責任

### 🌐 國際合作機制

建立跨境執法合作機制，共同打擊Deepfake濫用行為

# 來點微學術- 主要內容

## AI深偽詐欺技術發展趨勢

- 為什麼這技術這麼容易被使用?

## 近一年台灣與國際案例

- 那麼這個深偽造成什麼影響?

## 深偽內容的偵測技術與應對

- 可以用AI來偵測AI嗎? 技術是什麼?  
限制又是什麼?

## 與查核單位合作案例

- 共同挑戰真實世界問題，而不僅僅  
只是學術論文

## DeepFake: Synthesized Image/Video



# DeepFake: Easy to Use...

## DeepFaceLab

deepfake tutorial  
using whole\_face + XSeg

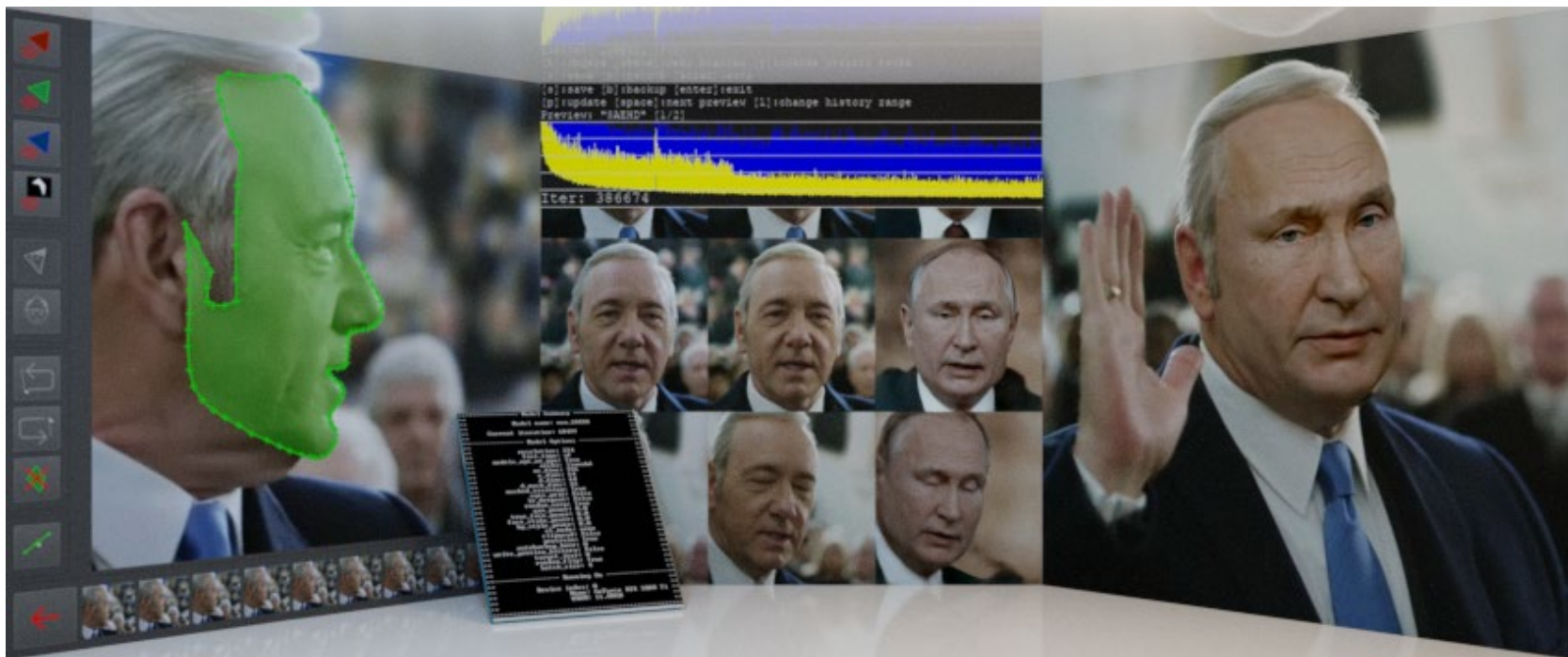
1. ~~multiple versions~~
2. extract images from src video
3. extract images from dst video
4. data\_src : extract faces
5. data\_src : view extracted faces
6. data\_dst : extract faces
7. data\_dst : view extracted faces
8. XSeg : mask few data\_dst faces manually
9. XSeg : mask few data\_src faces manually
10. XSeg : train model
11. XSeg : apply trained mask for data\_dst
12. XSeg : apply trained mask for data\_src
13. XSeg : check data\_src mask
14. XSeg : check data\_dst mask
15. XSeg : train more after mask correction
16. XSeg : apply trained mask again
17. Train SAEHD model
18. Interactive Merger
19. Build and view final video



data\_src.mp4

# DeepFake 最大宗的作法

- DeepFaceLab
  - <https://github.com/iperov/DeepFaceLab>
- 幾乎有程式能力基礎、有一點機器學習概念者皆有機會製作自己的DeepFake



好，就算不會用，隨便搜尋一下也很多...

Google 換臉工具

全部 圖片 影片 新聞 購物 地圖 網頁 更多 工具 已儲存

Swap Ai 技術 人工智能 Roop Picwish Ai 生成 Facefusion 臉神器 Ai 智能 Youtube Ai 圖片 臉替換

**AI 換臉 5 款神器分享**  
iMyFone AI 換臉神器推薦！5 款工具讓你一鍵變...

**FaceSwapper 免費 AI 換臉工具**  
電腦王阿達 FaceSwapper 免費 AI 換臉工具...

**AI Face Swap 線上 AI 換臉工具**  
三嚨行動哇 免費 AI Face Swap 線上 AI 換臉工具...

**如何解除限制？**  
YouTube 3 款常用的免費換臉工具，解除內容審...

**PicWish AI 換臉功能重磅上線**  
PicWish AI 換臉功能重磅上線，可...

**6 款 AI 換臉 App、...**  
訊連科技 6 款 AI 換臉 App、...

**AI 換臉工具 Rope...**  
AI 資源網 AI 換臉工具 Rope...

**最受歡迎的自拍編輯器**  
iFunnels-AI 數位行銷生態圈 什麼是 AI 換臉？4 個免費 AI 換臉工具推薦、如...

**2024 年你不應該錯過的 8 個免費線上換臉...**  
Mango Animate 2024 年你不應該錯過的 8 個免費線上換臉...

**FaceSwapper 免費 AI 換臉線上工具...**  
電腦王阿達 FaceSwapper 免費 AI 換臉線上工具...

**最強換臉工具 FACEFUSION VIP 版發布！**  
YouTube 最強換臉工具 FACEFUSION VIP 版發布！...

**免費 AI 影片製作工具 - Vidqu AI**  
適用於照片和影片的免費線... 免費 AI 影片製作工具 - Vidqu AI 適用於照片和影片的免費線...

**Reface - AI 智能在綫換臉...**  
AI 旋風 Reface - AI 智能在綫換臉...

**AI 換臉，想換誰就換誰**  
AI 資源網 精彩 AI 換臉秀：功能強大的 AI 換臉工具，支持...

**Face Swapper：一款由 Icons8 開發的基於...**  
ai, 人工智能, 1ai.net Face Swapper：一款由 Icons8 開發的基於...

**6 款 AI 換臉 App、免費線上 AI 換臉推薦，照片換臉好輕...**  
訊連科技 6 款 AI 換臉 App、免費線上 AI 換臉推薦，照片換臉好輕...

**科技新知 AI 換臉網站怎麼用？Face Swa...**  
傑昇通信 科技新知 AI 換臉網站怎麼用？Face Swa...

**AI 換臉工具推薦**  
iFunnels-AI 數位行銷生態圈 什麼是 AI 換臉？4 個免費 AI 換臉工具推薦、...

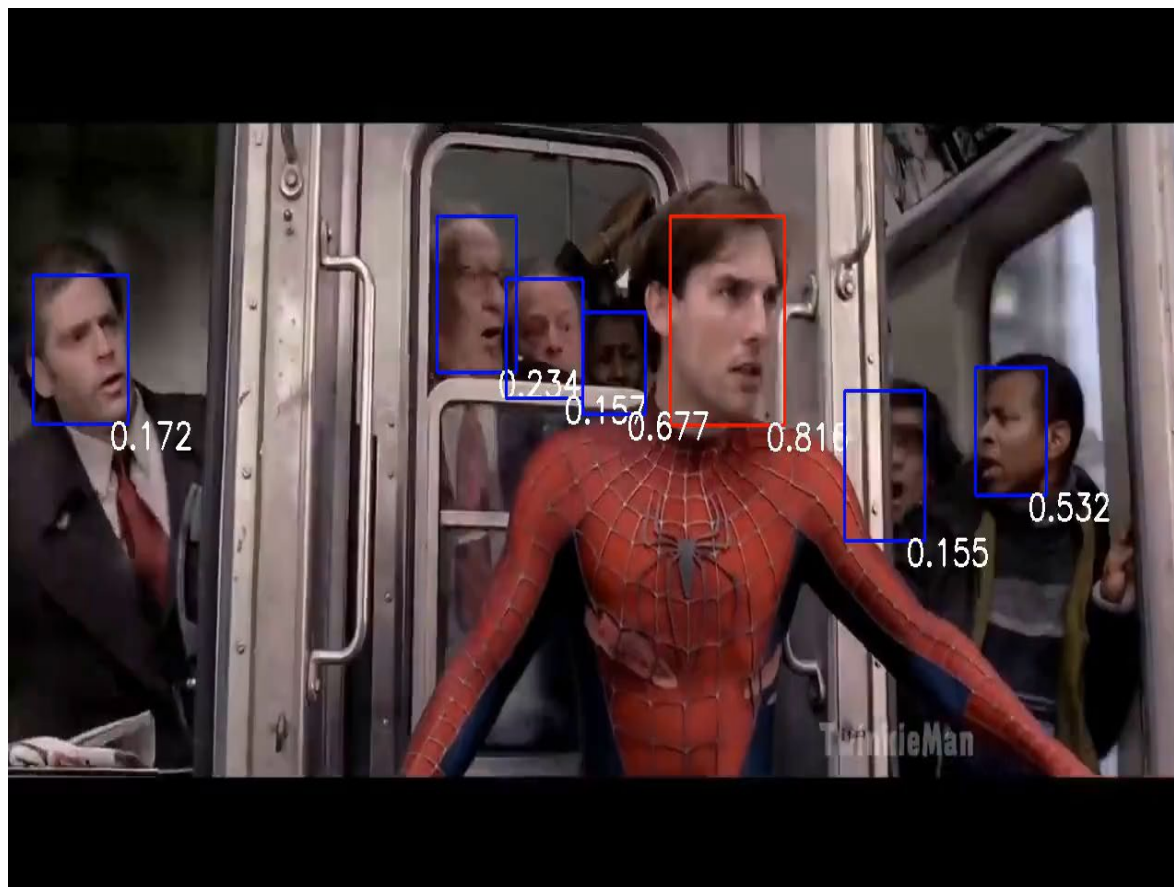


De-  
DeepFake!

Deepfake  
detection

---

## 那麼在防禦端呢? (DeepFake Detection)



X. Zhuang and C.C. Hsu, "Detecting generated image based on coupled network with two-step pairwise learning," *IEEE International Conference on Image Processing (ICIP, Top Conference on Multimedia)*, Taipei, Taiwan, 2019. (Best Student Award)

# GENERATIVE AI

GAN (生成對抗網絡)

Diffusion-based Models (擴散模型)

這些技術已經很成熟...



# Deepfake技術原理

## ⚙️ 自編碼器 ( Autoencoder )

自編碼器是Deepfake的核心技術，包含**編碼器**和**解碼器**兩部分：

- 1 **編碼器**：將人臉影像壓縮為低維度的特徵向量
- 2 **解碼器**：將特徵向量重建為人臉影像
- 3 **關鍵技巧**：訓練兩個人的編碼器共享，但解碼器分開

這使得系統能夠將一個人的臉部表情特徵應用到另一個人的臉上，同時保留目標人物的身份特徵。

## 🔗 換臉技術的工作流程



收集源人臉和目標人臉的大量影像

檢測並對齊人臉特徵點

訓練自編碼器或GAN模型

生成換臉結果並優化細節

## ⚖️ 生成對抗網路 ( GAN )

GAN由**生成器**和**判別器**

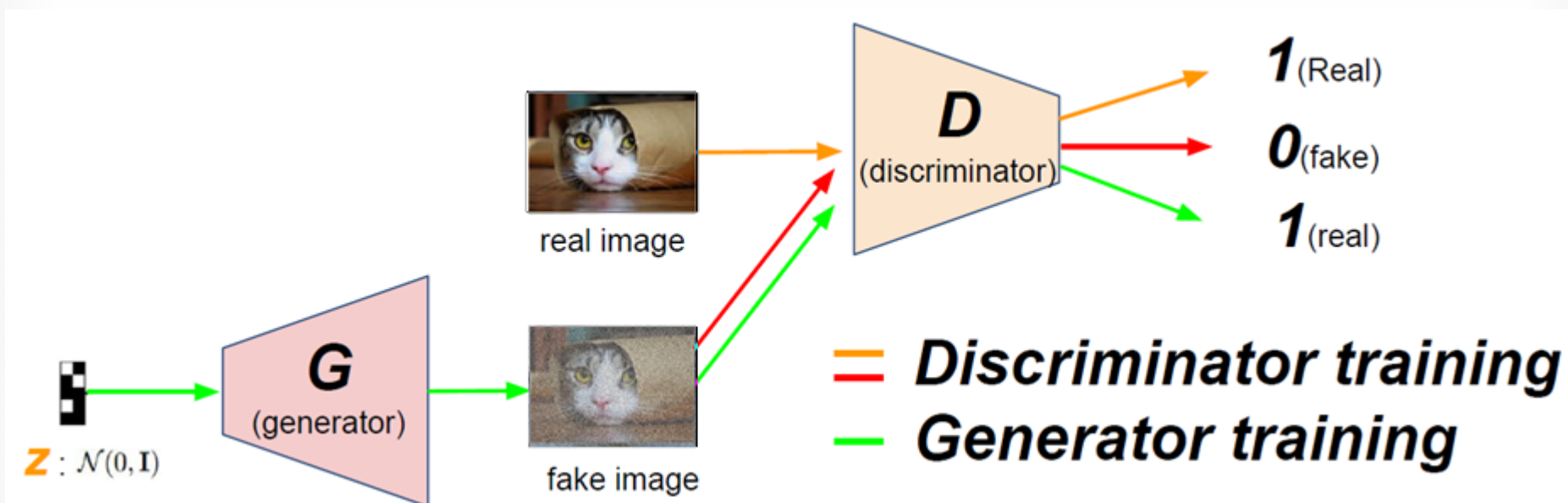
兩個網路組成，藉由對抗訓練提高合成品質：

- 1 **生成器**：負責生成逼真的假臉部影像
- 2 **判別器**：負責區分真實與生成的影像
- 3 **對抗過程**：兩個網路不斷競爭，互相提升

藉由這種方式，生成器能夠產生越來越逼真的換臉結果，最終達到以假亂真的效果。

## 歸功於生成式模型 Generative Model

- 如何生成高品質的影像？
  - 生成對抗網絡 ( GAN )



Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems*. 2014.



## Many GANs

- 對**GAN** (生成對抗網絡) 本質的深入探討
  - WGAN、WGAN-GP、RAGAN、BigGAN、StyleGANv2、Anycost GAN
  - ...etc
- 應用
  - 基於最先進的GAN模型，進行fine-tune.
- 重點在這些**GAN**有夠多...
- 現在還有**Diffusion**模型?!



# 甚至還有 Stable Diffusion Models (Midjourney & DALL.E)

---

# GAI的用途

這麼說來GAI很難預防被  
用來製作DeepFake了?

NO! 凡事一體兩面

- 每個東西都有它存在的理由，所以GAI也是。
- 那他到底可做什麼?

# 影像還原技術

bicubic  
(21.59dB/0.6423)



SRResNet  
(23.53dB/0.7832)



SRGAN  
(21.15dB/0.6868)



original



Ledig et al. CVPR 2017

# Label2Image



# HRGAN [Wang et al., CVPR2018]



(a) pix2pix



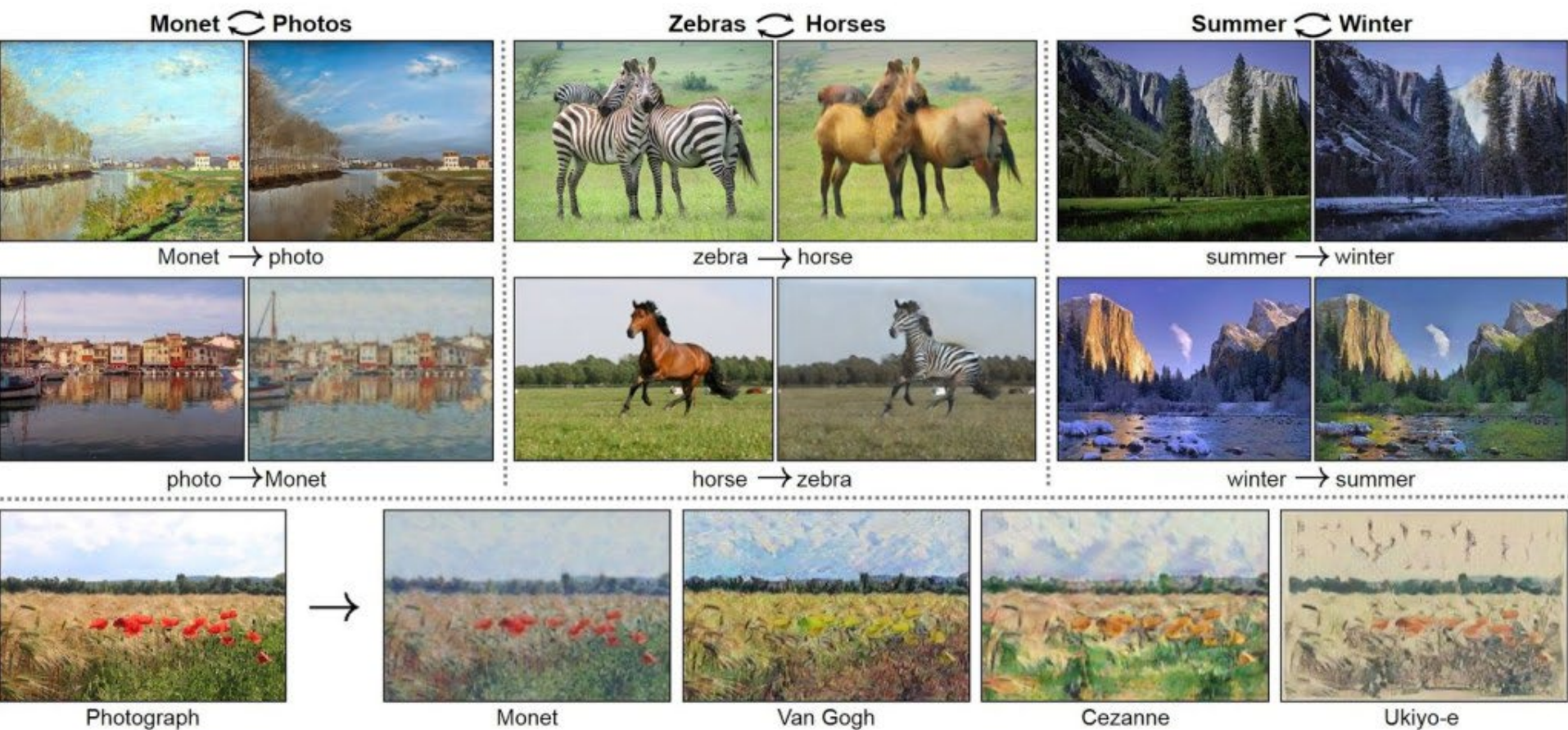
(b) CRN



(c) Ours (w/o VGG loss)



(d) Ours (w/ VGG loss)

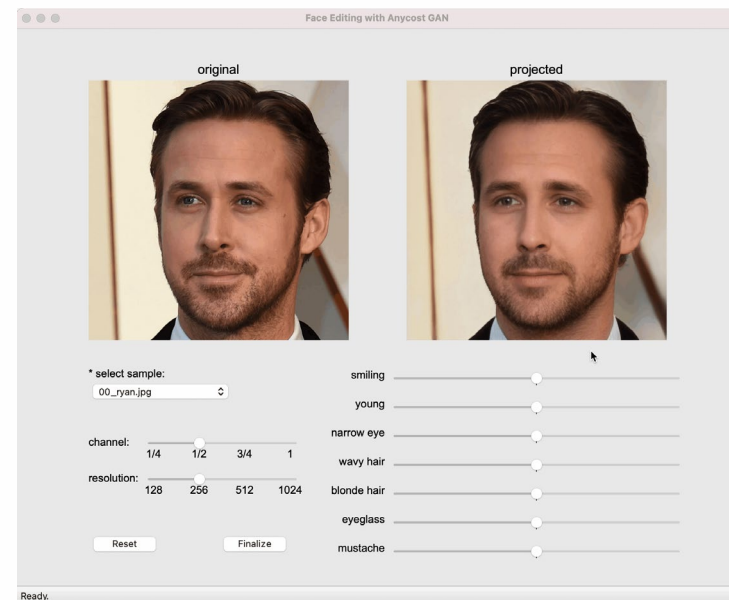


# Face Synthesis

- StyleGANv2 [CVPR2020]



- Anycost GAN [CVPR2021]



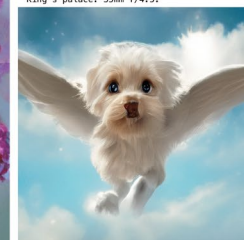
# GigaGAN: 瞬間產生高解析度圖片 [CVPR2023]



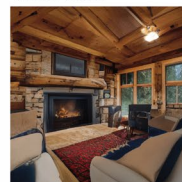
A portrait of a human growing colorful flowers from her hair. Hyperrealistic oil painting. Intricate details.



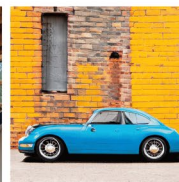
A golden luxury motorcycle parked at the King's palace. 35mm f/4.5.



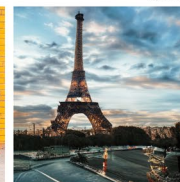
a cute magical flying mallepoop at light speed, fantasy concept art, bokeh, wide sky



A living room with a fireplace at a wood cabin. Interior design.



a blue Porsche 356 parked in front of a yellow brick wall.



Eiffel Tower, landscape photography



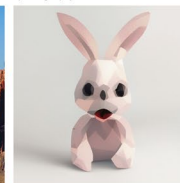
A painting of a majestic royal tall ship in Age of Discovery.



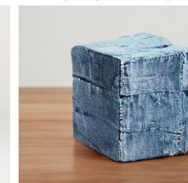
Isometric underwater Atlantis city with a Greek temple in a bubble.



A hot air balloon in shape of a heart. Grand Canyon



low poly bunny with cute eyes



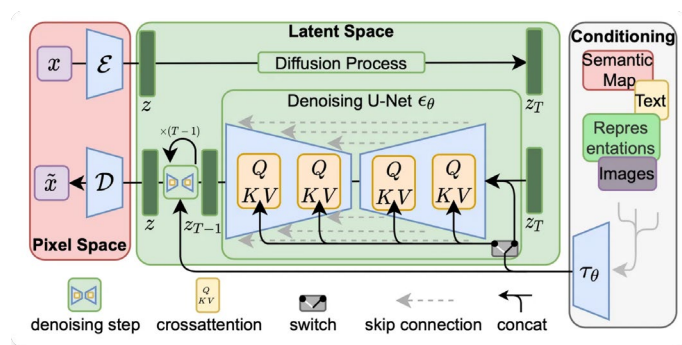
A cube made of denim on a wooden table

# Diffusion Model簡介

## 什麼是Diffusion Model？

**擴散模型** ( Diffusion Model ) 是一種基於深度學習的生成模型，其核心思想是透過**逐步添加噪聲**再**逐步去除噪聲**的過程來生成新的數據。

擴散模型的靈感來自於**熱力學中的非平衡過程**，模擬粒子從有序狀態到無序狀態，再從無序狀態恢復到有序狀態的過程。



## Diffusion Model的主要特點

- 高品質的影像生成能力，細節豐富且真實
- 強大的文本條件控制，可根據文字描述生成影像
- 多樣化的風格表現，可生成各種藝術風格的影像
- 訓練穩定，不易出現模式崩潰等問題

## 與其他生成模型的差異

| 模型類型            | 優點                   | 缺點                  | 代表模型                    |
|-----------------|----------------------|---------------------|-------------------------|
| Diffusion Model | 高品質生成結果、訓練穩定、強大的條件控制 | 生成速度較慢、計算資源需求高      | Stable Diffusion、DALL-E |
| GAN             | 生成速度快、資源需求相對較低       | 訓練不穩定、模式崩潰問題、細節表現較差 | StyleGAN、CycleGAN       |
| VAE             | 潛在空間連續、訓練穩定          | 生成結果較模糊、細節不足        | $\beta$ -VAE、VQ-VAE     |

# Diffusion Model原理

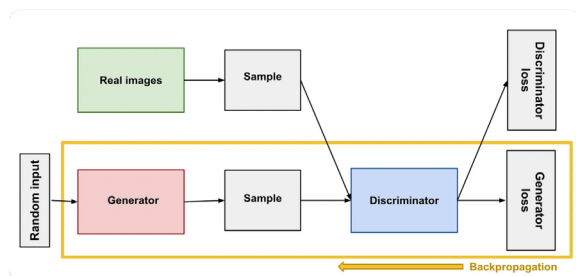
## → 前向擴散過程

前向擴散過程是一個 **逐步添加噪聲** 的過程，將原始數據轉換為純噪聲：

從原始影像  $x_0$  開始，逐步添加高斯噪聲

經過  $T$  步後，影像變為近似純噪聲  $x_t$

$$q(x_t|x_0) = N(x_t; \sqrt{\alpha_t}x_0, (1-\alpha_t)I)$$



## </> 訓練過程

從訓練數據中採樣真實影像  $x_0$

隨機採樣時間步  $t$  並添加相應噪聲得到  $x_t$

訓練神經網路  $\epsilon_\theta$  預測添加的噪聲

最小化預測噪聲與實際噪聲的均方誤差

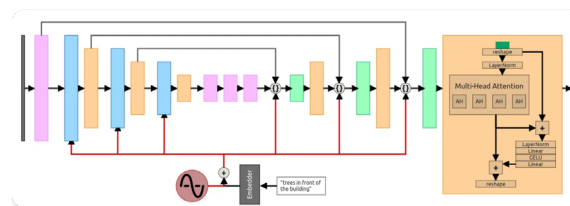
## ← 反向擴散過程

反向擴散過程是一個 **逐步去除噪聲** 的過程，從純噪聲生成有意義的數據：

從純噪聲  $x_t$  開始，逐步預測並去除噪聲

使用神經網路  $\epsilon_\theta$  預測添加的噪聲

$$p_\theta(x_0|x_t) = N(x_0; \mu_\theta(x_t, t), \sigma^2 I)$$



## 生成過程

從標準高斯分布採樣純噪聲  $x_t$

使用訓練好的神經網路逐步去除噪聲

每一步使用 DDIM 或 DDPM 採樣算法

最終得到生成的影像  $x_0$

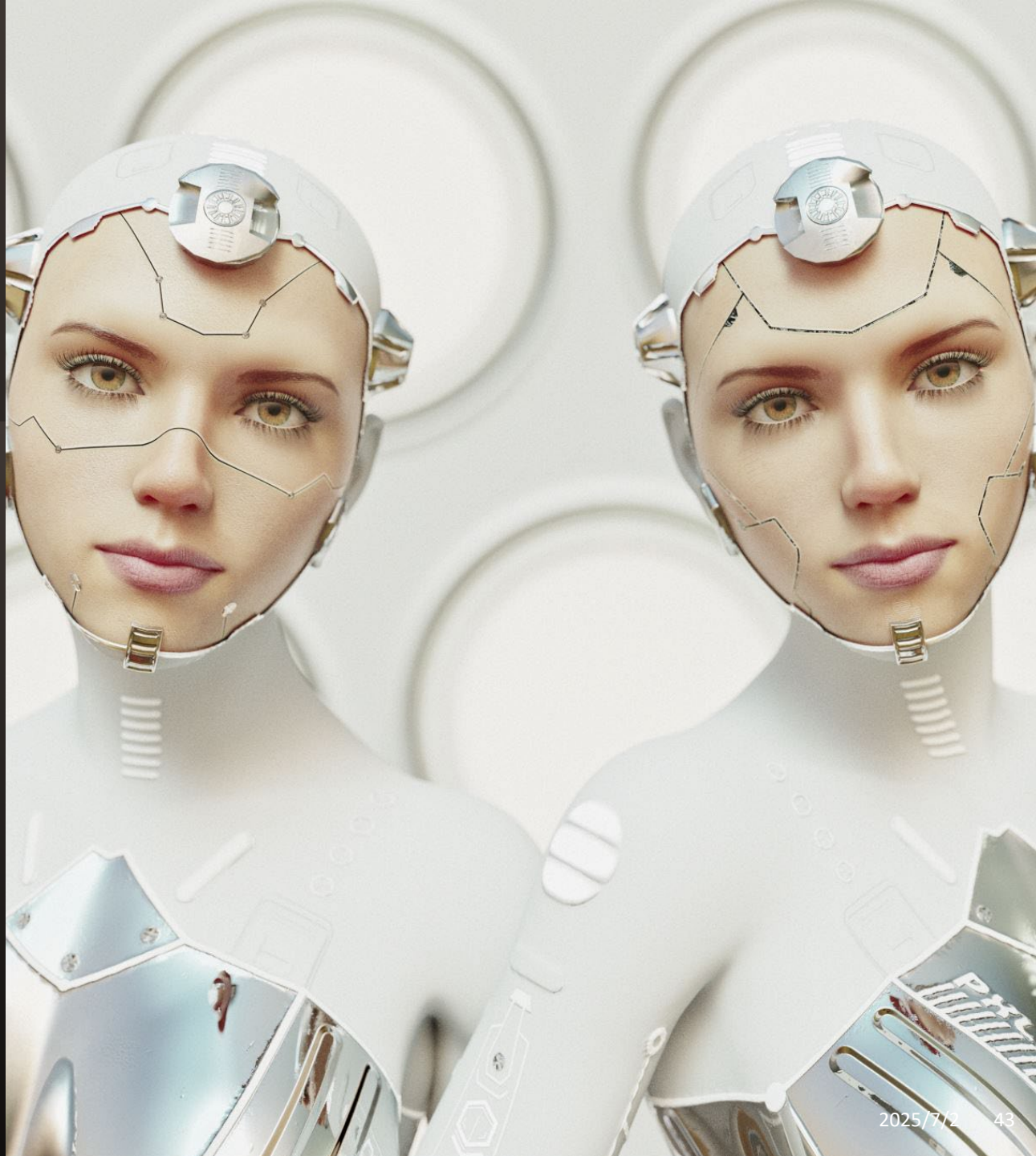
# Stable video diffusion [2024, arXiv]



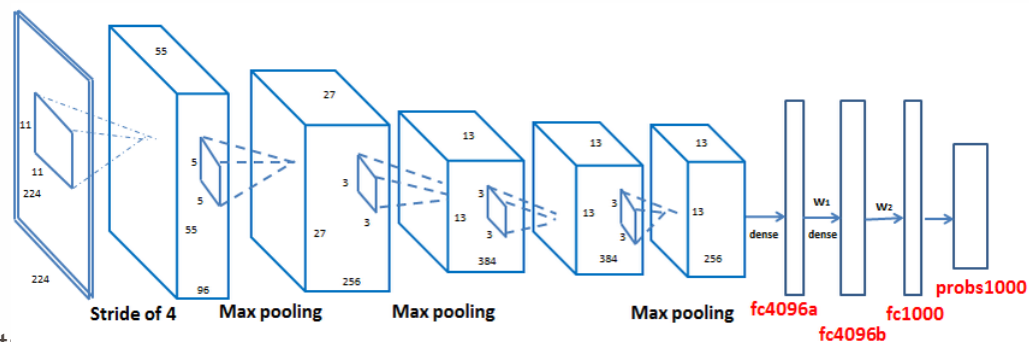
既然如此，以  
毒攻毒

運用AI對抗  
AI?

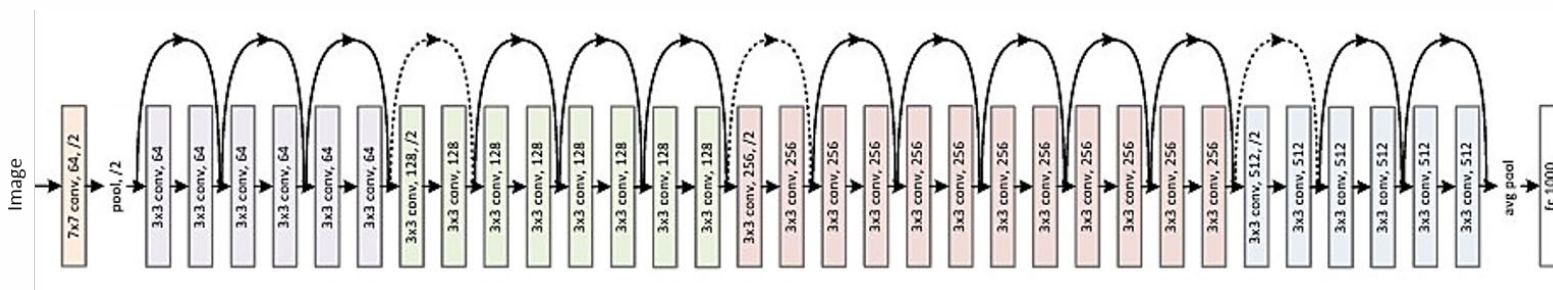
以影像分類為  
例



## Deeper Network



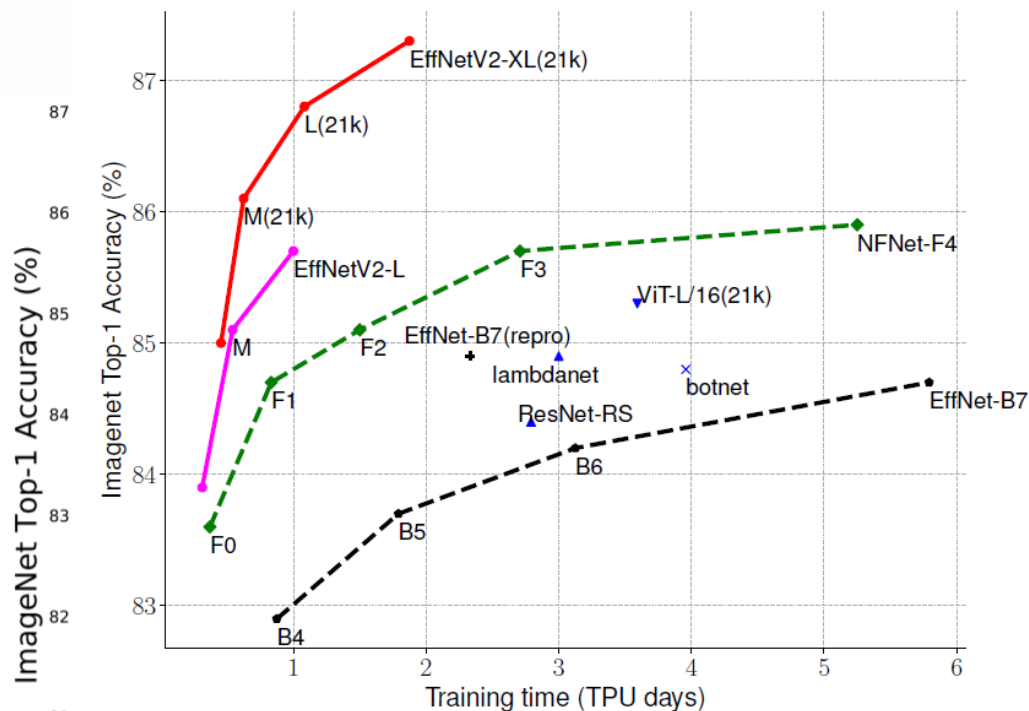
- 2013, AlexNet: 卷积神经网络
- 2016, Residual Net / DenseNet: 加深至152 層 ...
- 2017, Stochastic depth Net: 加深至1000 層 ...
- 2020, NAS (EfficientNet, NFN...)
- 2021, Transformers...attention...
- 2023-, Foundation models, LLM, MLM, AI Agent...



# SOTA for Image Recognition (2023)

- Meta pseudo label
- 自監督學習
- [Google 2021]
- DeepMine proposed NFNet
- 無需normalization!!
  - Normalizer-Free ResNet
- AGC (梯度剪裁) [2021]
- **EfficientNet v2**
  - [google 2021.4]
- **Swin-Transformerv2**
- **ConvNextV2**

Tan, Mingxing, and Quoc V. Le. "Efficientnetv2: Smaller models and faster training." *arXiv preprint arXiv:2104.00298* (2021).



(a) Training efficiency.

|            | EfficientNet<br>(2019) | ResNet-RS<br>(2021) | DeiT/ViT<br>(2021) | EfficientNetV2<br>(ours) |
|------------|------------------------|---------------------|--------------------|--------------------------|
| Top-1 Acc. | 84.3%                  | 84.0%               | 83.1%              | 83.9%                    |
| Parameters | 43M                    | 164M                | 86M                | 24M                      |

(b) Parameter efficiency.

# Good enough? Our preliminary study shows...

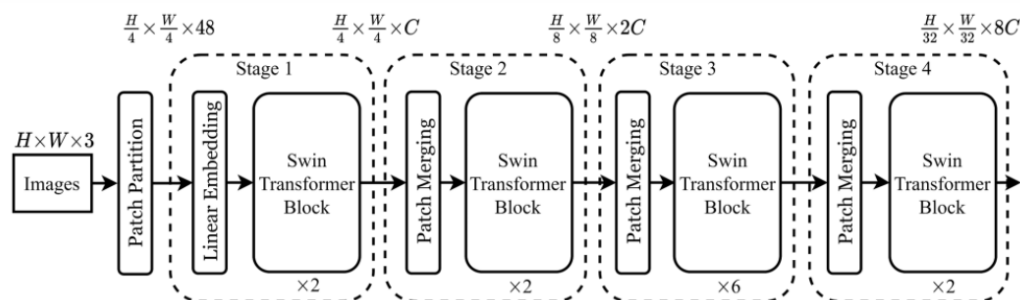
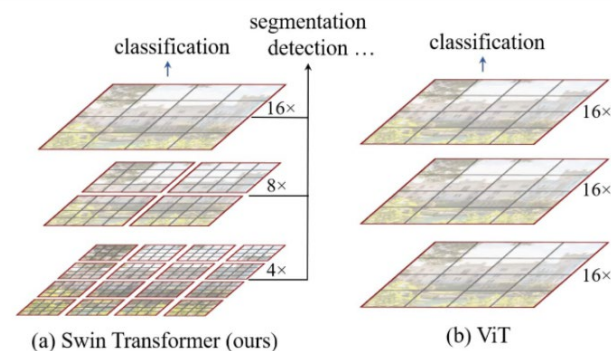
資料集: Faceforensics++ [1]

面部偵測: BlazeFace [2]

骨幹網路: Swin Transformer[4]

損失函數: Binary Cross Entropy

優化器: Adam

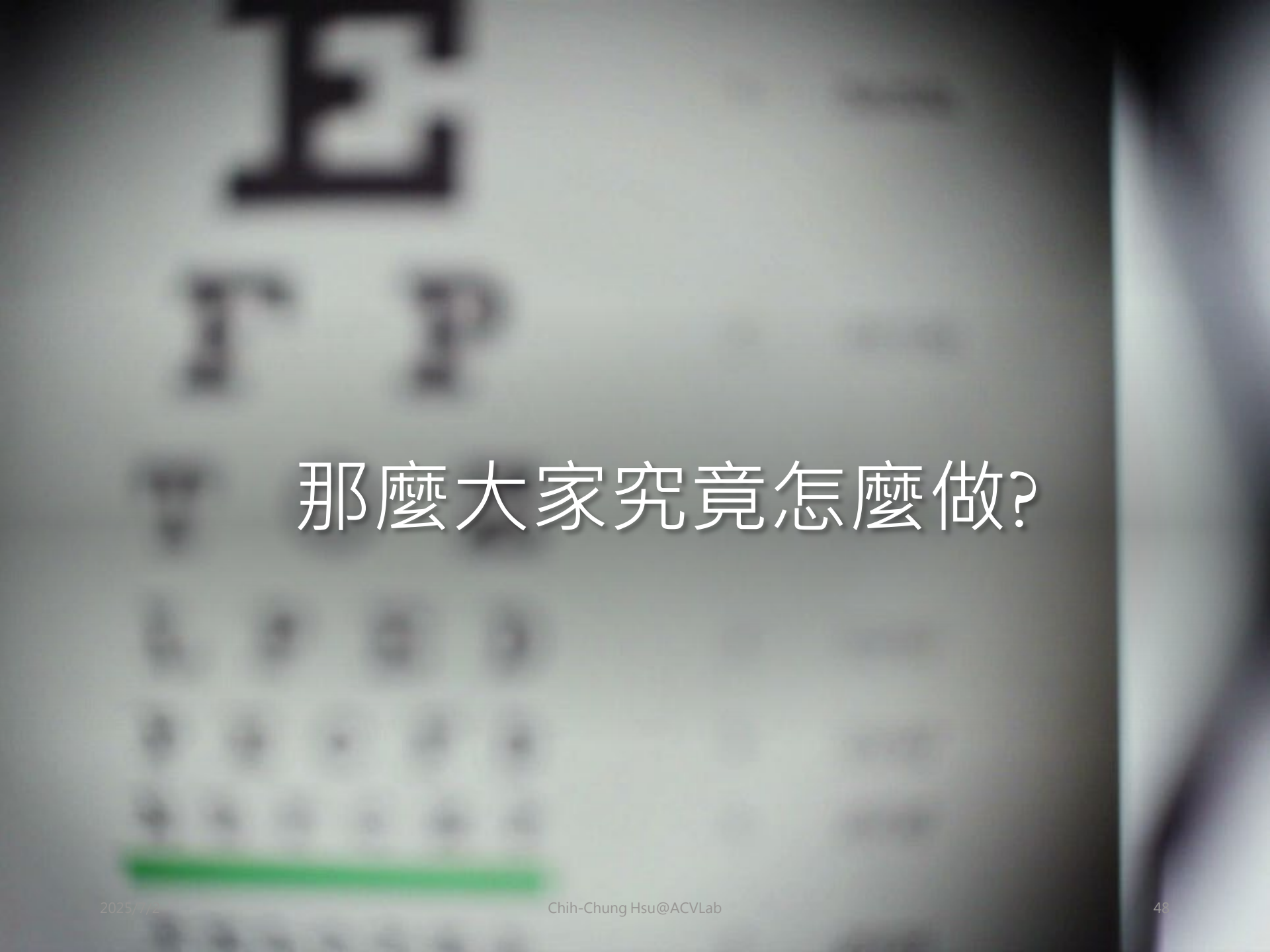


## 一個簡單的實驗來驗證

- 對Faceforensics++進行性能評估 [1]
  - 所有類別

| CR (training/testing sets) | RAW          | CR=23 | CR=40 |
|----------------------------|--------------|-------|-------|
| RAW                        | <b>0.987</b> | 0.936 | 0.772 |
| CR=23                      | 0.701        | 0.939 | 0.757 |
| CR=40                      | 0.576        | 0.648 | 0.735 |

| Method                       | Info | Deepfakes | Face2Face | FaceSwap | NeuralTextures | Pristine | Total        |
|------------------------------|------|-----------|-----------|----------|----------------|----------|--------------|
|                              |      | ▼         | ▼         | ▼        | ▼              | ▼        | ▼            |
| <a href="#">ZAntiFakeBio</a> |      | 1.000     | 0.920     | 0.971    | 0.907          | 0.936    | <b>0.940</b> |
| <a href="#">Leo</a>          |      | 1.000     | 0.861     | 0.971    | 0.853          | 0.922    | 0.917        |
| <a href="#">NoSenseAtAll</a> |      | 0.982     | 0.905     | 0.951    | 0.827          | 0.908    | 0.908        |
| <a href="#">Cancer</a>       |      | 0.964     | 0.781     | 0.942    | 0.780          | 0.952    | 0.903        |



那麼大家究竟怎麼做?

# Deepfake偵測方法概述

## 偵測技術分類

### 視覺特徵偵測

分析臉部不自然特徵、眨眼頻率、光影不一致、邊緣模糊等視覺線索

### 深度學習偵測

使用CNN、RNN等深度學習模型從大量真假樣本中學習區分特徵

### 生物特徵偵測

分析心跳信號、血流變化、眼球反射等生物特徵的一致性

### 元數據分析

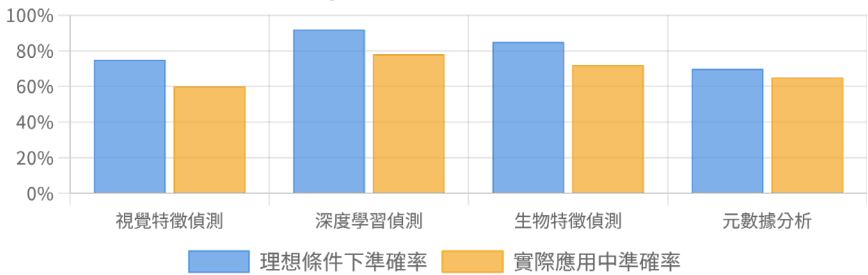
檢查文件元數據、壓縮痕跡、噪聲模式等技術指紋

## 偵測的挑戰

- 技術軍備競賽**：生成技術與偵測技術不斷進化，形成軍備競賽
- 壓縮與重新編碼**：社交媒體壓縮會破壞偵測線索
- 多樣性與泛化**：難以應對各種不同類型的Deepfake技術
- 對抗性攻擊**：專門設計來逃避偵測的Deepfake技術

## 偵測效能與準確率

各類Deepfake偵測方法的準確率比較



隨著Deepfake技術的進步，偵測難度也在增加。目前最佳的偵測模型在理想條件下可達**90%以上**的準確率，但在複雜場景中準確率可能降至**70%左右**。

## 偵測工具與平台

- |                                                                   |                                         |
|-------------------------------------------------------------------|-----------------------------------------|
| <b>Microsoft Video Authenticator</b><br>微軟開發的視頻真實性分析工具            | <b>Sensity</b><br>專業的Deepfake偵測平台       |
| <b>Deepware Scanner</b><br>開放使用的在線偵測工具                            | <b>FaceForensics++</b><br>學術研究用的偵測基準數據集 |
| 目前偵測工具主要應用於 <b>新聞媒體、社交平台、執法機構</b> 和 <b>網絡安全</b> 領域，但普通用戶可用的工具仍然有限 |                                         |

# Deepfake視覺特徵偵測

## 常見視覺偵測特徵



### 眨眼頻率異常

早期Deepfake模型難以模擬自然的眨眼動作，眨眼頻率過低或不自然

85%



### 臉部邊緣模糊或不自然

臉部與背景的融合處常有明顯的模糊、顏色不一致或硬邊緣

80%



### 光影不一致

臉部與場景的光源方向、陰影、高光等不一致

75%

## 動態特徵分析

### 臉部運動不協調

表情變化與說話內容不匹配

### 嘴唇同步問題

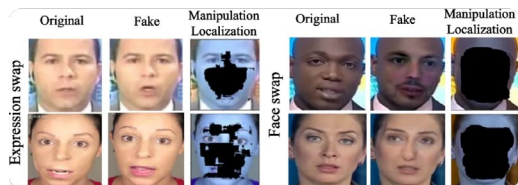
口型與聲音不同步或不自然

### 頭部姿態異常

頭部轉動時面部特徵變形

### 面部肌肉不自然

表情變化時肌肉運動不協調



## 更多視覺線索



### 牙齒與舌頭細節

牙齒、舌頭等口腔細節常顯得模糊或不自然

70%



### 眼鏡與配飾處理

眼鏡、耳環等配飾與臉部的交界處常有扭曲或變形

65%



### 頭髮邊緣處理

頭髮與背景交界處常缺乏自然的細節和過渡

60%

## 視覺偵測的局限性

隨著Deepfake技術的進步，純視覺偵測方法面臨越來越多挑戰：

**技術進步**：新一代Deepfake已能很好地處理眨眼、光影等早期問題

**高解析度**：高解析度Deepfake能呈現更多細節，減少視覺線索

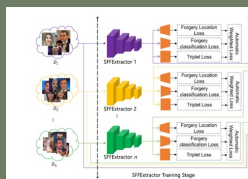
**後期處理**：專業後期處理可以修復大多數視覺瑕疵

**社交媒體壓縮**：影片壓縮會掩蓋許多視覺偵測線索

因此，現代Deepfake偵測通常需要**結合多種方法**，而非僅依賴視覺特徵。

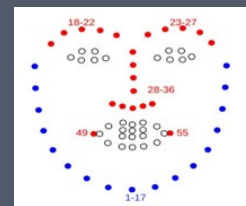
## 深度造假偵測的常規分類任務

- 影片、影像分類問題
- 混合方法 (影像+影片)



## 強健 (Robust) 的深度造假偵測：

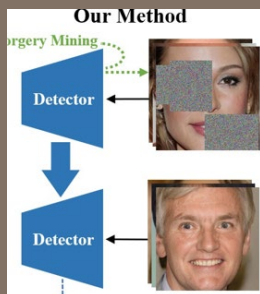
- 雜訊 (Noise) 壓縮 (Compression)
- 對抗攻擊 (Adversarial attack)



學術上  
我們怎麼做?

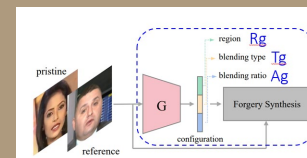
## 多模態的深度造假偵測

- 影像與語音的結合
- 其他 (例如生理訊號)



## 通用 (Generalized) 的深偽偵測

- 跨資料集 (Cross-dataset)
- 真實世界的場景



# Common Dataset

| 資料集                 | 真實影像   | 造假影像    | 年份   | Note                      |
|---------------------|--------|---------|------|---------------------------|
| UADFV               | 49     | 49      | 2018 | 專注在頭部姿勢                   |
| EBV                 | -      | 49      | 2018 | 注重於眼睛眨動                   |
| Deepfake-TIMIT      | 320    | 640     | 2018 | 基於 G A N                  |
| DFFD                | 1,000  | 3000    | 2019 | 多個 SOTA 方法                |
| DeepfakeDetection   | 363    | 3,608   | 2019 | 用公開的生成方法從行為者收集            |
| Celeb-DF (v2)       | 590    | 5639    | 2019 | 相對品質較好的偽造技術               |
| DFDC                | 23,564 | 104,500 | 2019 | Kaggle上的DFDC競賽（超大規模，真實場域） |
| FaceForensic++      | 1,000  | 5,000   | 2019 | 五種不同的生成方法，最主流             |
| FFIW-10K            | 10,000 | 10,000  | 2019 | 一幀多面                      |
| WLDR                | -      | -       | 2019 | Youtube上的相關人士影片           |
| DeeperForensics-1.0 | 50,000 | 10,000  | 2020 | 增加現實世界的擾動（攻擊等）            |
| Wild-Deepfake       | 3,805  | 3,509   | 2021 | 從網路上收集的真實偽造資料             |
| ForgeryNet          | 99,630 | 121,617 | 2021 | 8種影片偽造生成方法，添加對抗攻擊等擾動因素    |

- 應用面：no code!!
  - 給我應用程式，其餘免談!! >> 自己寫/跑code，對普羅大眾來說不需要
- 技術層面：現在學術上的最好方法 (State-of-the-art, SOTA)，足夠因應現實挑戰了嗎？
  - 未必，不然大家為什麼在這裡？



未知的作者的 [此相片](#) 已透過 [CC BY-SA-NC](#) 授權

# 真實場域：以既有的線上工具分析

## ▪ DeepFake 的偵測

### ▪ Sensity Visual Intelligence Platform

### ▪ Deepware

- 最易於使用，免費，效能一般

### ▪ Reality Defender

- 要申請 (還沒成功)

### ▪ Duck Duck Goose

### ▪ Truepic

- 針對AI修圖，不見得是 DeepFake

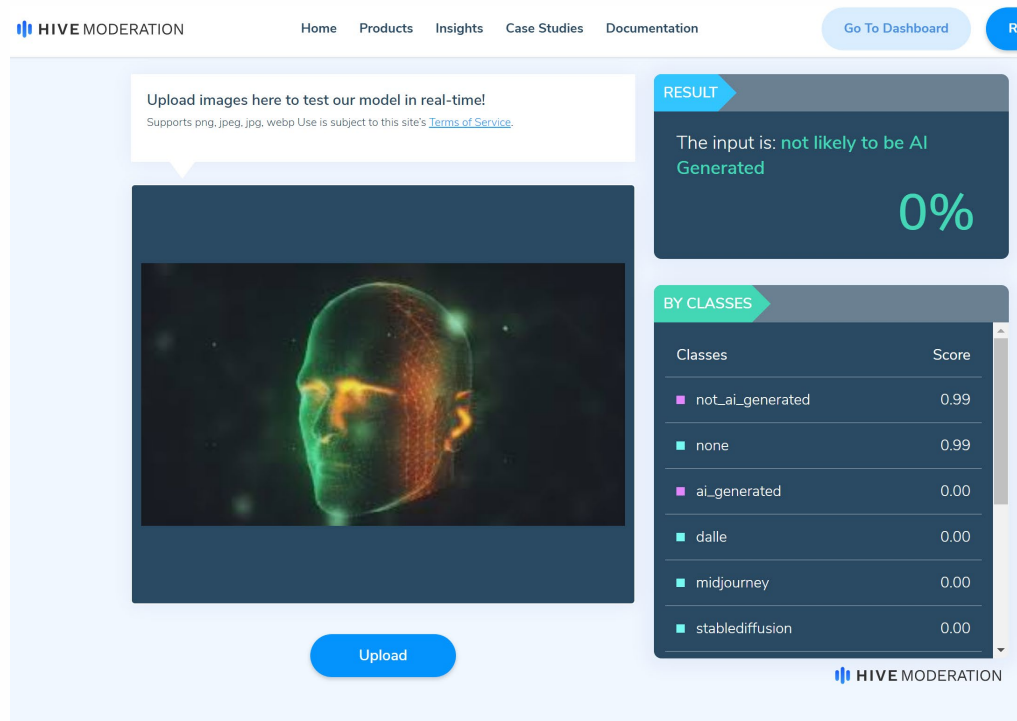
### ▪ WeVerify (Unavailable?)

### ▪ DeepFake-Detect (影像)

## ▪ 主流AIGC的技術 (不見得是換臉)

### ▪ Hive Moderation

- 大約包含40 種不同技術的偵測器
- 大多數主流的基本技術都可以偵測



# 研究與真實場域的差異



# 對抗真實場域下的深度偽造

能不能只鎖定特定人物來偵測就好?

- 基於身分特徵不一致性的深度偽造偵測方法

多模態的分析方法

- 是不是可以結合社群媒體資料來輔助判斷?

要怎麼樣可以增加偵測器的通用能力?

- 不用學習新的偽造影像，就能有一定辨識的能力

我們如何能夠始終擁有大量的篡改影像資料?

- 小樣本學習+ DeepFakde 偵測，快速適應新的偽造手法

偵測器遇上壓縮!!

- 結合rPPG訊號來強化影像受到各種不同壓縮時能保有穩定度的研究

不穩定的人臉偵測器？不穩定的人臉序列

- 即使因為種種原因沒有正確抓出所有人臉，仍可以穩定判斷是否偽造之研究

聯合攻擊？對抗攻擊 + 人臉操作

- 全球首創不受對抗攻擊影響的深度偽造偵測方法

# 真實場域下的偵測 I: 特定人物的偵測

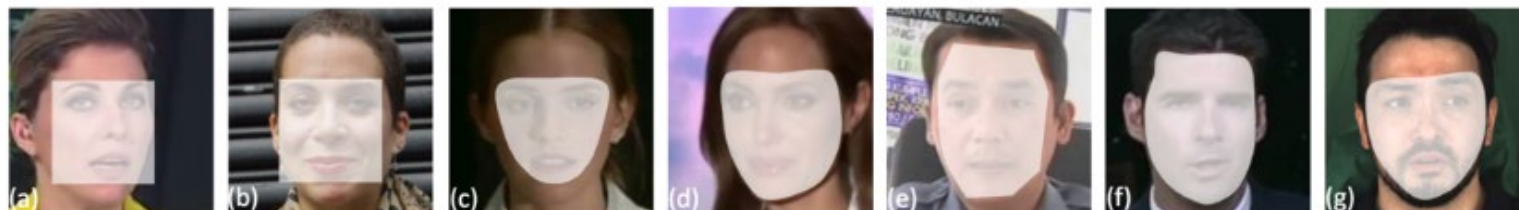
過往的方法都沒有討論到如何針對特定身分的偽造偵測

## 應用場域

- 所有論文偵測模型的效能是基於公開的 Dataset
- 效能可能優異，但是，是否能針對**特定重要人物**加強效能？



## 偽造過程的 Mask



圖片來源: [https://github.com/LightDXY/ICT\\_DeepFake](https://github.com/LightDXY/ICT_DeepFake)

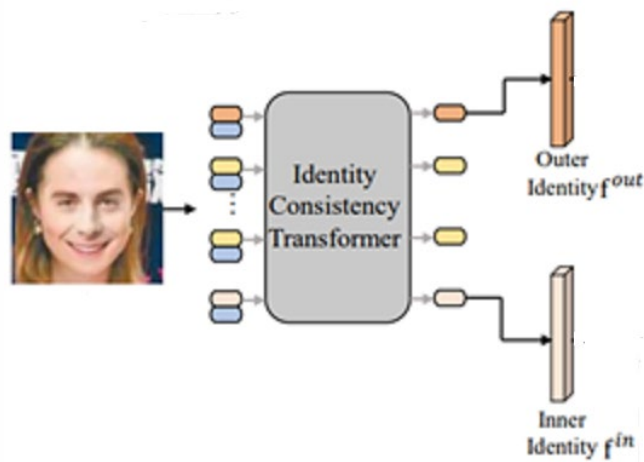
- 常見的DeepFake一般有一個mask(上圖白色的地方)，然後在mask的地方，進行換臉
- 為了之後說明方便，我們定義：

內臉:白色 mask 的地方

外臉:其他的地方

# 針對身分加強的方法：引入參考資訊

- 此方法主要是比對內臉跟外臉的身分一致性
  - 優點：不須用深偽影視訊的資料來進行訓練，比較彈性
  - 缺點：沒有將身分的特徵綁定；
- 引入人臉驗證 (Face verification) 的概念
  - 只針對”對的人”進行辨識
    - 用**真實照片**計算重要人物的內臉外臉的特徵，僅針對被便視為重要人物的影視訊進行剖析!!，
- 引入聲音**參考**資訊的辨識當作輔助資訊!



## 成果展示(偵測結果範例)

### result:

檢視的張數(每秒取2張) : 91張

**Method1**(取最相似的對照照片判斷) :

**chen** : 檢測到12張 , 人臉辨識信心度為0.14 , **chen**的造假影片機率為1.0

**la** : 檢測到91張 , 人臉辨識信心度為1.0 , **la**的造假影片機率為1.0

**su** : 未檢測到su的人臉

**tsai** : 未檢測到tsai的人臉

**Method2**(所有對照照片平均) :

**chen** : 比對照片有5張 , 內臉與5比對照片相符 , 平均檢測到44.40張 , 人臉辨識信心度為0.45 , **chen**的造假影片機率為1.0

**la** : 比對照片有5張 , 內臉與5比對照片相符 , 平均檢測到91.00張 , 人臉辨識信心度為1.0 , **la**的造假影片機率為1.0

**su** : 比對照片有5張 , 內臉與所有比對照片不符

**tsai** : 比對照片有5張 , 內臉與所有比對照片不符

---

### Summary:

檢測到影片有100%有**la**的人臉 , 造假率為1.0 , 此影片可能是**la**的假影片



考慮到通用：  
可以學代表  
性的深度偽  
造方法就好  
嗎？

---

ICIP' 19.

Best Student Paper Award

# 高通用性的偽造偵測器

- 引入 "共通偽造特徵" 的概念!!
  - 不同偽造方法之間，會不會共享相似的偽造特徵？
  - 如果會，那麼未來新的偽造方法，可能也會存有此瑕疵!!



# Objective Quality Comparison

提出方法與其他方法之間的性能比較

| Method/Test target  | LSGAN        |              | DCGAN        |              | WGAN         |              | WGAN-GP      |              | PGGAN        |              |
|---------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                     | precision    | recall       | precision    | recall       | precision    | recall       | precision    | recall       | precision    | recall       |
| Method in [5]       | 0.205        | 0.580        | 0.253        | 0.774        | 0.235        | 0.673        | 0.242        | 0.604        | 0.222        | 0.862        |
| Method in [7]       | 0.819        | 0.528        | 0.848        | 0.790        | 0.817        | 0.822        | 0.816        | 0.679        | 0.798        | 0.788        |
| Method in [8]       | 0.833        | 0.725        | 0.812        | 0.833        | 0.840        | 0.809        | 0.826        | 0.733        | 0.824        | 0.838        |
| Method in [15]      | 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        |
| <b>The proposed</b> | <b>0.981</b> | <b>0.956</b> | <b>0.986</b> | <b>0.986</b> | <b>0.895</b> | <b>0.881</b> | <b>0.876</b> | <b>0.881</b> | <b>0.951</b> | <b>0.936</b> |

## Visualization of the Unrealistic Details in the Fake Image

- 完全卷積網絡可用於可視化偽造的細節





好，如果真的要訓練，能不能只訓練  
少少資料，就可以學會辨識？

We only look once!

## 對抗少樣本情況下的深度偽造偵測

許多故意操縱的臉部可能結合了「多個」後處理方法

它可能是獨特的...

因此，您可能只能看一次!!



低樣本學習是必要的!!

結合自監督式學習 (chatGPT背後技術之一)



打造獨特的偵測網路系統

## Generalization?

- 交叉類型驗證
  - 在1%的訓練標籤設置下

|                                | FS           | F2F          | DF           | NT           |
|--------------------------------|--------------|--------------|--------------|--------------|
| Face X-ray [10]                | 0.440        | 0.413        | 0.444        | 0.451        |
| WA [9]                         | 0.510        | 0.531        | 0.589        | 0.453        |
| RNN-based [6]                  | 0.494        | 0.513        | 0.513        | 0.421        |
| XceptionNet [4]                | 0.516        | 0.493        | 0.463        | 0.434        |
| SSLM [11]                      | 0.522        | 0.520        | 0.468        | 0.489        |
| DeepFD (our previous work) [1] | 0.679        | 0.584        | 0.570        | 0.556        |
| ViT-DeepFake [33]              | 0.434        | 0.535        | 0.515        | 0.499        |
| <b>Proposed SCT</b>            | <b>0.710</b> | <b>0.772</b> | <b>0.806</b> | <b>0.695</b> |

# 真實場域：影片都有壓縮，沒人不壓縮的啦！

建立可以抵抗壓縮攻擊的深度偽造方法

2025/7/2

壓縮很常發生，並且會造成很大影響



■ 衆所周知...

- DeepFake 會產生非常逼真的細節 ..
- 這意味著造假的線索成分變得很少!

■ 壓縮，會破壞 細節和 造假的線索!!

# 那我們也結合生成式模型GAI

- 視覺訊號容易受影響!!
  - 引入不同的模態
  - 引入GAI當補充資訊



# 我們能做到更穩定的效能

- 不管什麼壓縮，都難不倒我們

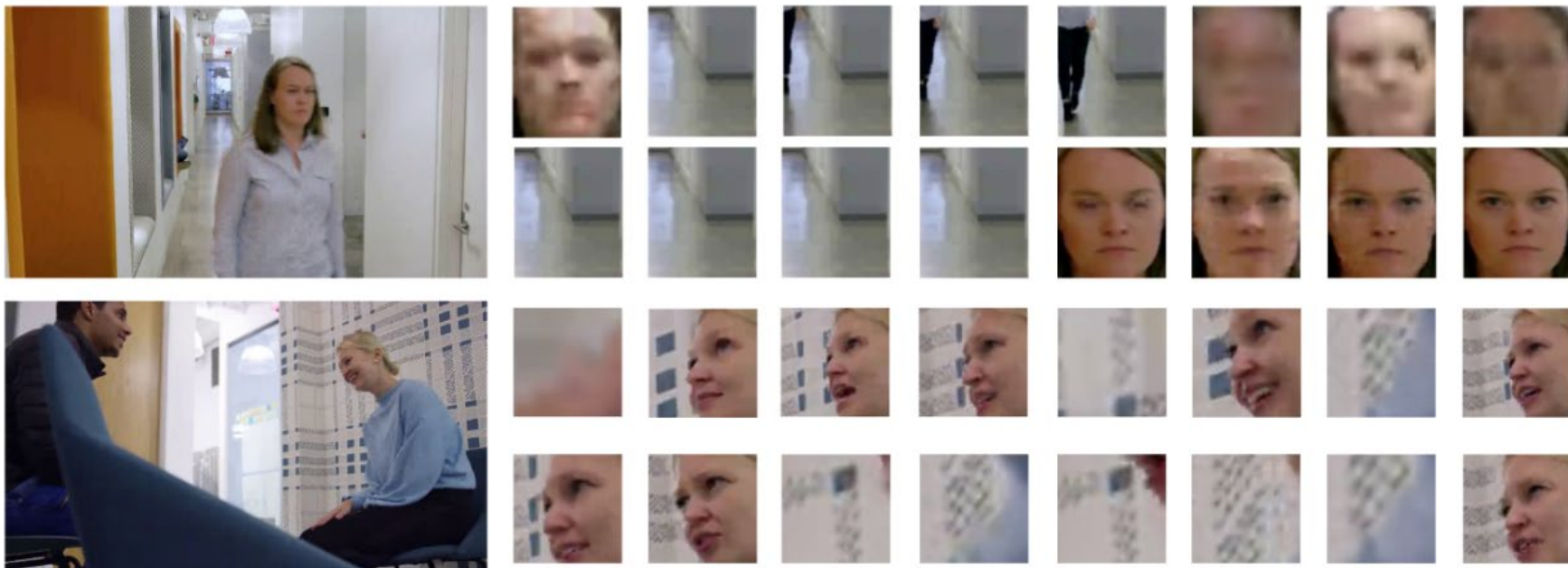
| Training dataset | Method            | Testing Set          |                      |                      |                      |                       |                      |
|------------------|-------------------|----------------------|----------------------|----------------------|----------------------|-----------------------|----------------------|
|                  |                   | FF++ (raw)           |                      | FF++ (c23)           |                      | FF++ (c40)            |                      |
|                  |                   | AUC(%)               | ACC(%)               | AUC(%)               | ACC(%)               | AUC(%)                | ACC(%)               |
| FF++ (raw)       | Xception [10]     | 96.32 (-)            | 96.25 (-)            | 74.92 (↓21.40)       | 79.25 (↓17.00)       | 63.49 (↓32.83)        | 62.25 (↓34.07)       |
|                  | LRNet [37]        | 99.10 (-)            | <b>98.24 (-)</b>     | 84.00 (↓15.10)       | 89.50 (↓8.74)        | 77.68 (↓21.42)        | 74.25 (↓23.99)       |
|                  | UIA-ViT [3]       | <b>99.33 (-)</b>     | 97.25 (-)            | 72.27 (↓27.06)       | 92.50 (↓4.75)        | 63.80 (↓35.53)        | 74.75 (↓22.50)       |
|                  | DFIL [5]          | 96.60 (-)            | 87.25 (-)            | 63.77 (↓32.83)       | 72.11 (↓15.14)       | 55.91 (↓40.69)        | 56.22 (↓31.03)       |
|                  | MaskRelation [46] | 97.50 (-)            | 82.55 (-)            | 68.28 (↓29.22)       | 78.75 (↓3.80)        | 58.49 (↓39.01)        | 58.75 (↓23.80)       |
|                  | UMCL (Ours)       | 98.21 (-)            | 92.49 (-)            | <b>95.69 (↓2.52)</b> | <b>96.53 (↑4.04)</b> | <b>80.21 (↓18.00)</b> | <b>84.47 (↓8.02)</b> |
| FF++ (c23)       | Xception [10]     | 73.76 (↓24.64)       | 67.20 (↓28.88)       | 98.40 (-)            | 96.08 (-)            | 63.49 (↓34.91)        | 56.50 (↓39.58)       |
|                  | LRNet [37]        | 84.01 (↓14.60)       | 90.51 (↓5.99)        | 98.61 (-)            | <b>96.50 (-)</b>     | 77.78 (↓18.83)        | 74.53 (↓21.97)       |
|                  | UIA-ViT [3]       | 75.53 (↓23.16)       | 88.88 (↓3.67)        | <b>98.69 (-)</b>     | 92.55 (-)            | 65.31 (↓33.38)        | 74.22 (↓18.33)       |
|                  | DFIL [5]          | 69.11 (↓26.03)       | 86.41 (↑1.16)        | 95.14 (-)            | 85.25 (-)            | 60.79 (↓34.35)        | 69.27 (↓15.98)       |
|                  | MaskRelation [46] | 68.10 (↓28.36)       | 82.75 (↑0.75)        | 96.46 (-)            | 82.00 (-)            | 61.36 (↓35.10)        | 62.75 (↓33.71)       |
|                  | UMCL (Ours)       | <b>96.46 (↓0.13)</b> | <b>97.40 (↑2.23)</b> | 96.59 (-)            | 95.17 (-)            | <b>80.57 (↓16.02)</b> | <b>86.57 (↓8.60)</b> |
| FF++(c40)        | Xception [10]     | -                    | -                    | -                    | -                    | 92.40 (-)             | <b>94.52 (-)</b>     |
|                  | LRNet [37]        | 77.59 (↓18.55)       | 84.25 (↓6.00)        | 77.79 (↓18.35)       | 84.25 (↓6.00)        | <b>96.14 (-)</b>      | 90.25 (-)            |
|                  | UIA-ViT [3]       | 66.57 (↓28.54)       | 86.71 (↑3.61)        | 66.70 (↓28.41)       | 85.71 (↑2.61)        | 95.11 (-)             | 83.10 (-)            |
|                  | DFIL [5]          | 62.51 (↓27.04)       | 80.00 (↑0.49)        | 63.48 (↓26.07)       | 73.13 (↓6.38)        | 89.55 (-)             | 79.51 (-)            |
|                  | MaskRelation [46] | 64.74 (↓25.31)       | 78.50 (↓1.11)        | 64.88 (↓25.17)       | 76.55 (↓3.06)        | 90.05 (-)             | 79.61 (-)            |
|                  | UMCL (Ours)       | <b>94.33 (↑3.63)</b> | <b>96.52 (↑4.77)</b> | <b>94.43 (↑3.73)</b> | <b>95.39 (↑3.64)</b> | 90.70 (-)             | 91.75 (-)            |



# 真實場域：現在都假設人臉偵測的到，如果抓不到呢?

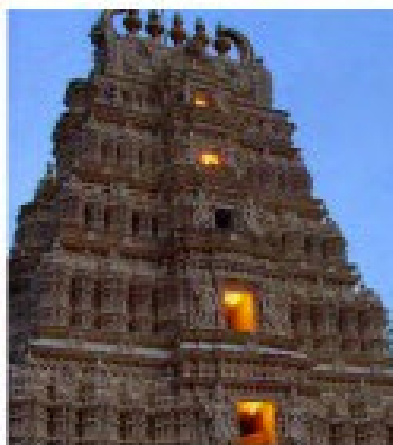
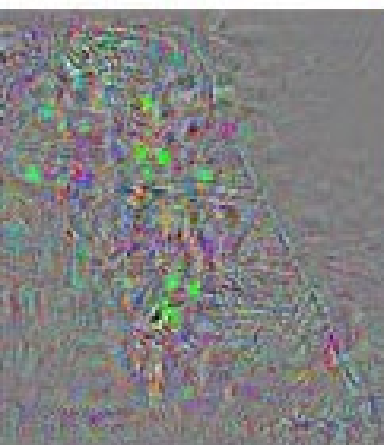
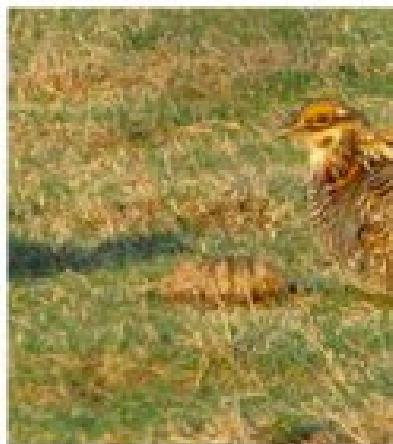
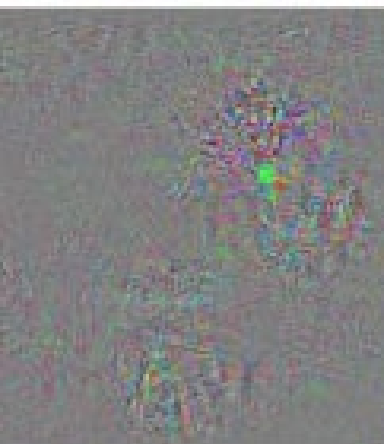
針對不穩定人臉序列下的偵測方法

# 老實說，人臉偵測器常常不穩定



我們提出一個特徵糾纏方法解決此問題，不管怎麼不穩，都不怕

| Method            | Venue     | Type        | $m_r$ | FF++ [7] |       |       | Celeb-DF [8] |       |       | DFDC [9] |       |       |
|-------------------|-----------|-------------|-------|----------|-------|-------|--------------|-------|-------|----------|-------|-------|
|                   |           |             |       | Acc.     | F1    | AUC   | Acc.         | F1    | AUC   | Acc.     | F1    | AUC   |
| Xception [1]      | CVPR2017  | Image-based | 0.0   | 0.925    | 0.894 | 0.972 | 0.861        | 0.806 | 0.910 | 0.953    | 0.910 | 0.981 |
|                   |           |             | 0.4   | 0.869    | 0.780 | 0.871 | 0.631        | 0.614 | 0.782 | 0.908    | 0.788 | 0.866 |
|                   |           |             | 0.8   | 0.814    | 0.594 | 0.654 | 0.398        | 0.389 | 0.604 | 0.864    | 0.598 | 0.647 |
| $F^3$ -net [2]    | ECCV2020  | Image-based | 0.0   | 0.950    | 0.928 | 0.986 | 0.965        | 0.957 | 0.993 | 0.957    | 0.921 | 0.986 |
|                   |           |             | 0.4   | 0.883    | 0.798 | 0.888 | 0.691        | 0.684 | 0.895 | 0.864    | 0.655 | 0.755 |
|                   |           |             | 0.8   | 0.818    | 0.599 | 0.662 | 0.418        | 0.407 | 0.664 | 0.850    | 0.539 | 0.595 |
| RECCE [5]         | CVPR2022  | Image-based | 0.0   | 0.938    | 0.911 | 0.979 | 0.941        | 0.925 | 0.985 | 0.940    | 0.872 | 0.973 |
|                   |           |             | 0.4   | 0.878    | 0.790 | 0.874 | 0.678        | 0.669 | 0.869 | 0.900    | 0.752 | 0.863 |
|                   |           |             | 0.8   | 0.817    | 0.599 | 0.655 | 0.414        | 0.404 | 0.648 | 0.861    | 0.579 | 0.648 |
| CORE [26]         | CVPRW2022 | Image-based | 0.0   | 0.948    | 0.925 | 0.984 | 0.953        | 0.940 | 0.989 | 0.950    | 0.903 | 0.977 |
|                   |           |             | 0.4   | 0.883    | 0.799 | 0.888 | 0.858        | 0.790 | 0.890 | 0.907    | 0.781 | 0.870 |
|                   |           |             | 0.8   | 0.818    | 0.601 | 0.663 | 0.764        | 0.572 | 0.661 | 0.863    | 0.595 | 0.651 |
| UCF [29]          | CVPR2023  | Image-based | 0.0   | 0.937    | 0.911 | 0.982 | 0.856        | 0.792 | 0.891 | 0.890    | 0.815 | 0.939 |
|                   |           |             | 0.4   | 0.875    | 0.790 | 0.882 | 0.626        | 0.607 | 0.642 | 0.871    | 0.733 | 0.812 |
|                   |           |             | 0.8   | 0.815    | 0.598 | 0.660 | 0.397        | 0.389 | 0.516 | 0.851    | 0.586 | 0.620 |
| DFIL [27]         | ACMMM2023 | Image-based | 0.0   | 0.954    | 0.939 | 0.987 | 0.957        | 0.954 | 0.964 | 0.940    | 0.881 | 0.955 |
|                   |           |             | 0.4   | 0.876    | 0.808 | 0.893 | 0.695        | 0.684 | 0.825 | 0.886    | 0.720 | 0.813 |
|                   |           |             | 0.8   | 0.759    | 0.603 | 0.665 | 0.518        | 0.350 | 0.644 | 0.855    | 0.565 | 0.621 |
| TALL-Swin [28]    | ICCV2023  | Video-based | 0.0   | 0.913    | 0.868 | 0.881 | 0.913        | 0.933 | 0.924 | 0.911    | 0.812 | 0.984 |
|                   |           |             | 0.4   | 0.867    | 0.767 | 0.740 | 0.847        | 0.789 | 0.825 | 0.872    | 0.758 | 0.786 |
|                   |           |             | 0.8   | 0.827    | 0.605 | 0.589 | 0.745        | 0.680 | 0.645 | 0.845    | 0.688 | 0.650 |
| DFGaze [30]       | TIFS2024  | Video-based | 0.0   | 0.946    | 0.926 | 0.986 | 0.956        | 0.954 | 0.972 | 0.915    | 0.881 | 0.968 |
|                   |           |             | 0.4   | 0.854    | 0.724 | 0.795 | 0.756        | 0.726 | 0.824 | 0.818    | 0.709 | 0.743 |
|                   |           |             | 0.8   | 0.785    | 0.652 | 0.656 | 0.612        | 0.669 | 0.659 | 0.798    | 0.648 | 0.596 |
| MaskRelation [33] | TIFS2023  | Video-based | 0.0   | -        | -     | -     | -            | -     | -     | -        | -     | -     |
|                   |           |             | 0.4   | -        | -     | -     | -            | -     | -     | -        | -     | -     |
|                   |           |             | 0.8   | -        | -     | -     | -            | -     | -     | -        | -     | -     |
| GRACE [Ours]      | -         | Video-based | 0.0   | 0.962    | 0.942 | 0.989 | 0.989        | 0.968 | 0.998 | 0.969    | 0.942 | 0.988 |
|                   |           |             | 0.4   | 0.958    | 0.936 | 0.987 | 0.970        | 0.920 | 0.998 | 0.969    | 0.940 | 0.988 |
|                   |           |             | 0.8   | 0.944    | 0.916 | 0.983 | 0.857        | 0.738 | 0.980 | 0.962    | 0.925 | 0.979 |



真實場域：攻擊者可以騙過人眼，也可以騙過AI吧

對抗攻擊 + 深度偽造

*IEEE TPAMI 23*

# 核心概念：對抗攻擊騙AI，我們騙攻擊者

AUC scores / **accuracy** (%)

| Dataset |             | Clean | FGSM               | PGD (iteration 5)  |                   |                   | PGD (iteration 15) |                   |                   | C&W               |
|---------|-------------|-------|--------------------|--------------------|-------------------|-------------------|--------------------|-------------------|-------------------|-------------------|
|         |             |       |                    | $\epsilon = 0.005$ | $\epsilon = 0.01$ | $\epsilon = 0.02$ | $\epsilon = 0.005$ | $\epsilon = 0.01$ | $\epsilon = 0.02$ |                   |
| FF++    | Meso        | 97.6  | 0.0 / <b>0.0</b>   | 0.0 / <b>0.0</b>   | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>   | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>  |
|         | Xception    | 99.8  | 0.0 / <b>14.29</b> | 0.1 / <b>72.6</b>  | 0.0 / <b>1.3</b>  | 0.0 / <b>0.0</b>  | 0.0 / <b>43.9</b>  | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>  |
|         | MesoD       | 97.6  | 100 / <b>100</b>   | 100 / <b>100</b>   | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>100</b>   | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>100</b>  |
|         | XD          | 99.8  | 98.8 / <b>99.1</b> | 100 / <b>100</b>   | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>100</b>   | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>100</b>  |
|         | MesoD (c23) | 75.1  | 100 / <b>100</b>   | 100 / <b>100</b>   | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>100</b>   | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>98.1</b> |
|         | XD (c23)    | 71.0  | 99.3 / <b>99.7</b> | 100 / <b>99.7</b>  | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>100</b>   | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>100</b>  |
|         | MesoD (c40) | 61.1  | 100 / <b>100</b>   | 100 / <b>100</b>   | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>100</b>   | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>99.5</b> |
|         | XD (c40)    | 66.1  | 99.6 / <b>99.0</b> | 100 / <b>99.9</b>  | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>100</b>   | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>99.1</b> |
| DFDC    | Meso        | 51.8  | 0.0 / <b>0.0</b>   | 0.0 / <b>0.0</b>   | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>   | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>  |
|         | Xception    | 63.0  | 0.0 / <b>14.5</b>  | 0.0 / <b>48.5</b>  | 0.0 / <b>9.5</b>  | 0.0 / <b>0.0</b>  | 0.0 / <b>36.0</b>  | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>  |
|         | MesoD       | 59.8  | 100 / <b>100</b>   | 100 / <b>100</b>   | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>100</b>   | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>100</b>  |
|         | XD          | 66.1  | 95.5 / <b>93.0</b> | 99.7 / <b>99.0</b> | 100 / <b>100</b>  | 100 / <b>100</b>  | 99.9 / <b>99.8</b> | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>100</b>  |
| CDF     | Meso        | 68.3  | 0.0 / <b>0.0</b>   | 0.0 / <b>0.0</b>   | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>   | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>  |
|         | Xception    | 51.6  | 0.0 / <b>14.5</b>  | 0.0 / <b>50.0</b>  | 0.0 / <b>3.0</b>  | 0.0 / <b>1.0</b>  | 0.0 / <b>45.0</b>  | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>  | 0.0 / <b>0.0</b>  |
|         | MesoD       | 61.2  | 100 / <b>100</b>   | 100 / <b>100</b>   | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>100</b>   | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>100</b>  |
|         | XD          | 55.4  | 100 / <b>97.5</b>  | 100 / <b>100</b>   | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>100</b>   | 100 / <b>100</b>  | 100 / <b>100</b>  | 100 / <b>99.5</b> |

真實世界的場域：通常 DEEPFAKE 出現在社群媒體上

那，就結合社群媒體資訊來做深度偽造偵測！

輔助資訊有用：那麼更多資訊呢？



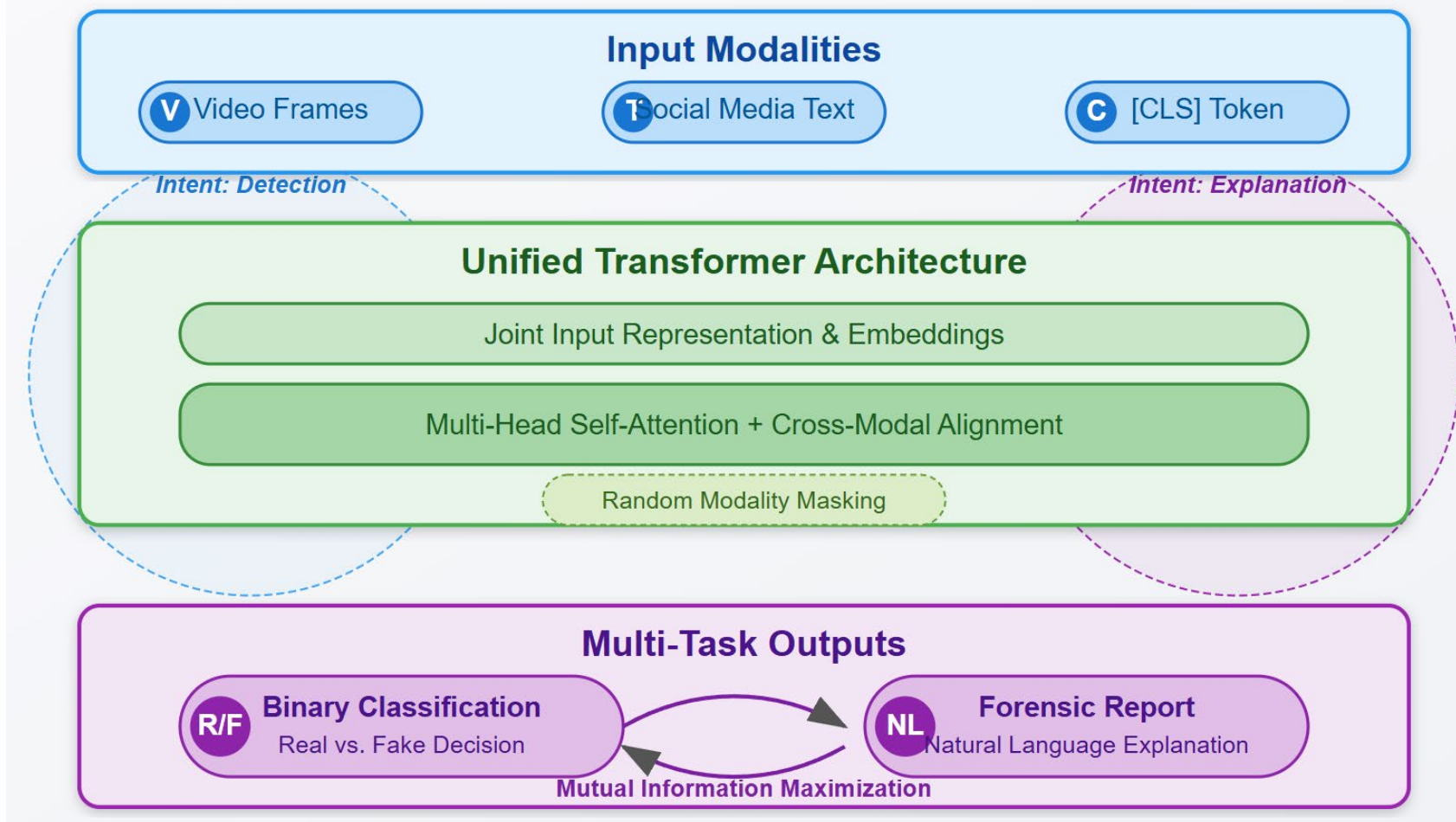
Deepfake detection的研究中，目前對於這方面的研究多是分別針對影像或文字的造假進行偵測，尚缺乏整體的研究及技術。



本研究預計透過多模資料分析、建構出結合社群資料與多媒體偽造偵測器的方法，提高在真實應用場景使用的便利性。

結合？容易嗎？

## DeepFake Detection & Explanation Framework





這麼說來，應該可以對抗 DEEPFAKE 了吧？



事實上，人的判斷能力往往不比 AI 差。

舉例來說，如果給你 (妳) 足夠的時間，大家應該有能力透過**常識**、其它**輔助資料**、甚至是**發文來源**就能判斷是否偽造

- 例如，教宗 + 大衣 → 常識推論不合理
- 例如，英國女皇不會很誇張的跳舞 → 不合常理



## DeepFake 還有太多待解的問題

\*真的壓縮很嚴重的情況，大概什麼方法都無效...

\*只有部分被偽造，或是刻意加入模糊或雜訊來混淆視聽

\*太多不同偽造的手法

\*AI永遠只是輔助工具，決策者仍在人

\*AI 是偽造，那麼手動 / 半自動 換臉，DeepFake仍能抓到嗎？

# So what can we do?

## ■ 生成初步查核內容，進一步合作

### MAIN FUNCTION 主功能

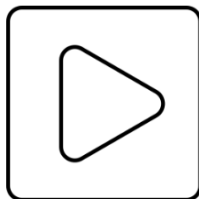
Upload an image/video to verify its authenticity. Note: This tool is for academic use only.

請上傳影像/影片以確認其真實性。註：本項目僅限學術用途。

Step 1: Choose an Image / Video

CHOOSE A FILE

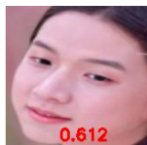
Only accept 'png', 'jpg', 'mp4', 'mov' file extension, and file size should be under 50MB.



Step2: Upload

### DETECTION RESULT 偵測結果

Visualized Results



This video could be fake..., (expand for more info 看更多)

#### Detection Details

[1-th Face] Fake Prob. of Our TwoStream Detector: 0.2520  
[1-th Face] Fake Prob. of SSSNet: 0.9900

本次查核針對影片XXX，透過多種深度偽造 (DeepFake) 偵測技術進行分析，結果顯示該影片存在高度偽造的可能性，並且在特定畫面中尤為明顯。從數據上來看，該影片的最終深偽造得分為 **0.766**，顯示偽造成分相當顯著。GRACE-Detector 和 Det-v2 皆給出 **1.0** 的偽造機率，這代表這些演算法對影片的偽造特徵有極高的信心。此外，SSSNet 的偽造機率為 **0.990**，而 Diffusion+SMI 亦達到 **0.729**，顯示不同演算法均偵測到影片的異常點。進一步分析偽造畫面，我們發現 Det-v1 在 **第 86 與 87 幀** 檢測到 **連續高偽造得分**，顯示該區段的畫面可能有特別明顯的偽造痕跡。此外，影片中還有 **16 個其他高偽造畫面**，但這些畫面並未形成明確的連續性，顯示偽造特徵在不同畫面中分佈較為零散。

面部尺寸的分析也提供了一些線索。影片中的面部大小範圍從 **最小 42x63 到最大 124x171**，平均面部大小為 **78.3x106.4**。雖然只有 **1 個面部的寬高小於 50 像素**，但這顯示影片中大多數面部應該是清晰可辨的，這也讓偽造痕跡更容易暴露。

#### 如何透過這些證據判斷偽造

##### 1. 關鍵畫面 (86-87 幀) 顯示出偽造特徵的連續性

1. 這表示在這兩幀之間，可能發生了人臉合成錯誤、換臉邊界異常或表情不自然的現象。
2. 這是影片中最值得關注的區域，應該逐幀檢視。

##### 2. 不同偵測器均顯示高偽造機率

1. GRACE-Detector 與 Det-v2 都給出滿分 1.0，SSSNet 也幾乎滿分，這代表即便使用不同演算法進行分析，仍然得出高度一致的偽造判定。
2. 這顯示影片中偽造的特徵並非個別檢測器的誤判，而是有一致性的偽造跡象。

##### 3. 偽造分數的分佈模式

1. Det-v1 雖然平均得分較低 (0.252)，但最高分達 **0.853**，顯示某些畫面特別可疑。
2. Diffusion-Det 的最高偽造分數為 **0.729**，平均達 **0.636**，顯示即便是在較平穩的畫面中，也有較高的偽造機率。

#### 結論

根據多種偵測器的高一致性結果，以及關鍵畫面 86-87 幀的異常特徵，該影片極可能為深度偽造內容。從統計數據來看，影片的偽造特徵分佈相對集中，特定畫面的異常更為明顯，因此透過適當的逐幀檢視與對比，可以進一步確認影片的偽造程度。

## (DeepFake偵測工具的侷限性) 研究方法上

能不能只鎖定特定人物來偵測就好?

- 以特定人物為主，因此對於一般民眾可能相對比較不好用

多模態的分析方法

- 適合Social media的相關公司、一般民眾不適用

要怎麼樣可以增加偵測器的通用能力?

- 通用型，但是辨識能力可能較低，要仔細看分析報告

我們如何能夠始終擁有大量的篡改影像資料?

- 通用型，辨識能力較低，要仔細看分析報告

偵測器遇上壓縮!!

- 相對穩定，但是仍要看分析報告

不穩定的人臉偵測器？不穩定的人臉序列

- 相對穩定，但是仍要看分析報告

聯合攻擊？對抗攻擊 + 人臉操作

- 相對穩定，但是仍要看分析報告

# 建議流程

影像來源鑑識

人為查證

AI模型分析 ， 查驗分析報告

總結結果

做出判斷

## 仔細查驗報告

- 例如以DeepWare來說，我們用一個很簡單的換臉程式來生成假影片上傳分析

 **DEEFAKE DETECTED**

New Scan



**Name:** BrianDF\_FirstTrain\_1109.mp4

**Size:** 73.2 MB

User

2021-12-16 09:03:08 UTC

Source

2 year(s) ago

DETAILS

Deepware aims to give an opinion about the scanned video and is not responsible for the result. As Deepware Scanner is still in beta, the results should not be treated as an absolute truth or evidence.

**Model Results**

**Avatarify:** Seferbekov

**Deepware:** Best performing single model selected from seferbekov's ensemble. Only the weights were used, our deepware scanner produces results.

**Seferbekov:**

**Ensemble:**

[Github](#)

**Video**

**Duration:** 58 sec

**Resolution:** 1920 x 1080

**Frame Rate:** 60 fps

**Codec:** h264

**Audio**

**Duration:** 58 sec

**Channel:** stereo

**Sample Rate:** 48 khz

**Codec:** aac

Request Expert Review

Request Takedown

2025/7/2

Chih-Chung Hsu@ACVLab

86

ACVLab

比方說以我們的分析技術為例



# DEEPFAKE DETECTION ONLINE

## 深偽鑑識技術





# DEEPFAKE DETECTION ONLINE

## 深偽鑑識技術



### MORE INFO

X

#### Video Information

Video Name: 子楊DF影片檔第1次BrianDF\_FirstTrain\_1109.mp4  
Video Resolution: 1920.0x1080.0  
Video FPS: 60.0  
Video Length: 600  
Number of Detected Person: 1  
Mean Number of Detected Faces: 298.0

#### Detection Details

[1-th Face] Fake Prob. of Traditional Deepfake: 0.1980  
[1-th Face] Fake Prob. of Face Manipulation: 0.2310  
**[1-th Face] Fake Prob. of Face Generation: 0.7950**  
[1-th Face] Fake Prob. of Facial Inconsistency: 0.0030

MORE INFORETRY

## 以HIVE偵測DeepFake為例

- 抓不出來？有時可以有時不行。我們無法完全確保這些 AI 技術 100% 可信，因此最終還是回歸到與人共同決策上

Upload images here to test our model in real-time!  
Supports png, jpeg, jpg, webp Use is subject to this site's [Terms of Service](#).



Upload

**RESULT**

The input is: **not likely to be AI Generated**

**6.2%**

**BY CLASSES**

| Classes          | Score |
|------------------|-------|
| not_ai_generated | 0.93  |
| none             | 0.93  |
| ai_generated     | 0.06  |
| stablediffusion  | 0.06  |
| dalle            | 0.00  |
| midjourney       | 0.00  |

HIVE MODERATION

## 現行方法的侷限性

大多數的目標還是著重在 DeepFake 偵測

特別是人臉影像



但是隨著現在 Diffusion-based AIGC 越來發展越好，以這為基礎的 DeepFake 很快就會出現

現行軟體也只有相對不到10%有支援

我們也有相關研究方法



但值得注意的是，觀察分析報告，才是重點

那麼，人眼能不能觀察出什麼線索？

- 可以的！要相信你自己的強大！熟能生巧。

- 例如.

- <http://pse.is/59mcwm>

- 上去多看看，熟能生巧

- Supported by **AI.FREE**



## 自然圖片分辨不出來，那不然人臉好了: Which face is real?

- <https://www.whichfaceisreal.com/>  
<https://farid.berkeley.edu/misc/AlorNotQuiz/>

You are **correct**. The image on the left is real.

[Play again.](#)



## 但請牢記幾點

### ■ AI技術仍不斷的演進

- 這些瑕疵，也會隨著時間漸漸變少
- 了解技術仍然還是重要的
- 要知道技術通常是寧可錯殺不可放過

### • 技術發展的侷限性：

- 跟人一樣，不會有什麼技術是100%可靠。
- 人機協作會是最重要的出路。



## 所以到底什麼是 DeepFake?

人臉身分交換  
(換臉)

整張人臉生成

整張影像 / 影片生成

影像對影像轉換

影像物件移除  
(Inpainting)

純AI

人工+AI

純人工

# 未來展望與挑戰

## 技術發展趨勢

### 生成品質持續提升

AI生成內容的品質將持續提高，逐漸難以與真實內容區分

### 即時生成能力增強

從批次處理到即時生成，使Deepfake和AI生成內容可即時應用

### 移動端應用普及

生成技術將更廣泛地應用於移動設備，降低使用門檻

### 多模態融合發展

影像、聲音、文字等多模態生成技術將更緊密結合

## 社會與倫理挑戰

### 隱私與肖像權保護

如何在技術發展的同時保護個人隱私和肖像權

### 資訊真實性危機

AI生成內容可能加劇假新聞和錯誤資訊傳播

### 法律與監管滯後

法律法規難以跟上技術發展速度，監管存在滯後性

### 社會信任危機

難以分辨真假內容可能導致社會普遍信任危機

## 技術應用前景

2030年AI生成內容應用領域預測



AI影像生成技術將在**創意產業、教育培訓、醫療健康、個人娛樂**、等領域有廣泛應用。預計到2030年，全球AI生成內容市場規模將達到 **1500億美元**，年複合增長率超過 **30%**

## 平衡發展建議

### 多方協作治理

政府、企業、學術界和公民社會共同參與AI治理

### 技術與倫理並重

在技術發展的同時加強倫理規範和自律機制

### 媒體素養教育

加強公眾對AI生成內容的辨識能力和批判思維

### 透明度與可追溯性

建立AI生成內容的標示和追溯機制，增加透明度

### 防禦技術發展

同步發展偵測和防禦技術，形成技術制衡

# 結語

## 重點總結

### 🕒 技術發展迅速

Deepfake和Diffusion Model技術正快速發展，生成內容品質不斷提高

### ⚖️ 雙面影響

這些技術既有創意應用潛力，也帶來隱私、安全和信任等社會挑戰

### 🛡️ 偵測與防禦

偵測技術正在發展，但與生成技術存在軍備競賽關係，需持續進步

### ⚖️ 法律與倫理

需要完善法律框架和倫理準則，平衡技術創新與社會保護

## 常見問題

### ❓ 如何辨識Deepfake或AI生成的內容？

注意不自然的臉部表情、眨眼頻率、光影不一致等視覺線索，並使用專業偵測工具輔助判斷

## 延伸閱讀

### 📖 學習資源

Stable Diffusion官方文檔：[stability.ai](https://stability.ai)

Midjourney使用指南：[docs.midjourney.com](https://docs.midjourney.com)

DeepFake Detection Challenge：[kaggle.com/dfdc](https://kaggle.com/dfdc)

### ✂️ 實用工具

AI內容偵測：[contentauthenticity.org](https://contentauthenticity.org)

開源Diffusion模型：[huggingface.co/models](https://huggingface.co/models)

台灣AI倫理協會：[taiea.org.tw](https://taiea.org.tw)

### 📰 最新研究與報導

台灣AI發展策略：[ai.nstc.gov.tw](https://ai.nstc.gov.tw)

AI生成內容監管趨勢：[law.gov.tw](https://law.gov.tw)

### ❓ 一般民眾如何保護自己免受Deepfake詐騙？

提高媒體素養，對突發緊急要求保持警惕，使用多重管道驗證身份，不輕信單一來源資訊

要試試看我們的偵測平台

(有限功能版)



關注更多新的技術發展請參考

<https://cchsu.info>

