

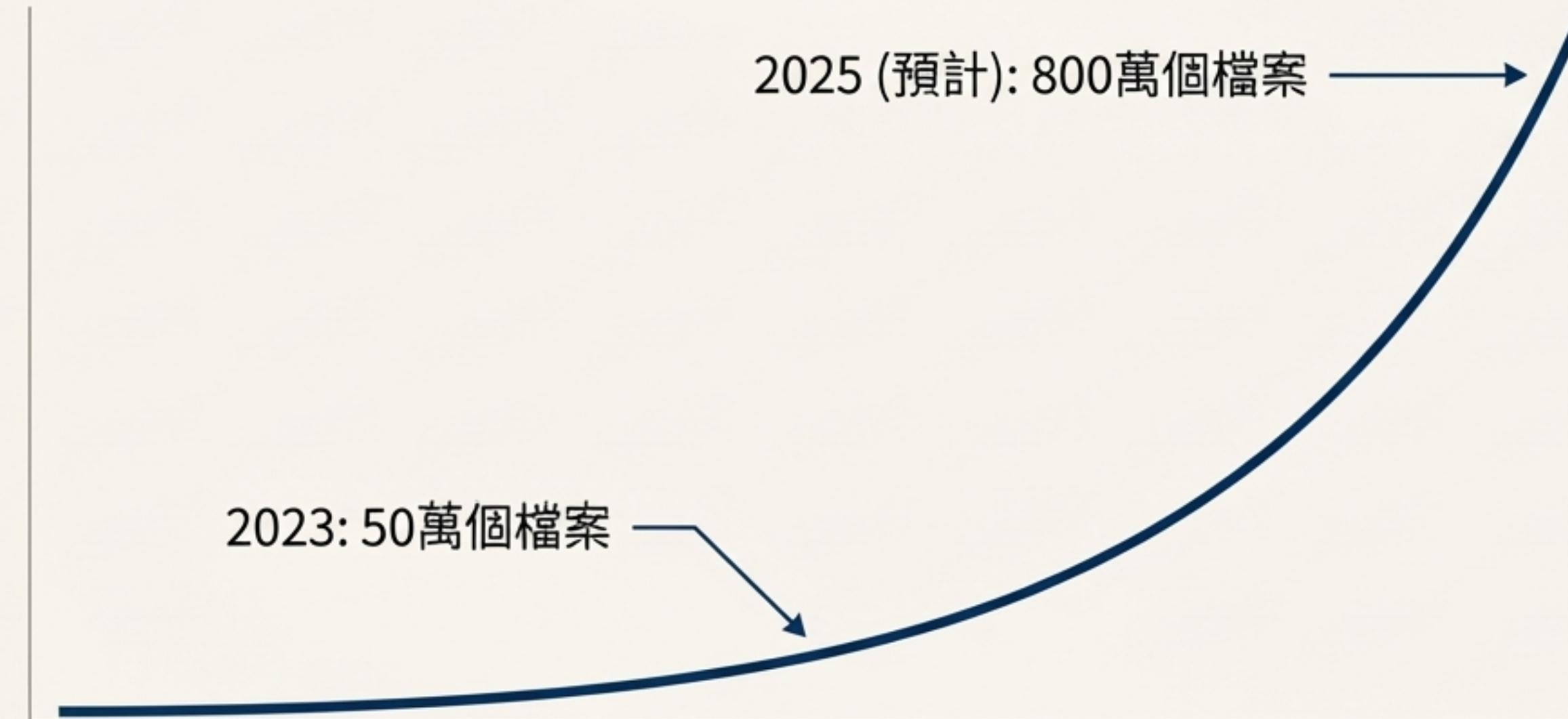


數位信任的終局： 在 Deepfake 時代下辨識真偽

剖析偽造影像的生成技術與偵測策略

威脅正以指數級速度增長

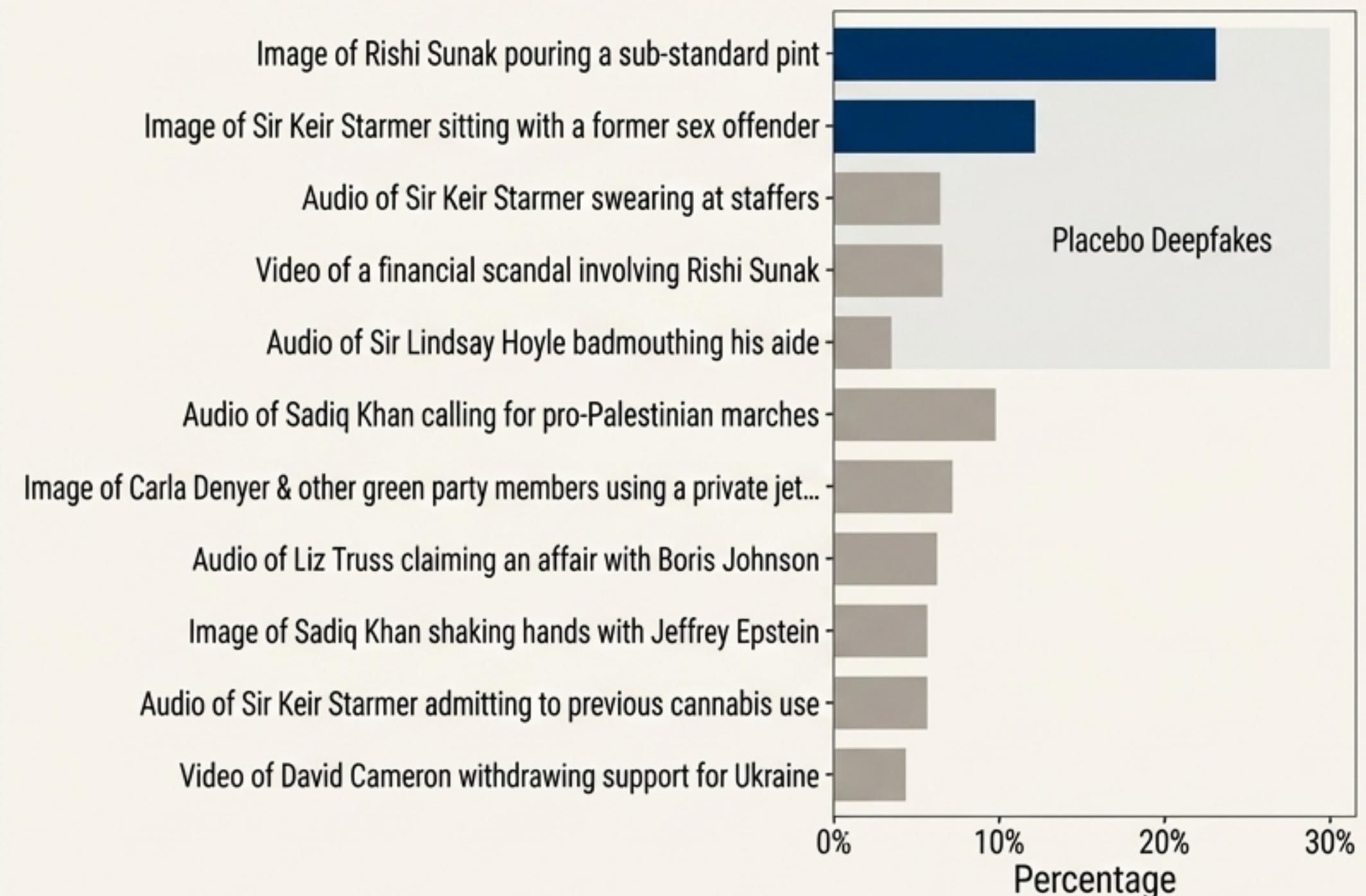
Deepfake 內容正呈指數級增長



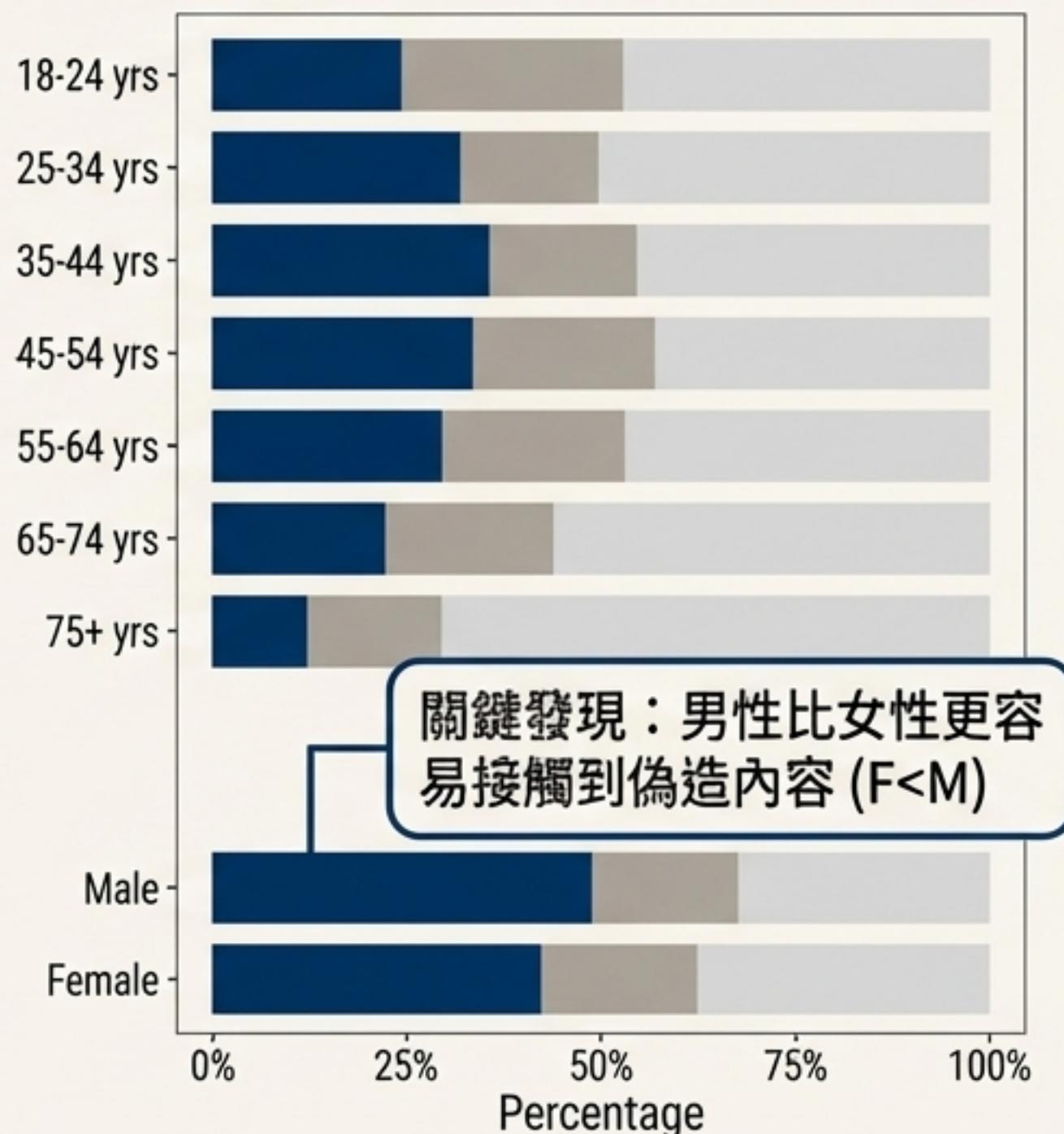
「這不僅是技術趨勢，更是對現實認知的結構性挑戰。」

從技術展示到政治武器

英國的政治輿論操作

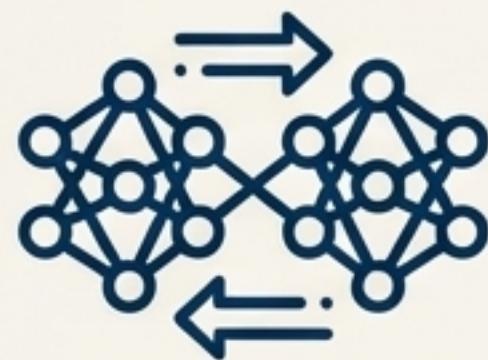


全面性的社會滲透

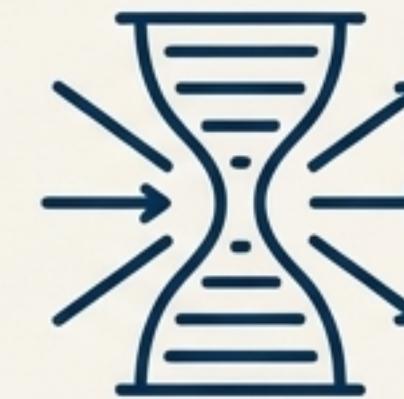


偽造的技術核心與我們的應對目標

偽造技術剖析 (Forgery Technology Breakdown)



生成對抗網路 (GAN) :透過兩個神經網路（生成器與判別器）的對抗學習，創造出極度逼真的偽造內容。



自編碼器 (Autoencoder) :學習將人臉壓縮成低維度編碼再重建，藉此實現臉部特徵的替換。

研究目標 (Research Objectives)

目標一：建立一個能自動偵測 Deepfake 換臉影像的深度學習分類模型。

目標二：提供模型可解釋的特徵視覺化，以分析 Deepfake 的生成特徵。

偵探的困境：為何辨識如此困難？

挑戰一：持續進化的軍備競賽 (Evolving Arms Race)



- Deepfake 生成技術不斷進化，今天的偵測模型可能明天就失效。
- 本次研究僅專注於換臉（Face Swap），但偽造類型涵蓋聲音、肢體等多種形式。

挑戰二：信任的崩潰—「騙子紅利」效應 (The Liar's Dividend)



- 當 Deepfake 變得無所不在，人們開始質疑所有真實影像的真實性，即使它們是真實的。
- 這導致公眾信任的侵蝕，使得偵測模型必須達到極高的準確度才能重建信任。

現有的鑑識武庫：兩種主流偵測思路



思路一：尋找頻譜指紋 (Frequency Domain Analysis)

核心假設：AI 生成的影像，即使肉眼無法分辨，其在頻率域中也存在著不自然的「指紋」。

方法：利用傅立葉轉換 (FFT) 等工具，分析影像的頻譜特徵。

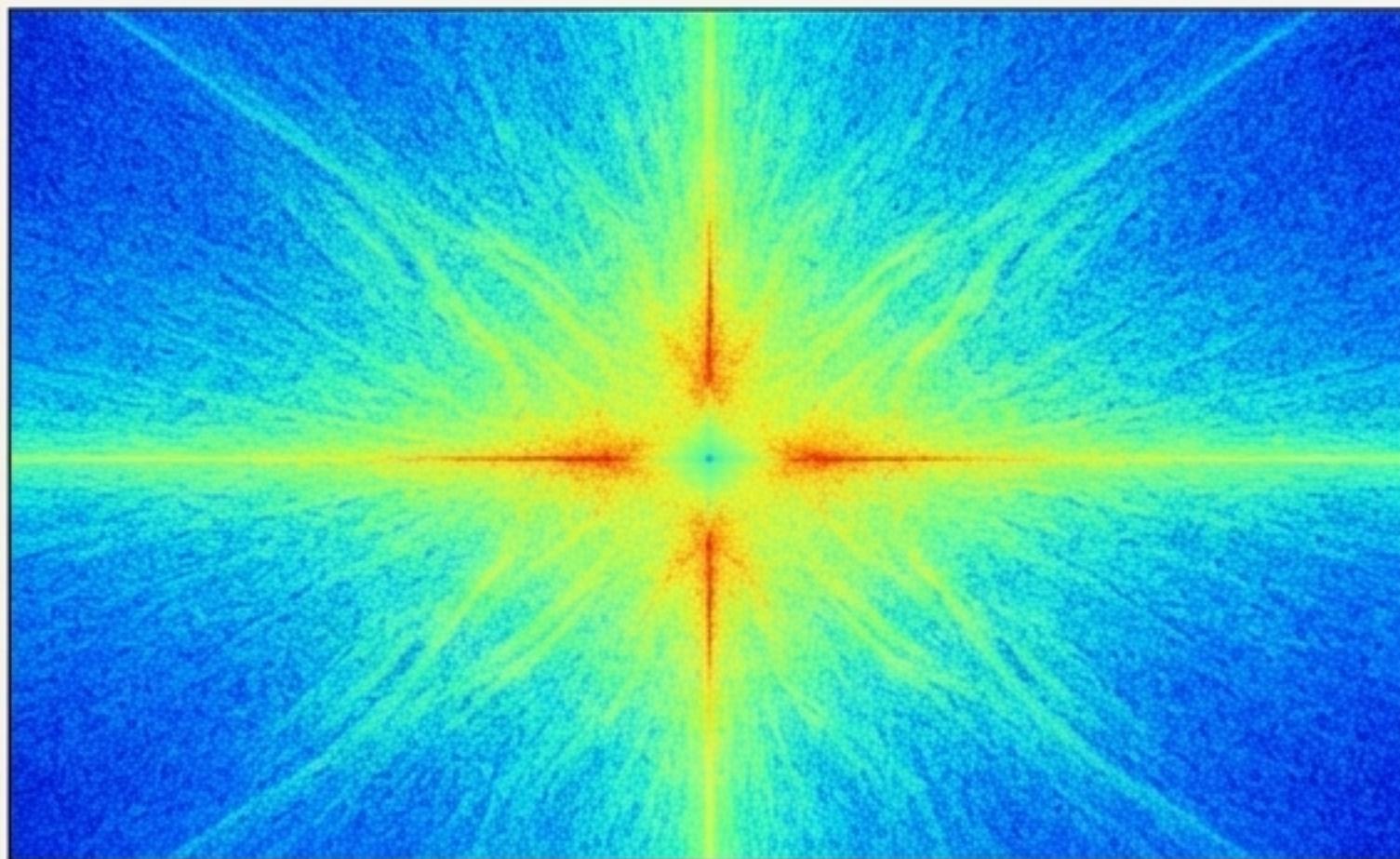


思路二：訓練 AI 獵犬 (Deep Learning Models)

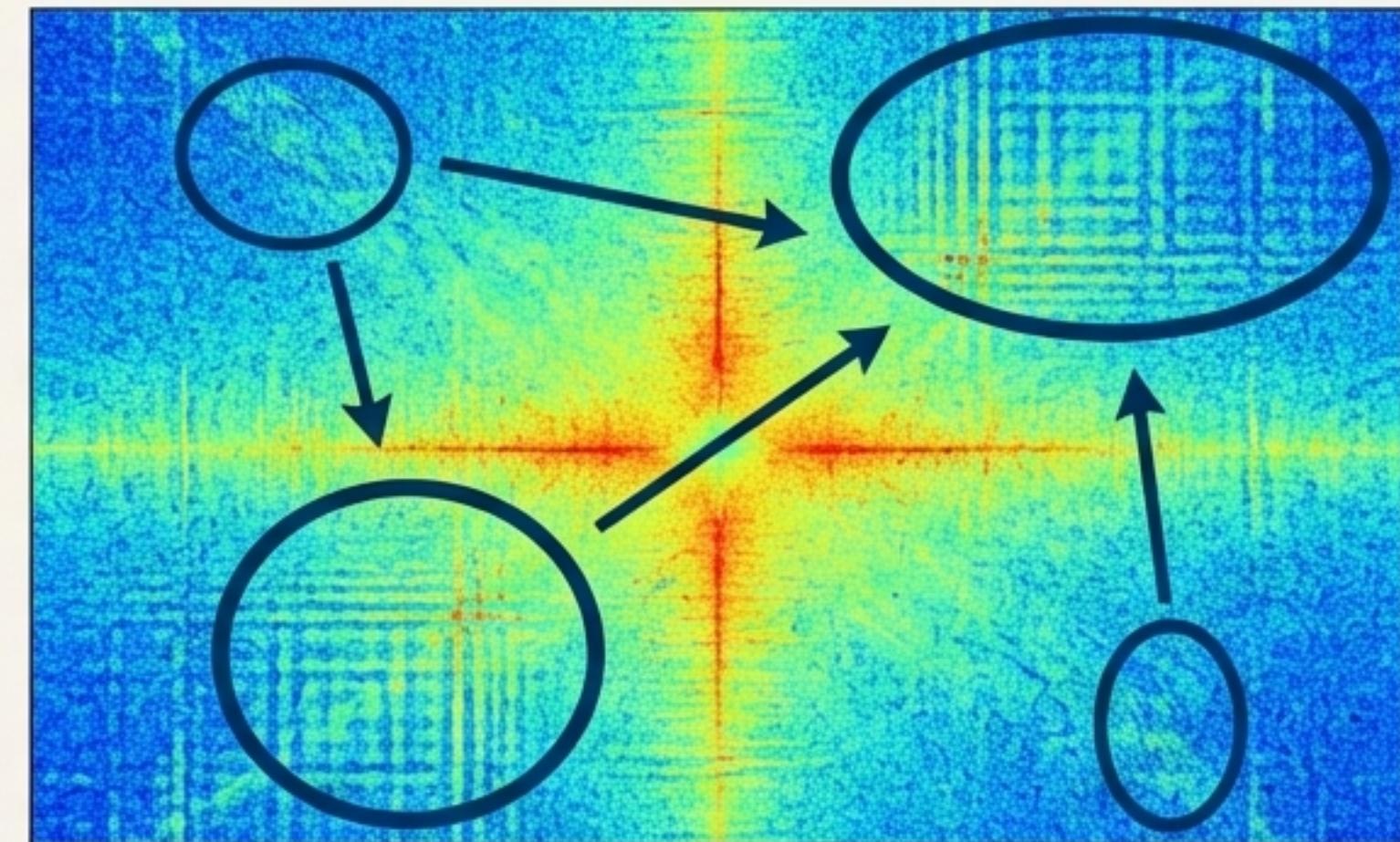
核心假設：深度神經網路能學習到人眼忽略的、偽造影像中的微觀不一致性。

方法：使用 CNN、Vision Transformer 等模型進行端對端的圖像分類。

鑑識方法一：藏在頻率中的幽靈指紋



真實影像頻譜 (Real Image Spectrum)



偽造影像頻譜 (Forged Image Spectrum)

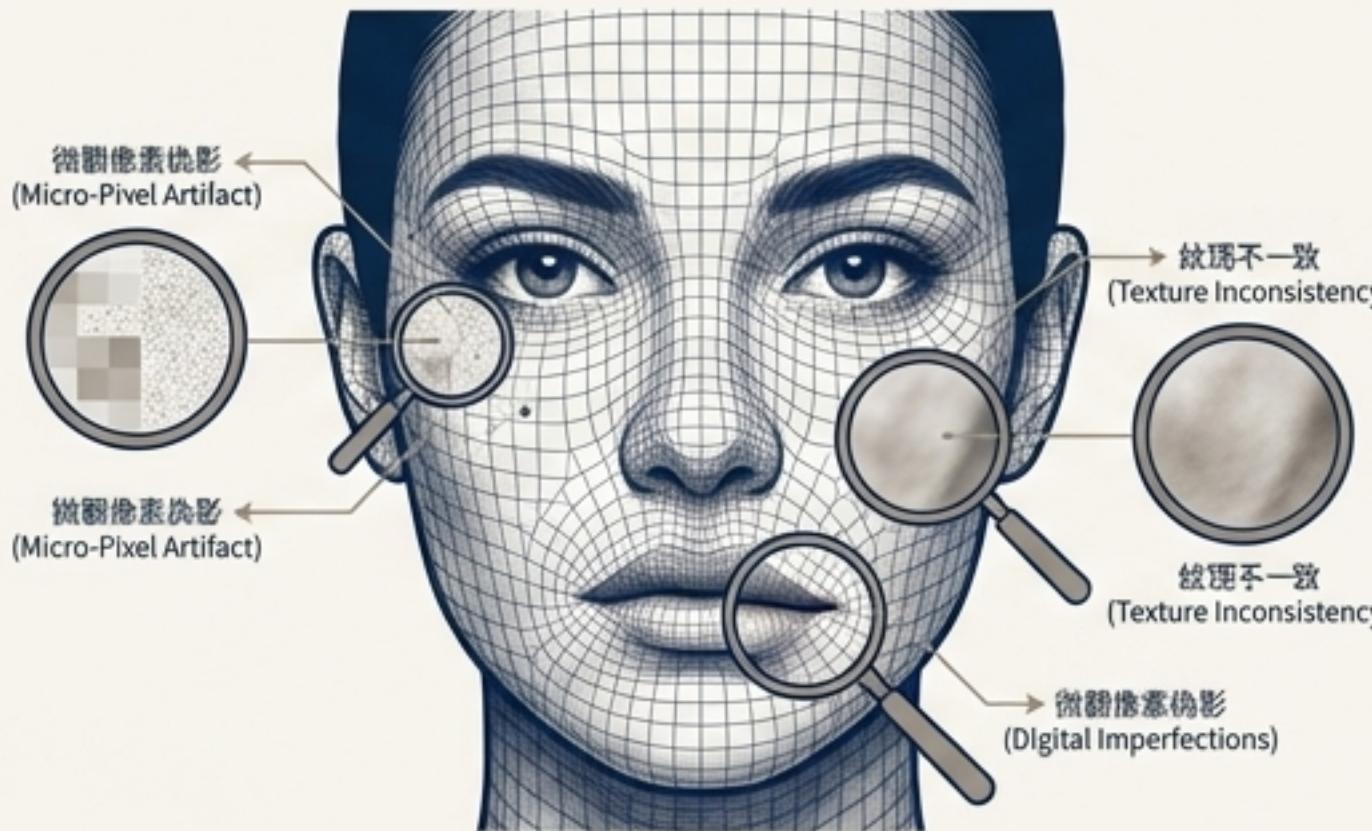
核心原理

生成模型（如 GAN）在處理高頻訊號時通常不夠自然，導致偽造影像的頻譜與真實影像存在統計性差異。

關鍵優勢

- **穩定性 (Stability)**：即使影像經過壓縮、縮放或濾鏡處理，頻率分佈中的異常特徵依然存在。
- **抗干擾性 (Robustness)**：將影像轉換至頻域後再進行卷積，能減少資料增強與壓縮造成的干擾，提升模型泛化能力。

鑑識方法二：能辨識微觀破綻的 AI 獵犬



核心原理 (Core Principle)

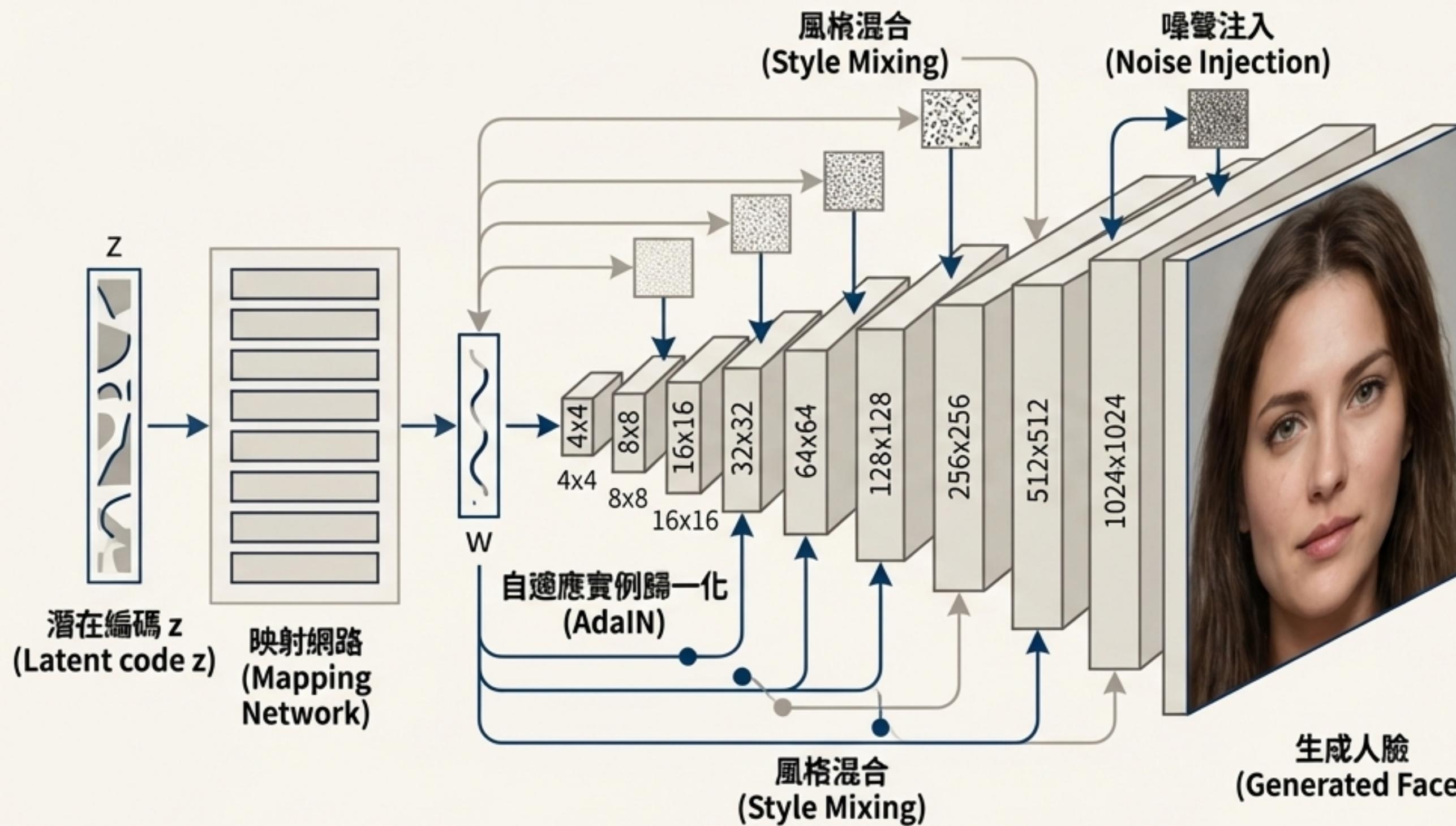
運用深度卷積神經網路 (CNN) 及 Vision Transformer (ViT) 架構，自動學習臉部區域的細微破綻與全域關聯性。

- **偵測目標：**眼睛閃爍頻率異常、嘴型不一致、像素級偽影等。

關鍵架構 (Key Architectures)

- **XceptionNet**：透過深度可分離卷積，有效提取臉部的細微紋理特徵。
- **EfficientNet**：自動平衡網路的寬度與深度，在效能與效率間取得最佳化。
- **Vision Transformer (ViT)**：捕捉臉部不同區塊間的長距離依賴關係，展現強大的泛化能力。

案例研究：解構偽造者的工具 (StyleGAN)



核心洞察 (Core Insight)：

StyleGAN 重新設計了生成器架構，實現了對臉部屬性（如姿態、身份）和隨機細節（如雀斑、髮絲）的精準、分層控制。

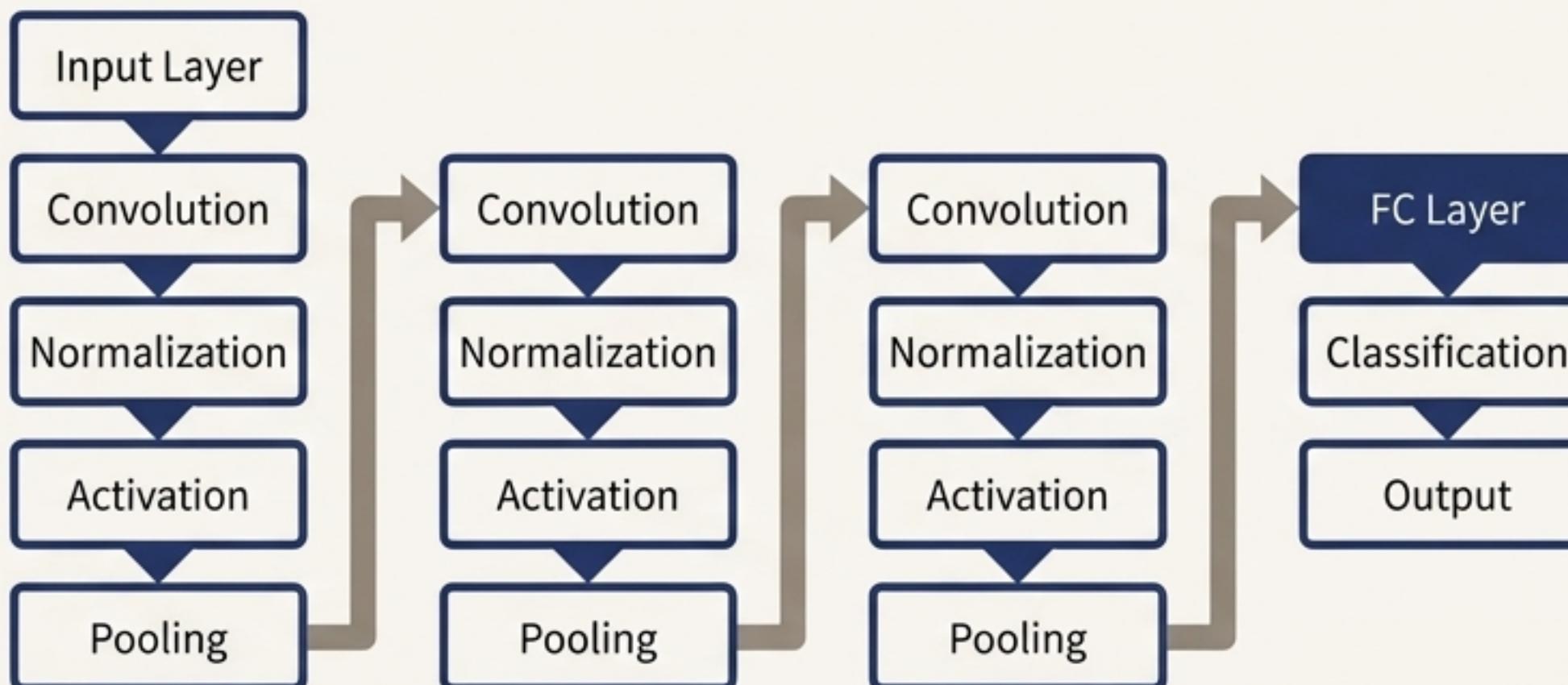
結論 (Conclusion)：

StyleGAN 的分層控制能力是其生成高品質 Deepfake 的關鍵。

案例 案例研究：追求效率與速度 (LightFFDNets)

核心目標 (Core Objective)

開發輕量級的 CNN 模型，在保持高準確率的同時，實現快速、高效的臉部偽造偵測。



模型架構 (Model Architecture)

- LightFFDNet v1：
2 個卷積層 + 1 個全連接層（極簡設計）
- LightFFDNet v2：
5 個卷積層 + 1 個全連接層（在準確性與速度間取得平衡）

LightFFDNets 的卓越表現

準確性指標 (Accuracy Metrics)
在 Fake-Vs-Real-Faces (Hard) 資料集上

1.0

F1 Score

完美

1.0

Recall

1.0

Precision

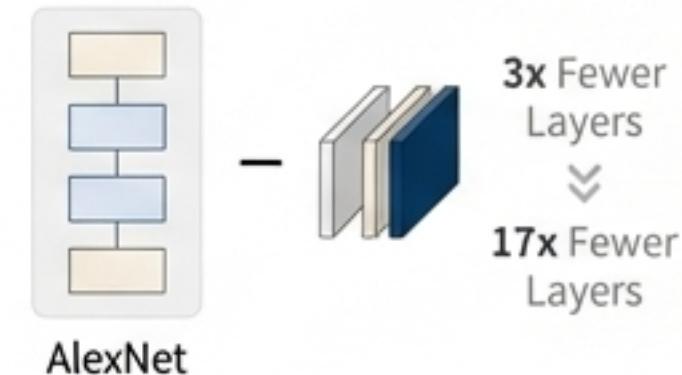
99.74%

驗證準確率
(Validation Accuracy)
(v2)

計算複雜度優勢 (Computational Complexity Advantage)

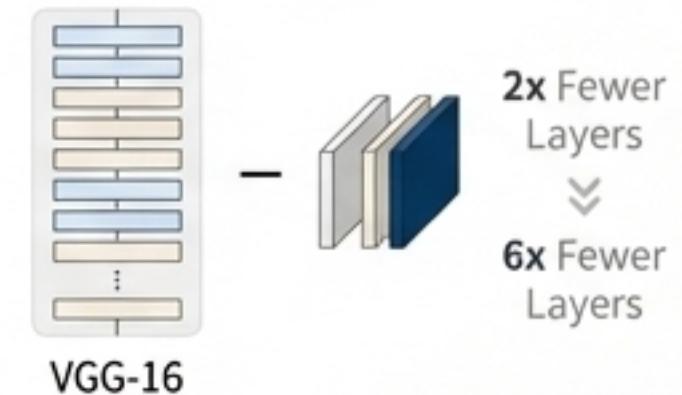
LightFFDNet v1

層數比 AlexNet 少 3 倍，
比 ResNet-50 少 17 倍。



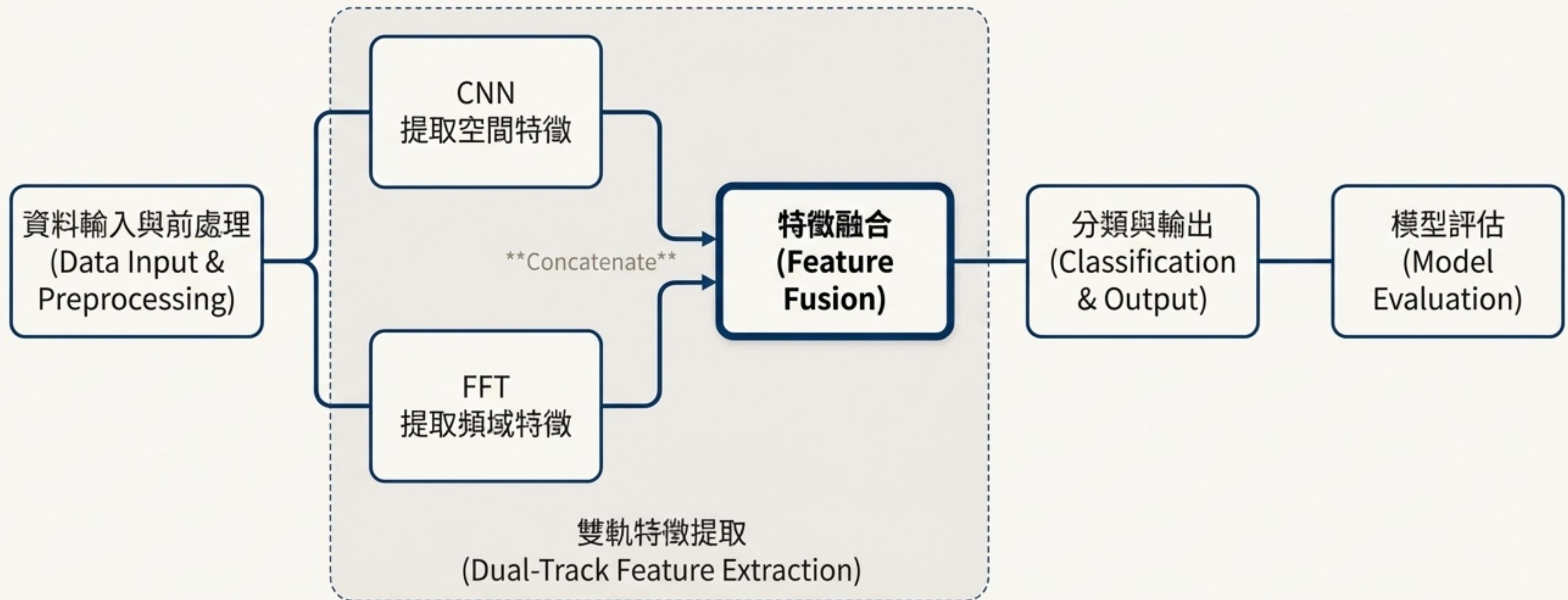
LightFFDNet v2

層數比 VGG-16 少 2 倍，
比 ResNet-50 少 6 倍。



結論 (Conclusion)：證明了輕量級模型在特定任務上可以達到甚至超越複雜模型的性能。

我們的藍圖：融合空間與頻譜特徵的偵測系統



步驟 1 & 2：數據基礎與雙軌特徵提取

Step 1. 資料蒐集與前處理 (Data Collection & Preprocessing)

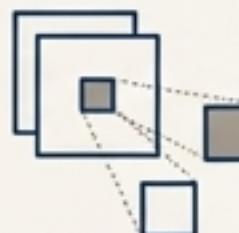
- **資料集** (Dataset) : "Real vs Fake Faces - 10k (來自 Kaggle)"

- **前處理** (Preprocessing) :

"統一影像尺寸至 224×224" (Unified image size to 224x224),

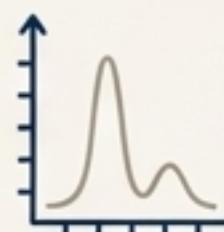
"影像像素值正規化" (Image pixel value normalization)

Step 2. 特徵提取 (Feature Extraction)



空間特徵 (Spatial Features)

使用卷積神經網路 (CNN) 自動學習臉部紋理、形狀與偽影等視覺特徵。

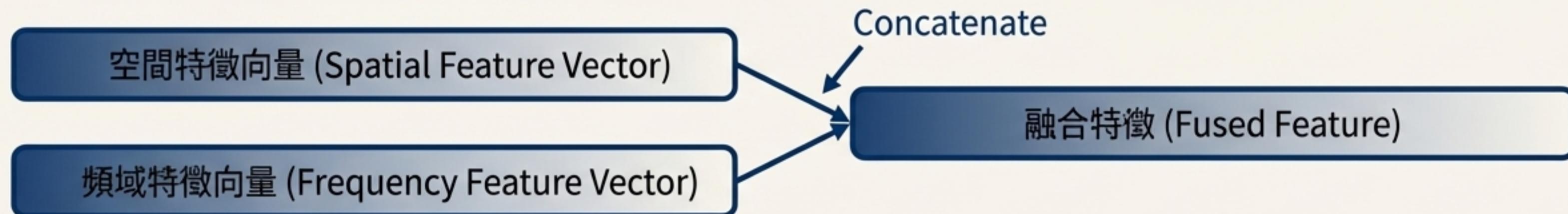


頻域特徵 (Frequency Features)

透過快速傅立葉轉換 (FFT) 分析 AI 生成時留下的週期性紋理與能量分佈異常。

步驟 3 & 4：融合核心與最終判決

Step 3. 特徵融合 (Feature Fusion)



核心優勢：融合後的特徵能同時捕捉到 **視覺上的不一致性**（空間特徵）與 **生成過程的底層痕跡**（頻域特徵），建立更全面的偽造證據鏈。

Step 4. 分類與輸出 (Classification & Output)



證明其價值：我們如何衡量成功

我們將透過一套全面的評估指標來驗證模型的效能與可靠性。

關鍵評估指標 (Key Evaluation Metrics) :

- **Accuracy (準確率)** : 整體分類正確的比例。
- **Precision / Recall / F1-score** : 評估模型在正負樣本上的查準率、查全率與綜合表現。
- **ROC-AUC (接收者操作特徵曲線下面積)** : 衡量模型在不同閾值下的整體辨識能力，是評估分類器優劣的黃金標準。

「透過嚴謹的方法論與客觀的評估，我們旨在為數位信任的保衛戰，提供一個更強大、更可靠的工具。」