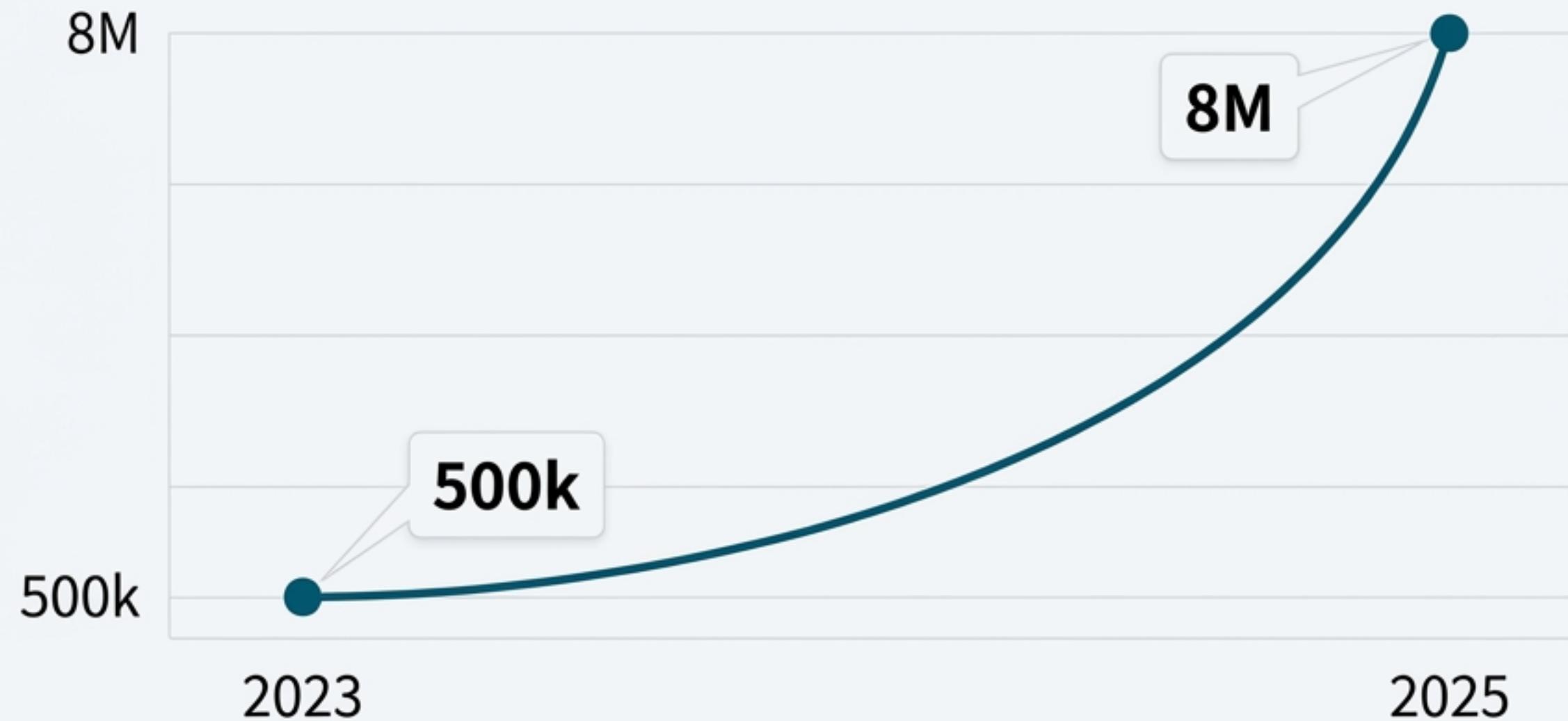


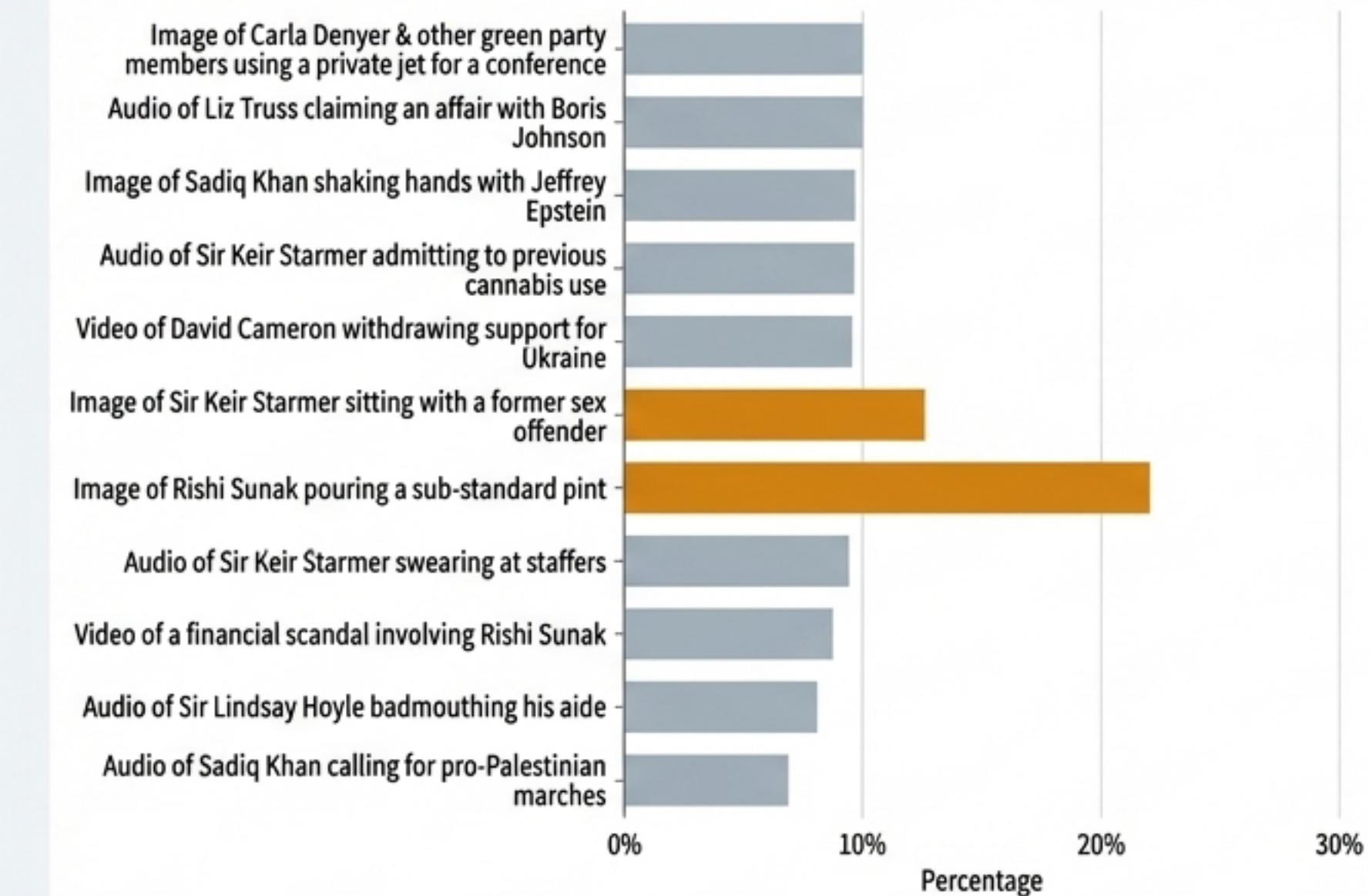
Deepfake 的指數級增長：一場信任危機



隨著生成式 AI 技術普及，Deepfake 內容正以驚人速度倍增，從 2023 年的 50 萬件預計在 2025 年激增至 800 萬件，對社會信任構成直接威脅。

從政治操控到個人詐騙：Deepfake 的具體危害

- **假新聞與政治操控：**利用偽造影像和音訊影響選舉、抹黑政治人物。
- **隱私侵犯與惡意影片：**將名人或普通人臉部移植到不當內容中。
- **網路詐騙與偽造證據：**用於金融詐騙、製造虛假法律證據。



英國研究顯示，大量民眾已接觸過用於政治目的的 Deepfake 內容，凸顯其在現實世界中的影響力。

偵測的雙重挑戰：技術進化與信任侵蝕



技術挑戰 (Technical Challenges)

- **技術持續進化**：Deepfake 生成模型不斷迭代，現有的偵測模型容易過時失效。
- **偽造類型多樣**：挑戰不僅限於換臉 (Face Swap)，還包括聲音偽造 (Voice Swap)、頭部重現 (Head Reenactment) 等。



社會挑戰 (Societal Challenges)

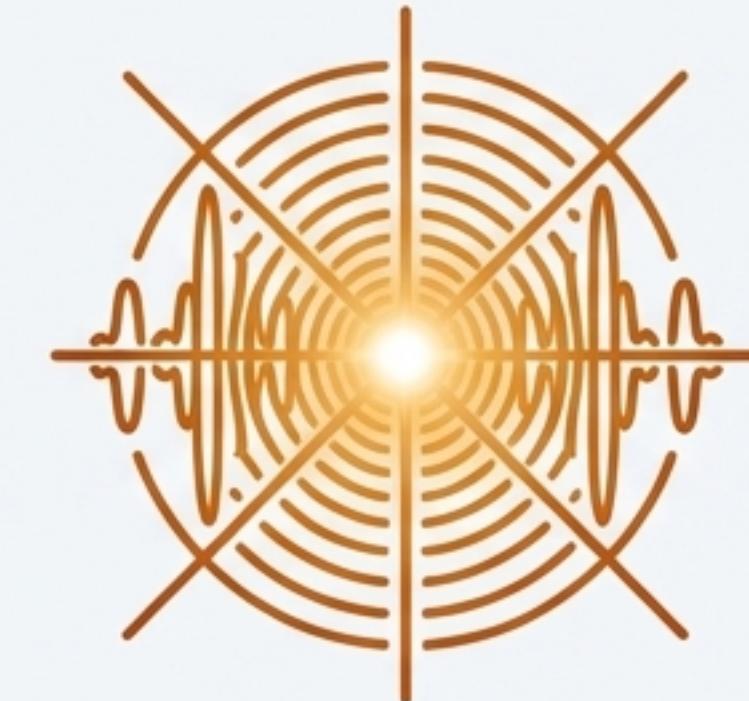
- **反真實效應 (Liar's Dividend)**：Deepfake 的普及讓大眾開始懷疑真實影像的真實性，騙子可以輕易地將真實證據斥為偽造。
- **重建信任的門檻**：偵測模型必須達到極高的準確率，才能有效重建公眾對數位內容的信任。

辨識偽造的兩大戰場：空間域與頻率域



空間域 (Spatial Domain)

肉眼可見的世界。分析影像的像素、紋理、光影和細微的視覺不一致性，例如不自然的眼睛閃爍或臉部邊緣瑕疵。



頻率域 (Frequency Domain)

隱藏在像素背後的世界。透過傅立葉轉換 (FFT) 分析影像的頻譜特徵，揭露生成模型在處理高頻訊號時留下的獨特『指紋』。

CNN

Vision Transformer

像素分析

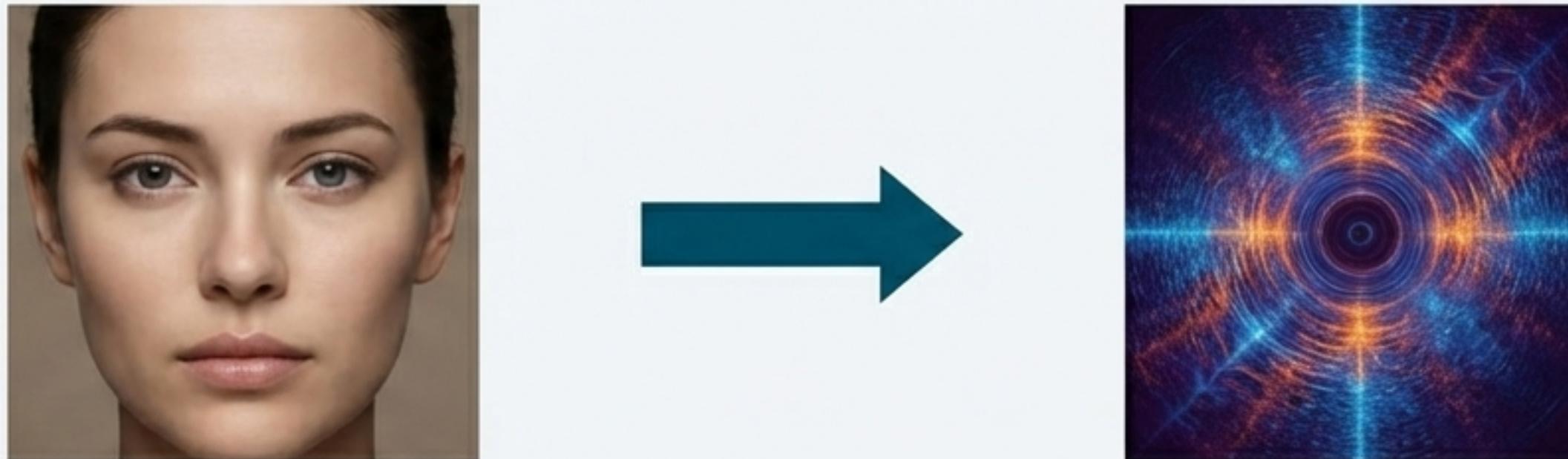
FFT

頻譜

週期性紋理

頻率域 分式分析：揭示肉眼不可見的生成痕跡

雖然 Deepfake 影像在視覺上幾可亂真，但在頻率分佈 (Fourier Spectrum) 上卻與真實影像存在顯著差異。這是因為生成模型（如 GAN）在複製高頻細節時往往不夠自然。



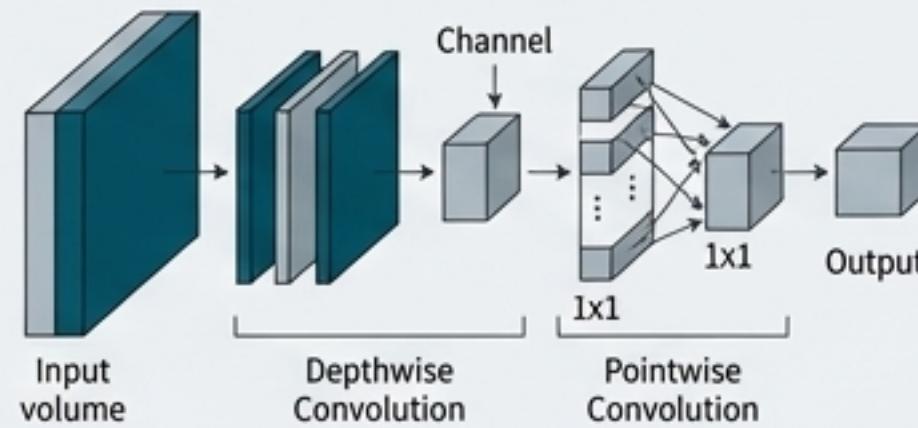
- **揭露統計性差異**：生成影像的高頻分佈與真實影像存在根本差異，可透過快速傅立葉轉換 (FFT) 進行量化辨識。
- **提升泛化能力**：在頻域進行卷積操作，能有效減少壓縮、縮放等干擾，讓模型更具通用性。
- **穩定性高**：即使影像經過後製處理，換臉操作留下的頻率異常特徵依然存在，比單純的像素分析更為可靠。

深度學習模型：捕捉像素級的微小破綻

卷積神經網路 (CNN) 與視覺轉換器 (Vision Transformer) 等先進架構，能自動學習並識別臉部區域的微小不協調之處，例如不自然的眼睛閃爍頻率或嘴型變化。

XceptionNet

其深層可分離卷積結構，能高效提取臉部細微特徵。



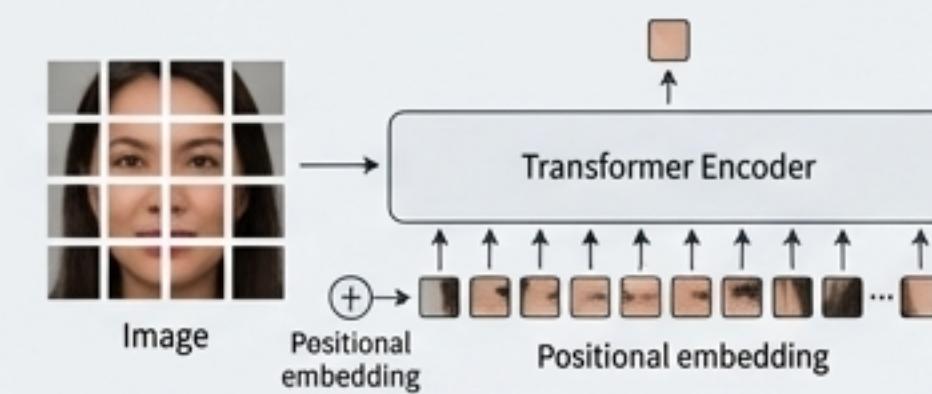
EfficientNet

透過自動化模型縮放，在維持高準確率的同時，大幅減少參數數量。



Vision Transformer (ViT)

將影像視為序列處理，能捕捉臉部各區域間的全局關聯性，在跨資料集的測試中展現出強大的泛化能力。



案例研究一：知己知彼 — 深入 StyleGAN 生成架構

*A Style-Based Generator Architecture for Generative Adversarial Networks (Tero Karras, et al., 2018)

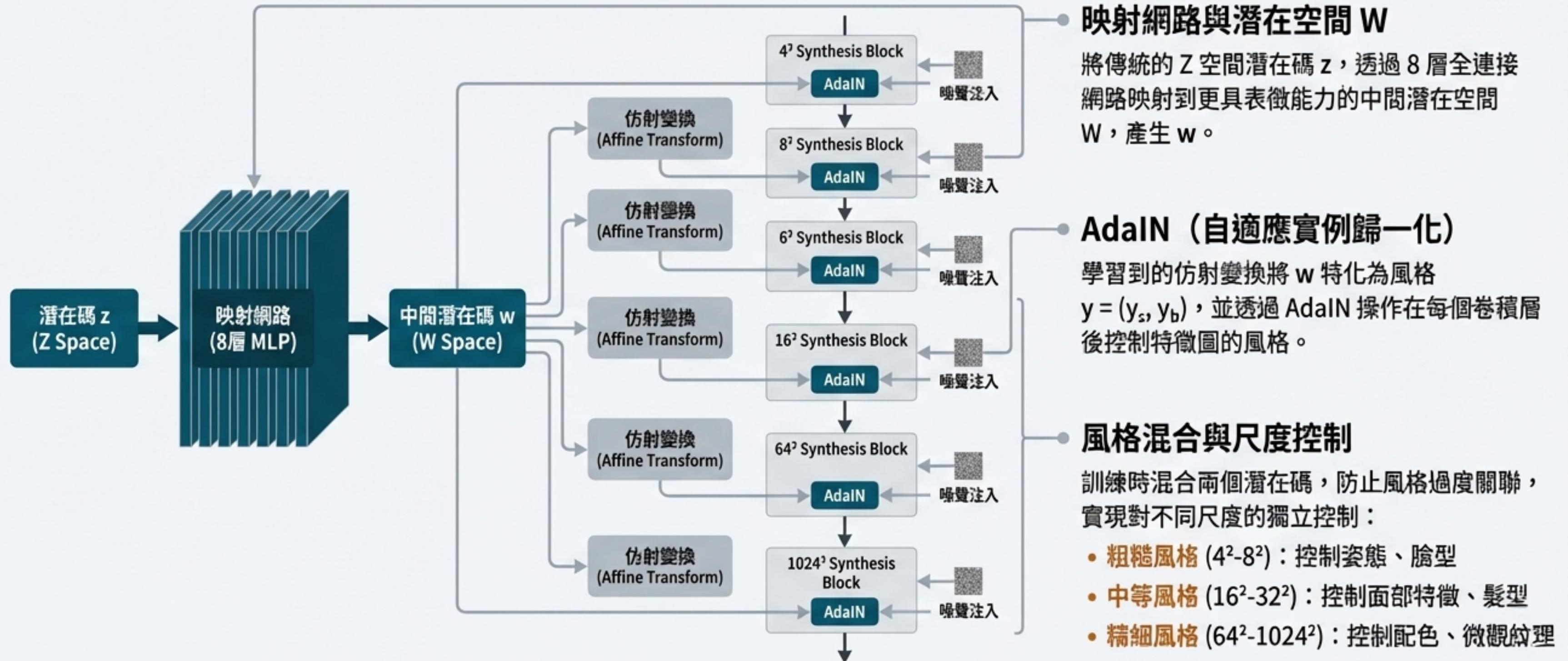
StyleGAN 重新設計了生成器架構，借鑒風格遷移的概念，實現了對生成影像高層次屬性（如姿態、身份）和隨機細節（如雀斑、髮絲）的無監督分離與精準控制。

很定創新之的副大簡介：

1. 映射網路 (Mapping Network) 與中間潛在空間 W
2. 自適應實例歸一化 (AdaIN)
3. 噪聲注入 (Noise Injection)
4. 風格混合 (Style Mixing)



StyleGAN 的核心機制：從映射網路到尺度控制

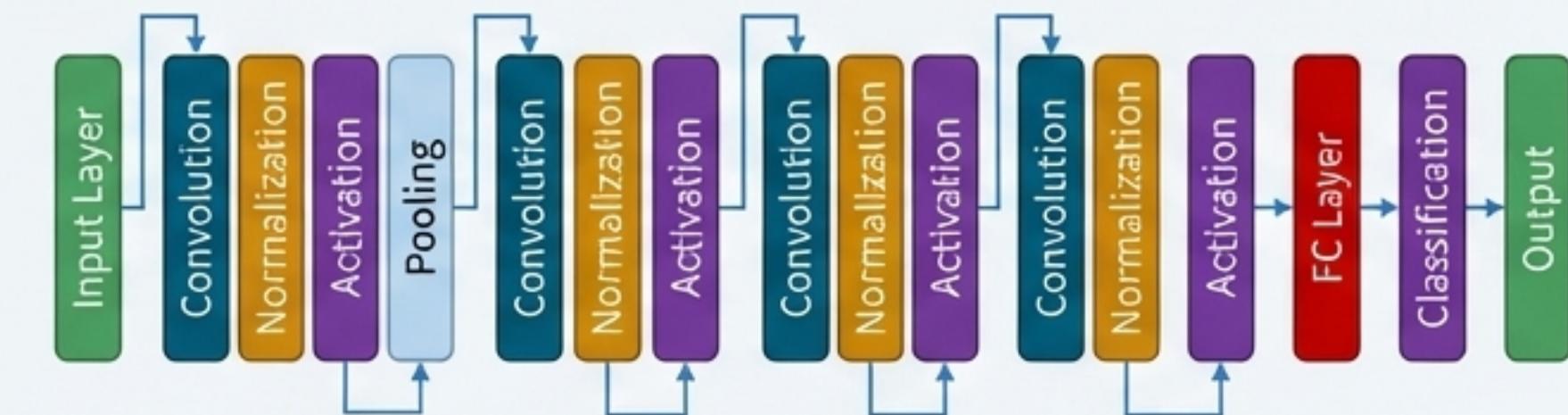


案例研究二：追求效率 — 輕量級 CNN 偵測模型

LightFFDNets: Lightweight Convolutional Neural Networks for Rapid Facial Forgery Detection
(Günel Jabbarlı, Murat Kurt, 2024)

此研究專注於開發輕量級的 CNN 模型，證明在不犧牲過多準確率的前提下，可以實現極高的計算效率，達成快速、準確的人臉偽造偵測。

- **LightFFDNet v1**：極簡設計，僅 2 個卷積層，追求最快速度。
- **LightFFDNet v2**：包含 5 個卷積層，1.5 個卷積層，在準確性與速度間取得更佳平衡。



LightFFDNets 的卓越表現：準確性與計算效率的平衡

準確性表現 (Accuracy Performance)

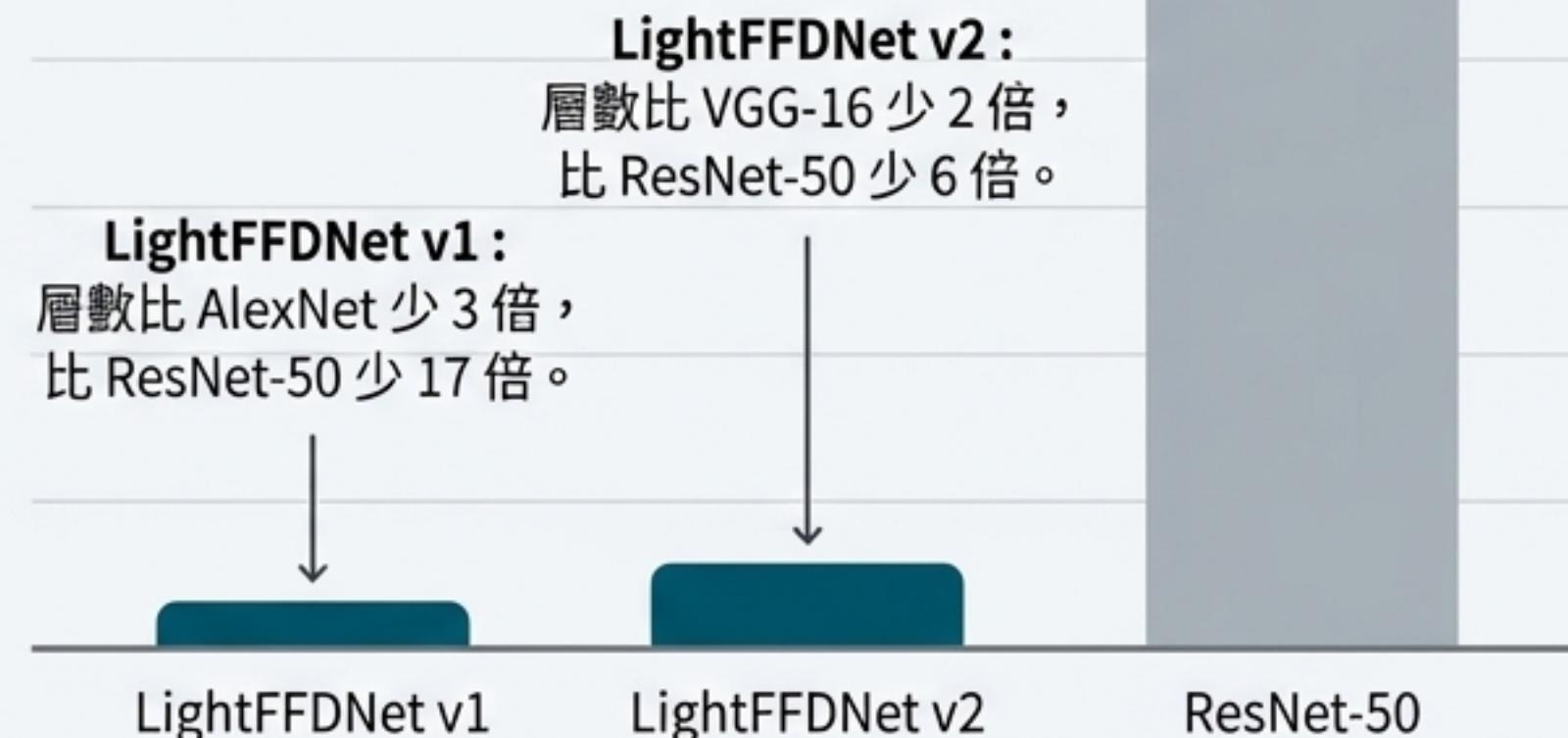
Dataset: Fake-Vs-Real-Faces (Hard)

v1: **99.48%** v2: **99.74%**

Test Metrics

- **F1 Score:** 1.0
- **Recall:** 1.0
- **Precision:** 1.0
- **Test Accuracy:** 最高達 99.87%

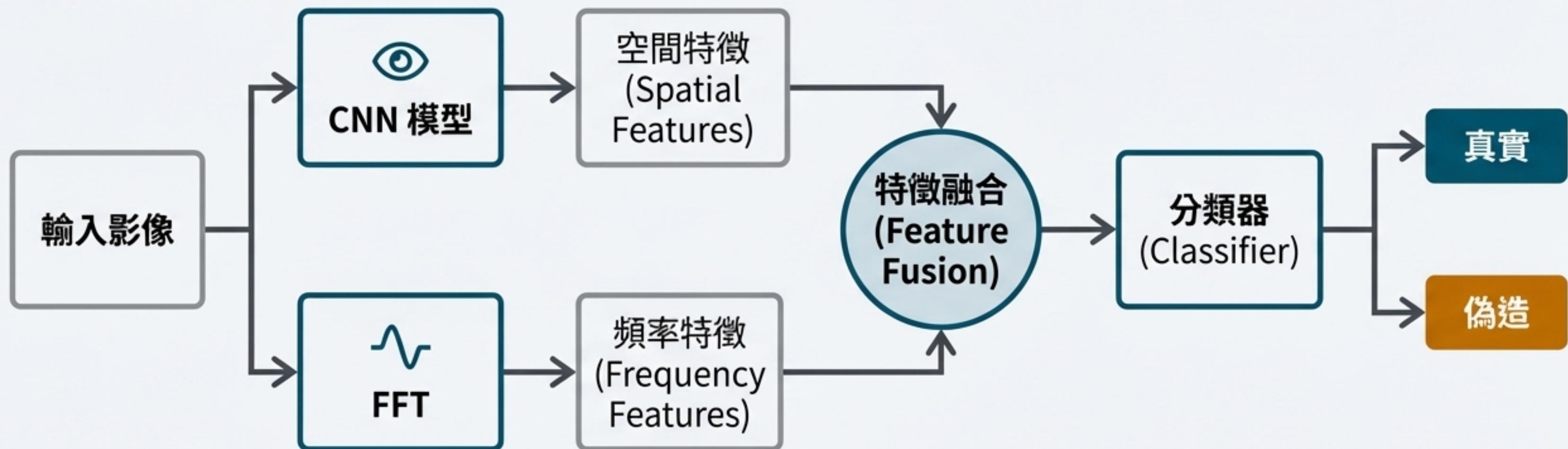
計算複雜度 (Computational Complexity)



結果證明，特製的輕量級模型能夠在特定任務上，以極低的計算成本達到與大型預訓練模型相媲美甚至超越的性能。

融合的力量：為何不結合空間與頻率特徵？

單獨依賴空間域或頻率域分析都有其局限性。空間模型可能被壓縮干擾，而頻率模型可能忽略細微的視覺瑕疵。真正的強大防線，在於整合兩者的優勢。



我們的設計核心是將 CNN 提取的空間特徵與 FFT 提取的頻域特徵進行融合，創造一個能同時感知視覺紋理與底層生成痕跡的混合特徵表示。

我們的設計流程：一個五步驟的整合偵測框架

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

建立包含真實與偽造影像的標準化資料集。

1

2

3

4

5

特徵提取 (Feature Extraction)

並行提取影像的空間域與頻率域特徵。

特徵融合 (Feature Fusion)

將兩種特徵向量拼接成一個強大的混合表示。

分類與輸出 (Classification & Output)

使用分類模型對融合後的特徵進行判斷。

評估指標 (Evaluation Metrics)

以多維度指標嚴謹評估模型的性能。

分類與輸出 (Classification & Output)

使用分類模型對融合後的特徵進行判斷。

評估指標 (Evaluation Metrics)

以多維度指標嚴謹評估模型的性能。

步驟 1-2：從資料準備到雙軌特徵提取

資料蒐集與前處理 (Step 1)

資料集 (Dataset): Real vs Fake Faces - 10k

前處理步驟 (Preprocessing)

- 統一影像尺寸 (e.g., 224×224)
- 正規化處理 (Normalization)

特徵提取 (Step 2)

👁️ 空間特徵 (Spatial Features)

方法: 使用預訓練的卷積神經網路 (CNN) 作為主幹

目的: 捕捉臉部細節、紋理不一致性與邊緣瑕疵

⚡ 頻域特徵 (Frequency Features)

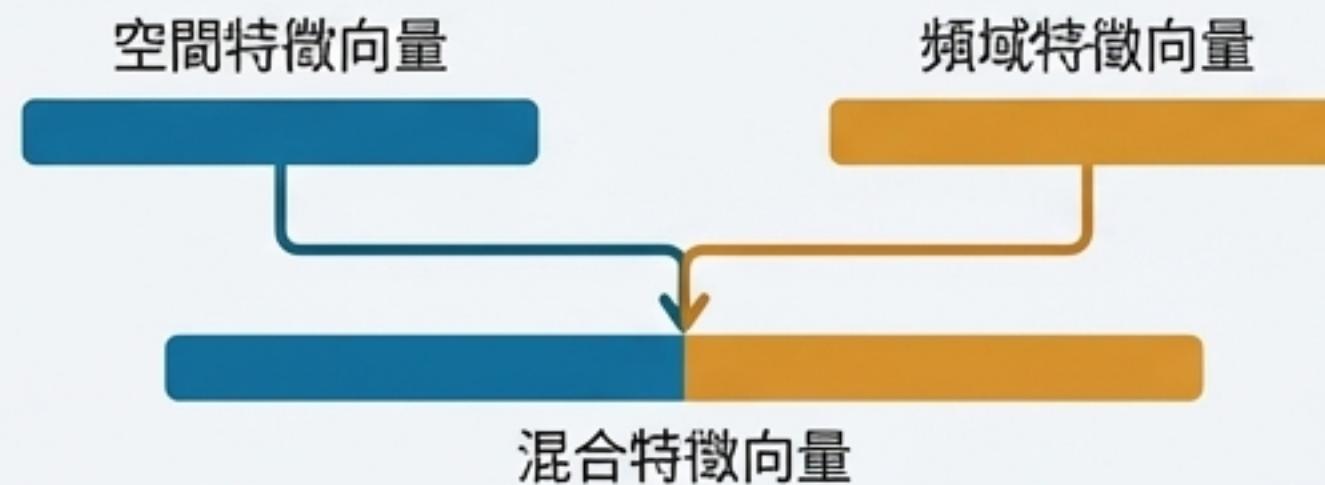
方法: 對影像進行快速傅立葉轉換 (FFT)

目的: 分析 AI 生成過程中留下的週期性紋理與能量分佈異常

步驟 3-4：特徵融合與最終分類

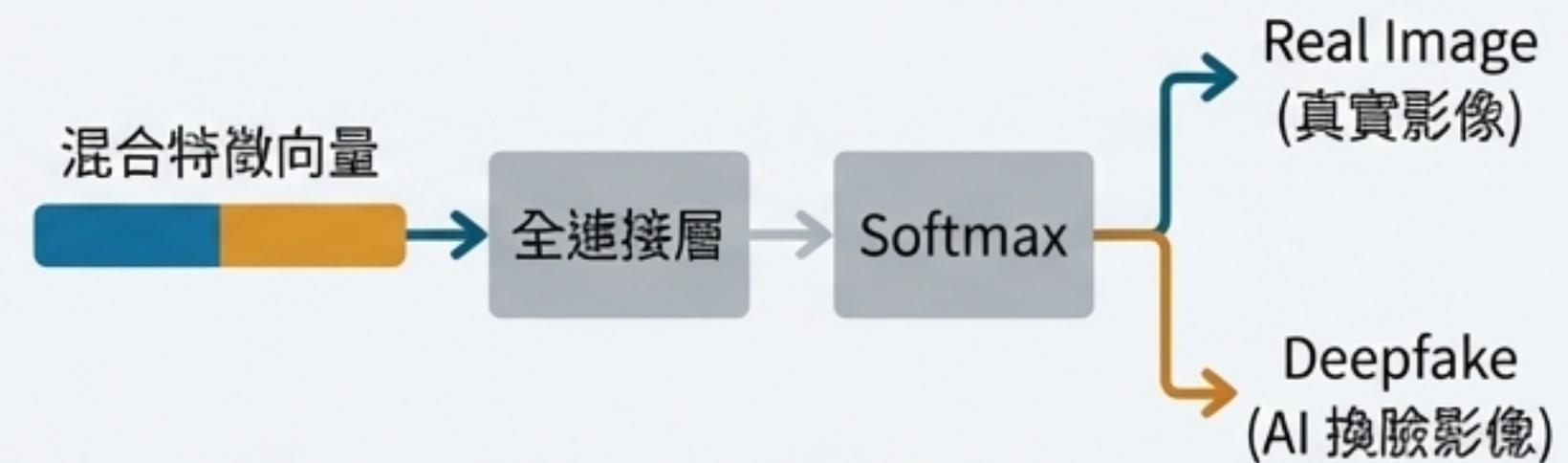
特徵融合 (Step 3)

方法 (Method)：拼接 (Concatenation)



分類與輸出 (Step 4)

分類器 (Classifier)：全連接層 (Fully Connected Layer)



優勢 (Advantage)：

融合後的特徵能同時反映影像的視覺紋理與底層的生成痕跡，提供比單一特徵更全面的判斷依據。

輸出結果 (Outputs)：

- Real Image (真實影像)
- Deepfake (AI 換臉影像)

步驟 5：以嚴謹指標評估效能，建立可信賴的防線

一個偵測模型的價值最終取決於其性能。我們採用一套完整的評估指標，全面檢驗模型的準確性、可靠性與泛化能力。

評估指標 (Evaluation Metrics)



Accuracy (準確率)

整體正確分類的比例。



Precision (精確率)

在所有被預測為 Deepfake 的樣本中，真正是 Deepfake 的比例。



Recall (召回率)

在所有真正的 Deepfake 樣本中，被成功偵測出的比例。



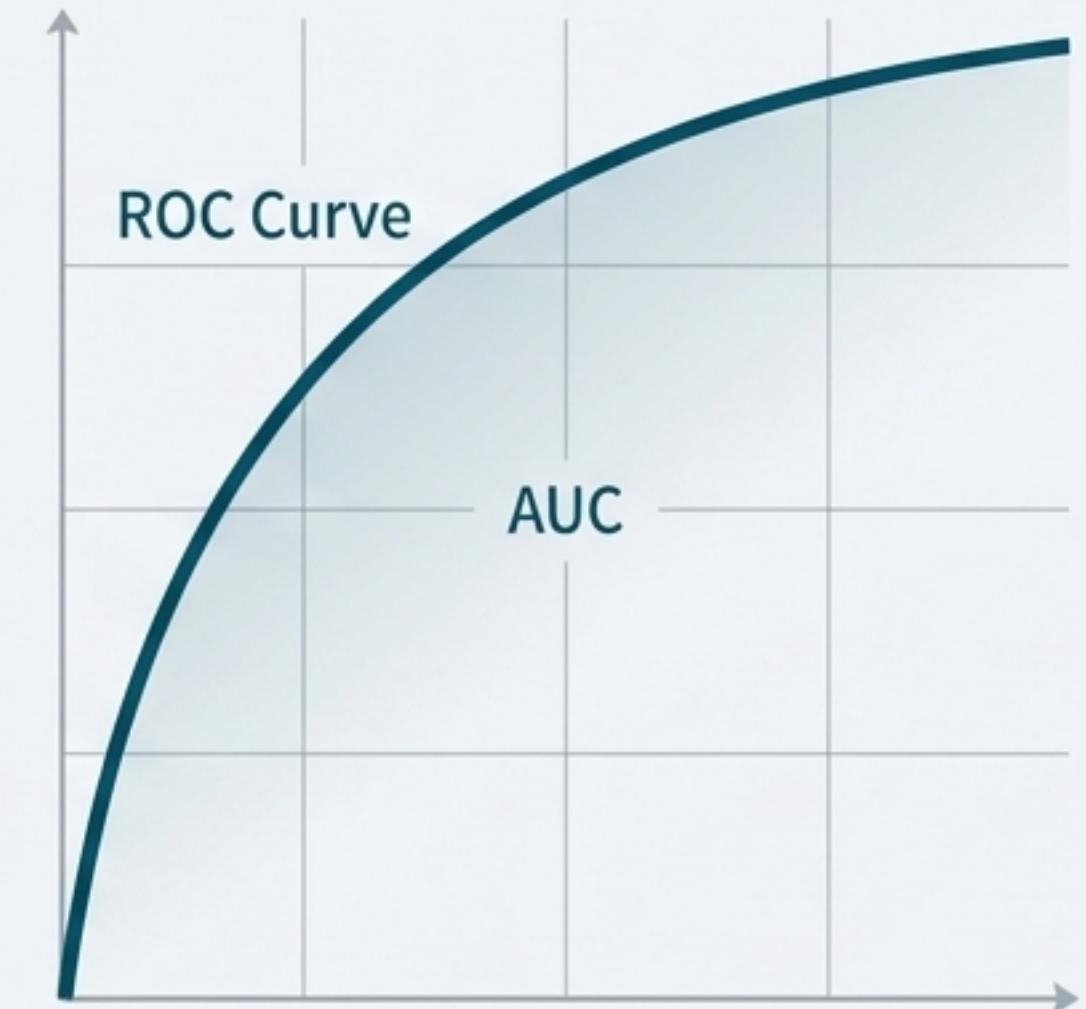
F1-Score

Precision 與 Recall 的調和平均數，綜合評估模型的穩健性。



ROC-AUC

衡量模型在不同閾值下的整體辨識能力，是評估分類器性能的關鍵指標。



唯有透過嚴謹且透明的評估，我們才能建立真正有效的 Deepfake 防線，為重建數位世界的信任貢獻一份力量。