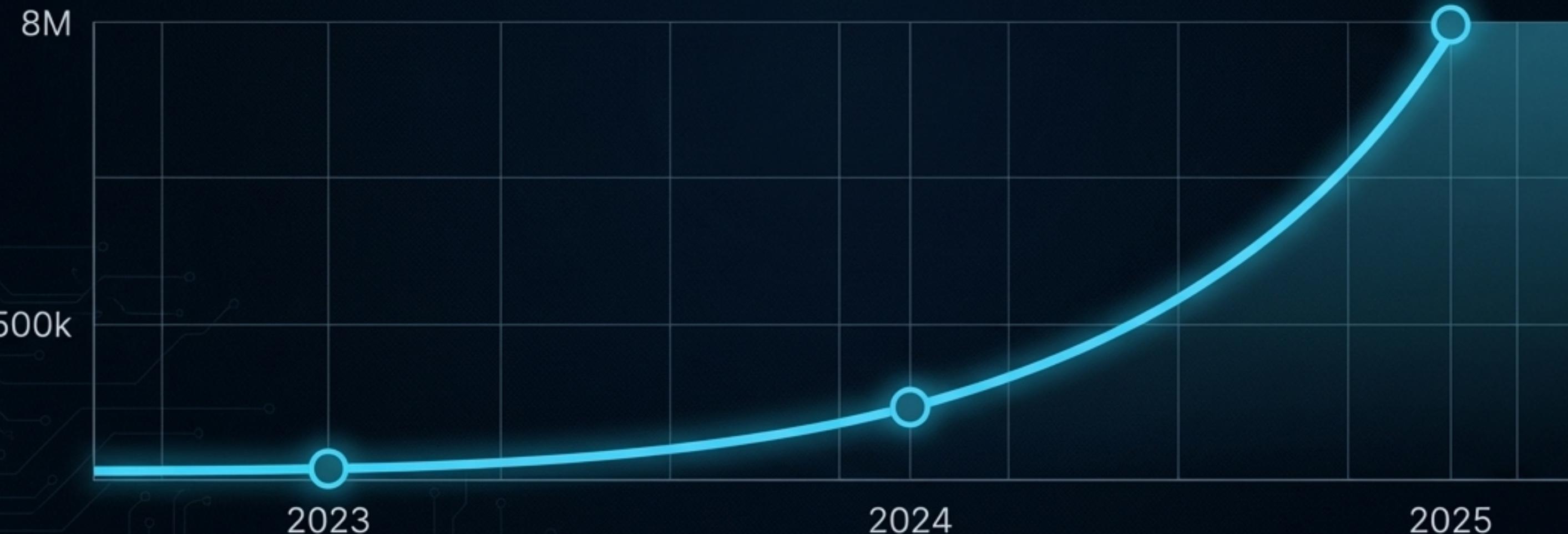


數位偽造的指數級增長

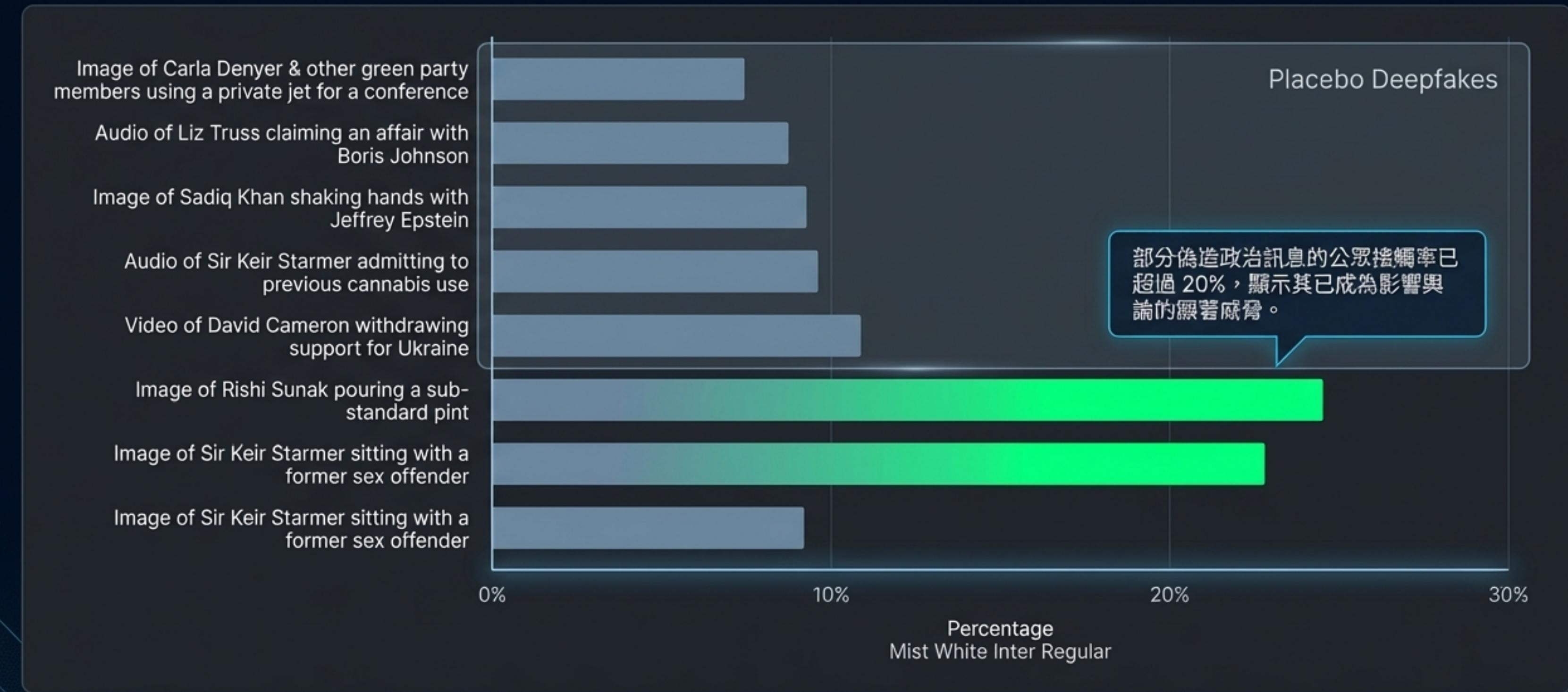
8,000,000

預計至 2025 年，Deepfake 檔案數量將從 2023 年的 50 萬激增至 800 萬，呈現失控的擴散趨勢。



侵蝕信任：偽造資訊的政治武器化

Deepfake 技術被濫用於製造假新聞、操縱政治話語，嚴重破壞媒體公信力與社會信任基礎。



核心挑戰：在「反真實效應」中重建信任



技術持續進化：生成模型不斷迭代，今日的偵測模型可能在明日失效。



信任度危機：Deepfake 普及引發「反真實效應 (Liar's Dividend)」，即使是真實影像也可能被質疑，侵蝕公眾信任。



高精準要求：偵測模型必須達到極高準確率，才能有效反擊並重建資訊生態的可靠性。

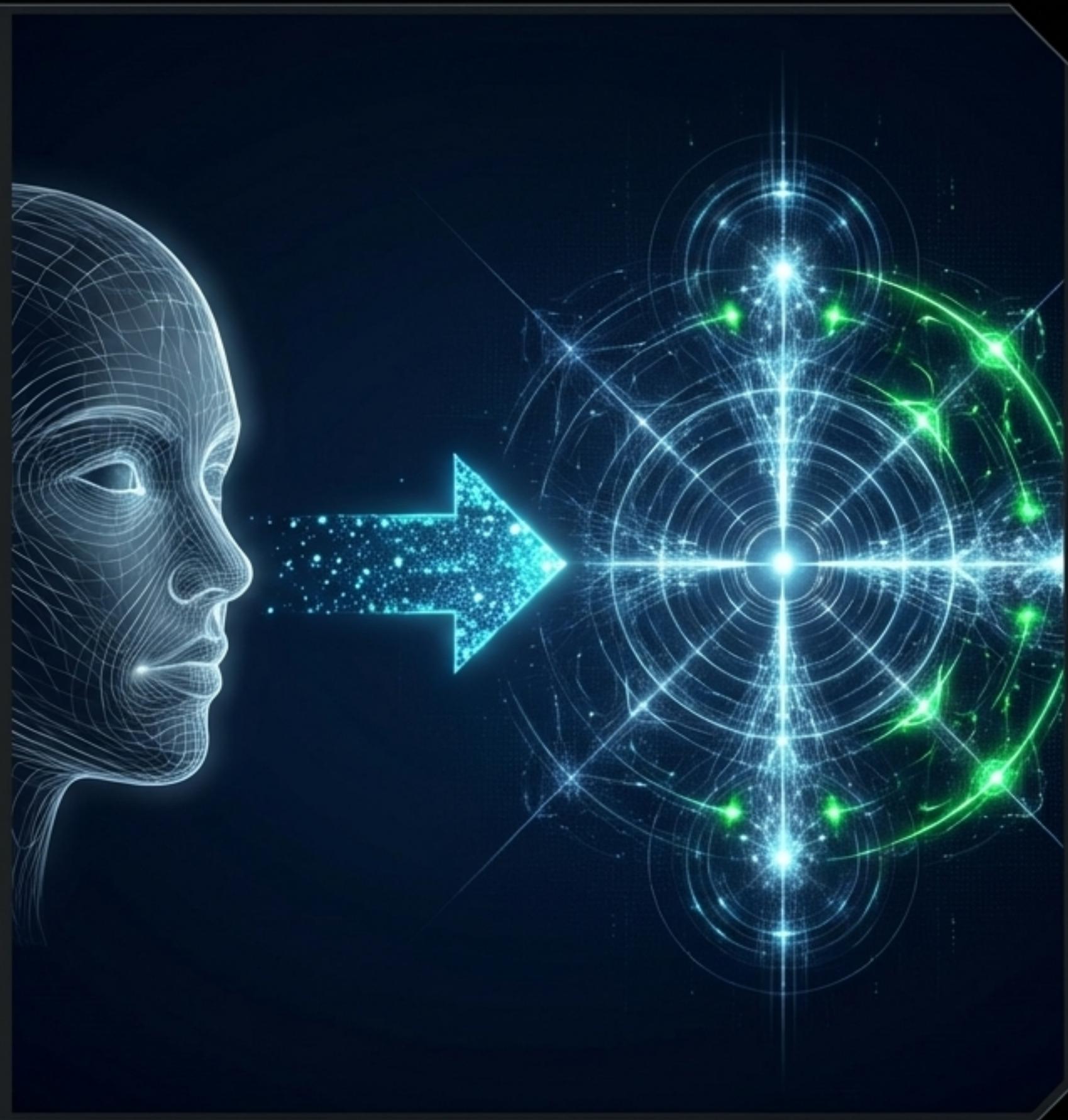
01 // 現有軍火庫

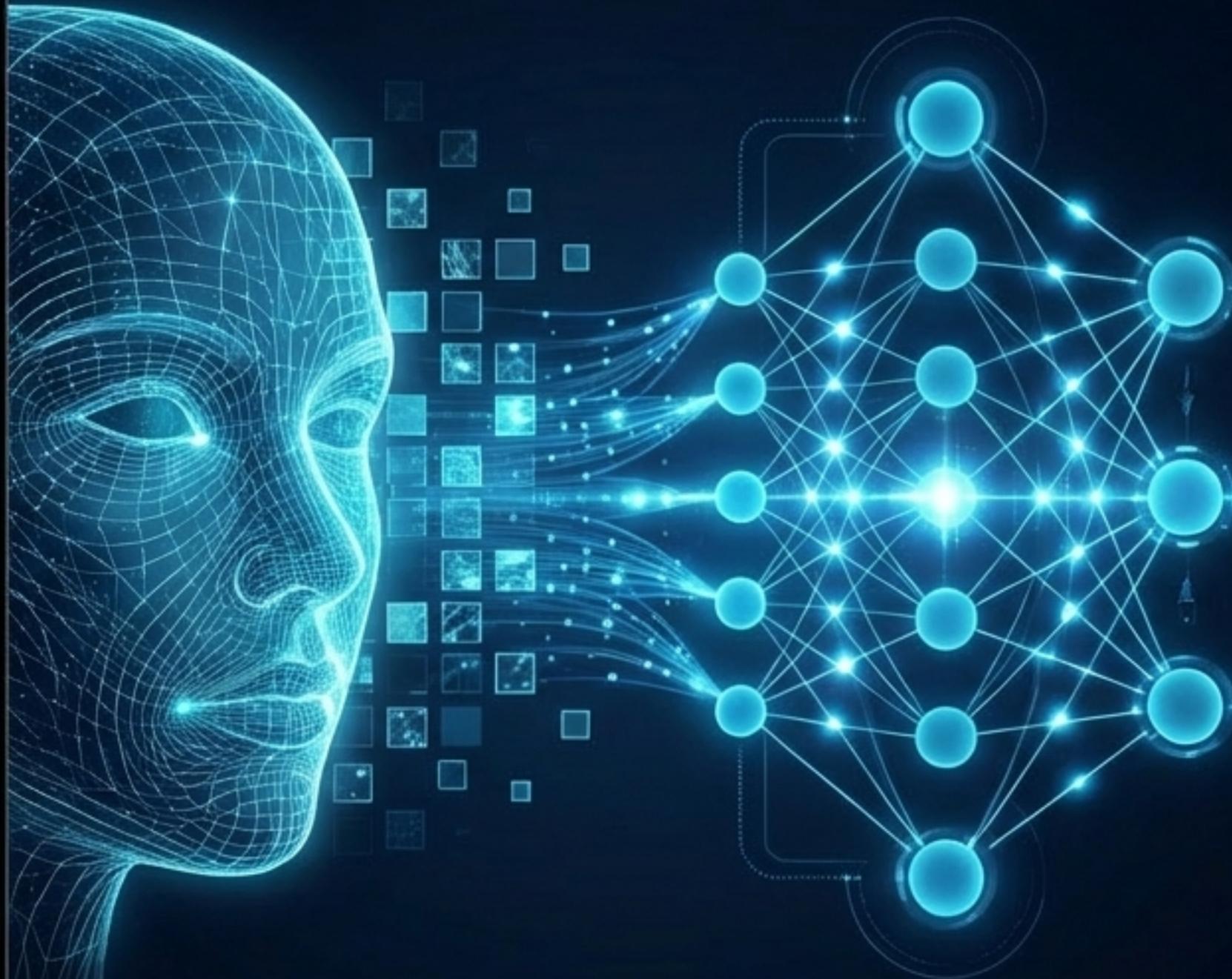
當前防禦範式分析 (ANALYSIS OF CURRENT DEFENSE PARADIGMS)

方法一：頻譜指紋分析

儘管 Deepfake 在視覺上極為逼真，其生成過程在頻域 (Frequency Spectrum) 中會留下非自然的統計性差異。

透過快速傅立葉轉換 (FFT)，我們能偵測這些「生成痕跡」。此特徵對影像壓縮、縮放或濾鏡處理具有高度穩定性，比單純的像素分析更具韌性。





方法二：神經模式辨識

運用卷積神經網路 (CNN) 與 Vision Transformer (ViT) 架構，模型能自動學習人臉中微小的空間不一致性。

- **XceptionNet**
透過深度可分離convolution，有效提取臉部細微特徵。
- **Vision Transformer (ViT)**
捕捉臉部區域間的全域關聯性，展現強大的泛化能力。

防禦間隙：單一方法的侷限性



優勢

對抗壓縮與干擾。

Inter Medium

侷限

可能忽略複雜的空間偽造

線索。

Inter Regular

如何結合兩者優勢，
建立一個更全面的
防禦系統？



優勢

精準捕捉細微視覺異常。

Mist Medium

侷限

對未見過的生成技術泛化能

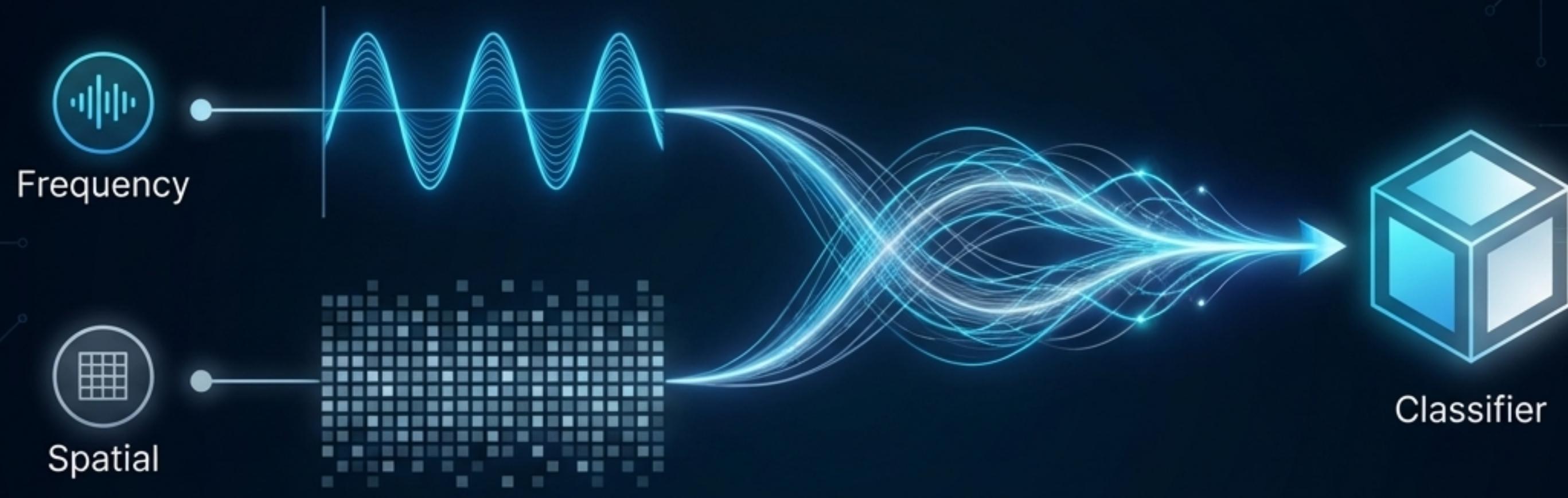
力可能受限。

Inter Regular

02 // 融合式智能

我們的設計藍圖 (OUR DESIGN BLUEPRINT)

核心論點：融合頻域與空間特徵



Thesis Statement
Inter Medium

我們提出一個混合式偵測模型，不依賴單一特徵，而是將頻域的「生成痕跡」與 CNN 提取的「空間紋理」進行特徵融合。

Benefit
Inter Regular

這種融合表示 (Hybrid Representation) 能同時捕捉視覺偽造的細微破綻與生成模型的底層統計異常，大幅提升偵測的準確性與穩健性。

步驟 01：資料集與前處理

Dataset



Mist White
Real vs Fake Faces - 10k

Silver-Gray
包含真實影像與 Deepfake 影像的
平衡資料集。

Preprocessing



影像尺寸統一

所有影像縮放至 224×224 像素。

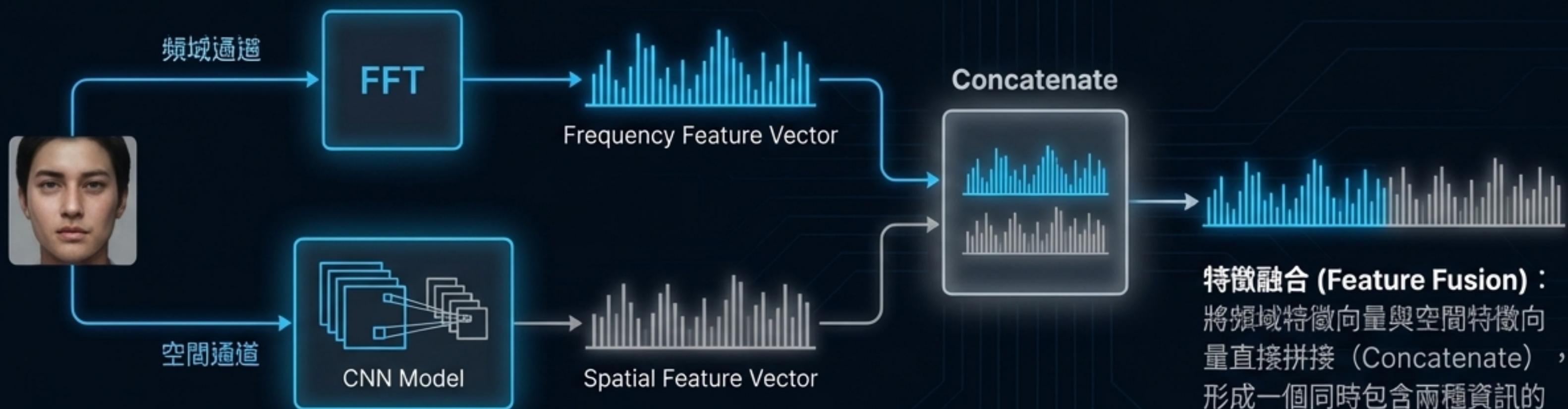


正規化處理

標準化像素值以利模型穩定訓練。

步驟 02-03：雙通道特徵提取與融合

頻域特徵 (Frequency Features)：透過 FFT 轉換，分析 AI 生成時遺留的週期性紋理與能量分佈異常。



特徵融合 (Feature Fusion)：
將頻域特徵向量與空間特徵向量直接拼接 (Concatenate)，形成一個同時包含兩種資訊的強大混合特徵。

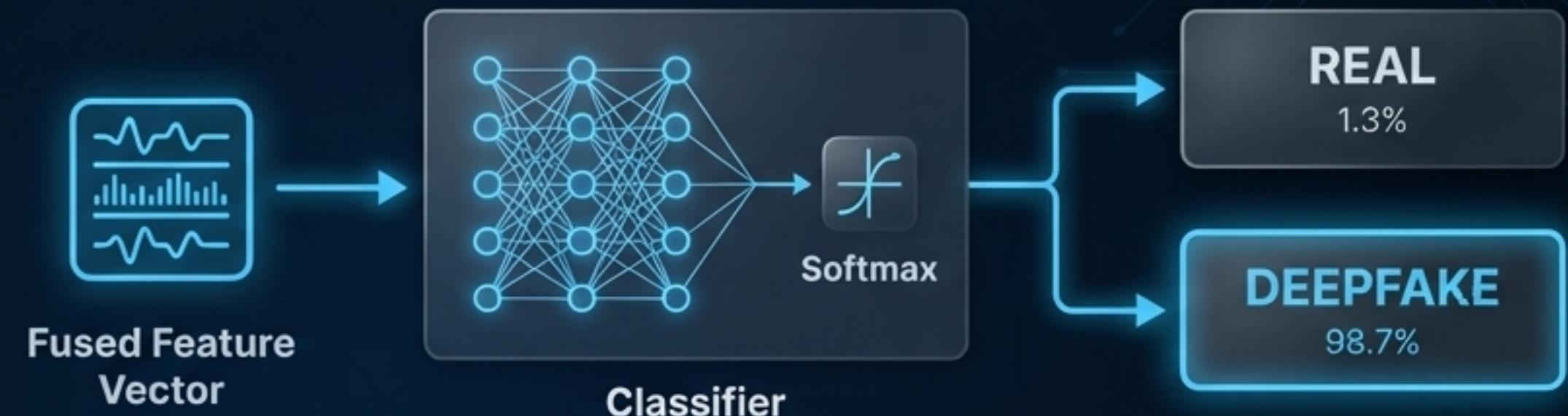
空間特徵 (Spatial Features)：運用預訓練的 CNN 模型，提取臉部的高階語義特徵，如紋理、光影與幾何不一致性。

步驟 04-05：分類輸出與效能評估

分類器 (Classification)

將融合後的特徵輸入至全連接層 (Fully Connected Layer)。

使用 Softmax 函數輸出最終分類機率：
「真實影像」或「Deepfake」。



評估指標 (Evaluation Metrics)



Accuracy

準確率



Precision / Recall / F1-Score

精確率 / 召回率 / F1 分數



ROC-AUC

整體辨識能力

展望：建立更具韌性的數位信任生態

透過融合頻譜與空間智能，我們不僅僅是在 偵測偽造，更是在
為一個可驗證、可信賴的資訊未來，建立演算法層級的防禦工事。