deepfake

313831007 傅靖茹

一、背景

隨著人工智慧與深度學習技術的迅速發展，影像與影片的生成技術變得越來越先進。Deepfake 技術，利用深度神經網路合成或篡改人臉與語音，已廣泛應用於娛樂、廣告、甚至政治領域。然而，這項技術也帶來了嚴重的資訊安全與社會信任危機，例如偽造名人言論、製作虛假新聞，甚至可能被用於詐騙與勒索。

傳統的 deepfake 偵測方法多數依賴於特定偽造手法的特徵，導致模型在遇到新型或未見過的 deepfake 類型時，準確率大幅下降。為了提升模型的泛化能力，學界與業界開始關注於更具通用性的特徵學習與跨模態技術。Vision-Language Foundation Models, VLMs 近年來展現出強大的語意理解與跨領域遷移能力。這類模型能夠同時處理圖像和文字資訊，並學習到更高層次的語意關聯，為 deepfake 偵測任務帶來新的突破契機。因此，如何有效運用 VLMs 於 deepfake 偵測，並驗證其在跨類型偽造場景下的泛化表現，成為當前值得深入探討的研究方向。

結合 CLIP 與 LoRA 於 Deepfake 偵測任務之動機

CLIP（Contrastive Language-Image Pretraining）作為大型視覺-語言基礎模型，具備強大的語意理解與跨模態特徵提取能力，已被證實在 deepfake 偵測任務中展現優異的泛化表現，特別是對未見過的偽造手法亦能維持穩定準確度[1]–[3]。然而，CLIP 的原始嵌入空間並非專為 deepfake 分類而設計，若直接應用於偵測任務中，可能難以捕捉偽造內容的微觀差異。

此外，在現實應用場景中，deepfake 樣本數量有限，獲得大規模標註資料頗具挑戰。若採用傳統的全參數微調，不僅訓練成本高，且在小樣本條件下極易出現過擬合現象，導致泛化能力下降 [2]。

Low-Rank Adaptation（LoRA）是一種參數高效的微調技術，透過引入低秩矩陣進行適應性調整，僅需訓練極少量新增參數，便能顯著提升任務適應性，並降低過擬合風險。近期研究指出，結合 LoRA 的 CLIP 模型，即使僅使用約 50 對真實與偽造影像進行微調，也能在多種 deepfake 生成器上達成優於傳統方法的準確率，並展現更佳的跨方法泛化能力 [4]，LoRA 大幅降低訓練所需的參數量與計算成本，適用於資源受限或需快速部署的場景 。

二、實驗方法

Dataset

本研究採用 FaceForensics++（FF++）資料集，該資料集為當前 Deepfake 偵測領域中最具代表性的基準數據集之一，涵蓋多種由先進人臉偽造技術生成之真實與偽造影片，具備高度多樣性與挑戰性。

為嚴謹評估模型在面對未知偽造手法時的泛化能力，採用 cross-manipulation 的資料切分策略，具體如下表所示：

	Real_youtube	FaceSwap	NeuralTextures	Frame Count
訓練集	70%	90%	-	7070 | 9090 | 0
驗證集	20%	10%	-	2020 | 1010 | 0
測試集	10%	-	100%	1010 | 0 | 10100

上述配置中，訓練與驗證資料僅包含 Real_youtube 與 FaceSwap 類型，NeuralTextures 類型資料**僅於測試階段出現**，以模擬真實應用場景下模型需面對新型未知偽造的挑戰。測試集中特別保留 Real_youtube 類型的真實樣本，作為與未見過之偽造類型之對照組，幫助評估模型是否能在完全未知的偽造情境下，仍準確分辨真實與偽造樣本。

此外，Real_youtube真實影片樣本均勻分布於訓練、驗證與測試各階段，可提升模型對真實樣本多樣性的學習能力，進而有效降低誤判率並增強模型魯棒性。

Model

為驗證所提方法之有效性，我們設計兩種比較模型：

- **Baseline**：使用預訓練 CLIP 模型萃取圖像特徵，並在其基礎上訓練線性分類器。
- **CLIP + LoRA**：在 CLIP 模型基礎上引入 LoRA（Low-Rank Adaptation）進行參數高效微調。

Baseline 模型代表無需額外調整大模型權重的低資源設定，其表現可作為 CLIP 特徵在 Deepfake 偵測任務上的基準參考。而 CLIP + LoRA 則旨在透過極小化參數更新量，提升模型對任務的適應性與泛化能力，特別是在面對未知偽造技術時的表現。

兩者的比較將可全面評估參數高效微調策略在模型效能與應用潛力上的影響，為後續研究提供設計依據。

Ablation Study

為進一步探討 LoRA 組態對模型效能之影響，我們設計一系列消融實驗，針對以下四個關鍵變因進行系統性分析：

- **秩（Rank）**：選用 rank ∈ {8, 16, 32} 等不同設定，觀察低秩表示能力對表現之影響。
- **縮放因子（Alpha）**：搭配不同 rank 測試alpha ∈ {16, 32,64 }，評估權重放大對微調效果的調節作用。
- **Dropout 機率**：設定 0.0、0.05、0.1 三種情況，以分析正則化強度與過擬合控制的關聯。
- **訓練資料比例**：調整可用訓練資料比例為 `20%`、`50%`、`100%`，模擬低資源場景下模型之泛化能力。

在每組 ablation 中，其餘超參數均保持一致，僅調整單一變因，以確保變異因素控制與比較公平性。

三、實驗結果

Ablation study

在本次 ablation study 中，共設計了 27 組不同 LoRA 參數組合，為了更清楚呈現模型效能的變化，特別挑選數組具代表性的結果進行說明。這些組合涵蓋了最佳表現、不同 dropout 設定，以及參數量差異等情境，有助於全面理解各參數對模型表現的影響。

組合	AUC	EER	F1	Accuracy	rank	alpha	dropout
最佳表現	0.886	0.185	0.9394	0.8909	8	64	0.05
最大參數量	0.836	0.195	0.8984	0.8273	32	64	0.05
最小參數量	0.839	0.180	0.9388	0.8909	8	32	0.05
Dropout=0	0.872	0.190	0.9490	0.9091	8	64	0.00
Dropout=0.05	0.886	0.185	0.9394	0.8909	8	64	0.05
Dropout=0.1	0.848	0.210	0.9293	0.8727	8	16	0.10
AUC最低	0.786	0.245	0.8229	0.7182	16	64	0.05
AUC中位數	0.833	0.190	0.8852	0.8091	32	64	0.00

分析結果顯示，最佳的參數組合為 rank=8、alpha=64、dropout=0.05，該組合在 AUC、F1、Accuracy 等指標上皆達到最高分（AUC 0.886，F1 0.9394，Accuracy 0.8909），同時僅需約 0.20M 的參數量，展現出優異的效能與模型效率。相較之下，將 dropout 設為 0 雖然 AUC 也可達 0.872，但略低於最佳組合，顯示適度的 dropout 有助於提升模型泛化能力。此外，參數量較大的組合（如 parameter_count=0.77M）在效能上未見明顯優勢，反而部分組合（如 AUC 0.836）表現不如最佳組合。AUC 最低的組合（0.786）則驗證了部分參數設定對模型有負面影響。

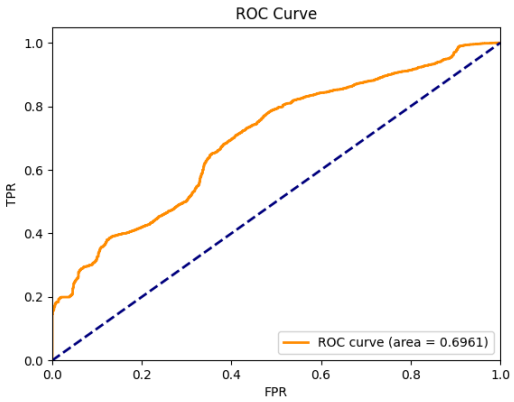
綜體來看，最佳的 LoRA 參數組合能在效能與模型規模間取得良好平衡，後續實驗將以此組合作為預設設定，進行進一步驗證與比較。

Performance

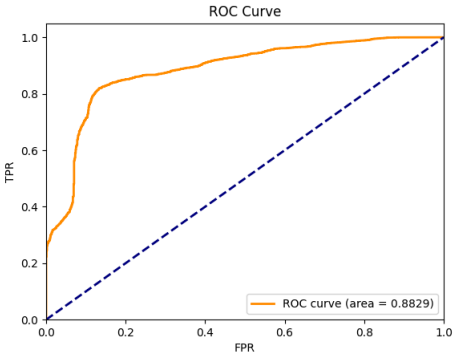
本節詳細說明在 testing dataset 上，針對 frame-level 與 video-level 兩種層級進行深偽檢測模型的效能評估。Frame-level 測試是將每一個影格作為獨立樣本，分別預測其真偽，適合檢視模型對於細微畫面變化的敏感度與辨識力。Video-level 測試則是針對同一部影片，將所有影格的預測分數進行加總並取平均，作為該影片的整體預測結果，這種方式更貼近實際應用場景，能反映模型對於整體語境與時序資訊的綜合判斷能力。

Frame-level Performance

model	ACC	F1	EER	AUC	Run time
frozen CLIP linear probe	0.8405	0.9112	0.3525	0.6961	~ 2 min
CLIP + LoRA	0.8898	0.9384	0.1628	0.8829	~ 2 min



frozen CLIP linear probe



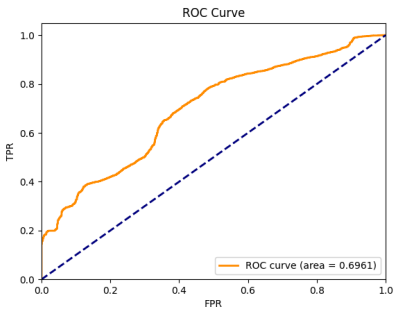
CLIP + LoRA

Video-level Performance

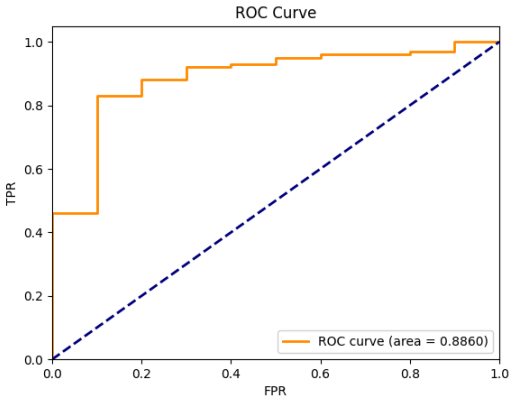
在本研究的實驗結果中，針對相同 CLIP + LoRA 模型於 frame-level 與 video-level 之 AUC 表現進行比較，發現 video-level 評估顯著優於 frame-level。具體而言，CLIP + LoRA 在 frame-level 的 AUC 為 0.8829，而於 video-level 則提升至 0.8860，顯示即便模型架構與權重一致，影片層級的整合策略仍能進一步增強深偽檢測效能。

此現象可歸因於 video-level 評估方法將同一部影片內所有影格的預測分數進行加總並取平均，有效平滑單一影格可能出現的誤判或極端值，進而提升模型在整體影片層級上的穩定性與準確度。

model	ACC	F1	EER	AUC	Run time
frozen CLIP linear probe	0.8182	0.898	0.38	0.703	~ 2 min
CLIP + LoRA	0.8909	0.9394	0.185	0.886	~ 2 min



frozen CLIP linear probe



CLIP + LoRA

根據測試結果，CLIP + LoRA 模型在 frame-level 及 video-level 的各項指標（Accuracy、F1、AUC）均明顯優於 frozen CLIP linear probe。以 frame-level 為例，CLIP + LoRA 的 AUC 達到 0.8829，明顯高於 frozen CLIP linear probe 的 0.6961，顯示其對個別影格的偽造特徵有更強的辨識能力。進一步觀察 video-level，CLIP + LoRA 的 AUC 亦達 0.8860，同樣顯著優於 frozen CLIP linear probe 的 0.7030，說明該模型在整體影片層級也能有效整合時序與空間資訊，提升對深偽影片的整體判斷準確度。此外，CLIP + LoRA 在 EER與 F1 分數上也展現出更佳的平衡與穩定性，顯示模型在提升準確率的同時，能有效降低誤判率並兼顧召回率。

四、實驗分析與討論

Identity Leakage

本研究發現，資料切分的粒度對深偽檢測模型的評估結果具有顯著影響。當採用 frame-level 切分時，同一部影片的影格會被分散至訓練、驗證與測試集，導致模型在測試階段仍可見到訓練時出現過的人物與背景資訊，進而產生 identity leakage 現象。這種情況下，模型可能僅記住特定人物的臉部特徵或場景風格，而非真正學會深偽圖像中的偽造特徵，導致評估指標如 AUC被嚴重高估，甚至出現接近 1.0 的不切實際分數，無法真實反映模型的泛化能力。

相對地，video-level 切分策略則確保每位人物僅出現在訓練、驗證或測試集其中之一，徹底杜絕人物重疊，使得模型在測試階段會遇到完全陌生的人物，無法再依賴訓練時學到的 identity cue 來判斷真假，雖然會導致AUC相對下降，但這種下降實際上反映了更貼近真實應用場景的模型泛化挑戰，亦即模型必須具備辨識從未見過的新人物是否被偽造的能力。

綜合上述結果，建議深偽檢測模型的訓練與評估過程中，必須明確說明並嚴格控制資料切分方式。未經控制的 frame-level 切分會造成評估指標失真，誤導模型泛化能力的判斷；唯有採用 video-level 切分，才能避免 identity leakage，並更真實反映模型在實際應用中的表現與挑戰。因此，未來相關研究應將資料切分策略作為標準報告項，並以此作為模型泛化能力驗證的重要依據。

CLIP + LoRA 相較於 CLIP Baseline 的優勢

本研究分析 CLIP 結合 LoRA 在深偽檢測任務中的泛化優勢。CLIP 屬於大型視覺語言預訓練模型，具備強大表徵能力，但其原生特徵空間未針對深偽優化，直接全參數微調易造成過擬合，LoRA 以低秩適應方式，只需調整少量參數，能有效針對深偽特徵進行強化，同時保留原模型的泛化潛力。

這種設計促使模型聚焦於與人物身份無關、但能區分真偽的細節特徵，減少對訓練集中已見過人物或場景的依賴，提升對新資料的辨識能力。實驗結果顯示，CLIP + LoRA 在少量樣本下即可於多種深偽生成器和未見資料上展現優異的泛化表現，明顯優於傳統微調或僅用 frozen backbone 的線性探針。

五、related work

Adapting Vision-Language Models for Universal Deepfake Detection

本研究探討如何自適應視覺語言基礎模型如 CLIP，以進行通用型深偽檢測。作者主張保留 CLIP 的文本分支並結合 prompt tuning 技術，使模型能夠整合語意線索進行判斷。實驗顯示，在訓練資料有限的情況下，該方法在多個資料集上均優於僅使用視覺分支的傳統作法，顯示文本語意在提升泛化能力方面具關鍵作用。

Standing on the Shoulders of Giants: Reprogramming VLM for General Deepfake Detection

此研究提出以輸入擾動為基礎的重編程方法，能將預訓練的視覺語言模型轉化為強大的深偽檢測器。其核心設計為結合可學習的視覺擾動與文字提示，使模型無需修改參數即可完成分類任務，並透過下式實現類別預測機率的計算：

$$p(y \mid x) = \frac{\exp\big(\cos(f_I(x'),\, f_T(T))/\tau\big)}{\sum_k \exp\big(\cos(f_I(x'),\, f_T(T_k))/\tau\big)}$$

其中$f_I(x')$表示經加入學習型擾動後的影像特徵，$f_T(T_k)$則是對應第 k 類文字提示 T_k 的向量嵌入，τ 為溫度參數。此設計展現高度的跨資料集與跨類型泛化能力，顯示出以提示驅動的零微調檢測策略之潛力。

Facial Feature-Guided Adaptation for Foundation Models

本方法針對基礎模型提出臉部特徵引導策略，強化模型對人臉區域的空間學習能力。透過結合時序與空間資訊，並引入對比學習損失，顯著提升模型在未見過的偽造樣本上的辨識效能與泛化能力。該方法參數需求低，並適用於影片級深偽檢測任務。

Can ChatGPT Detect Deepfakes

本研究探討多模態大型語言模型如 GPT-4V 在零樣本深偽檢測中的表現。雖然其準確率仍不及專用模型，但具備自然語言解釋能力，可指出影像異常位置並提供判斷依據。此特性在法律、新聞等需要可解釋性的應用場景中具潛在價值。

六、參考資料

[1] J. Wang et al., “Foundation Model with Temporal Prompt for Generalizable Deepfake Detection,” *arXiv preprint arXiv:2503.19683*, 2024. [Online]. Available: https://arxiv.org/html/2503.19683v1

[2] M. Grillini, M. Zago, L. Baroffio, F. G. B. De Natale, and P. Bestagini, “Few-shot Fine-tuning of Foundation Models for Deepfake Detection,” in *Proc. ICPR*, 2024. [Online]. Available: https://iris.unimore.it/bitstream/11380/1343567/2/2024_ICPR_Deepfakes.pdf

[3] GoatStack, “Universal deepfake detection with CLIP,” *GoatStack AI Topics*, 2024. [Online]. Available: https://goatstack.ai/topics/universal-deepfake-detection-with-clip-oygmxq

[4] E. J. Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models,” *arXiv preprint arXiv:2106.09685*, 2021. [Online]. Available: https://arxiv.org/abs/2106.09685