Cross-Manipulation Deepfake Detection with Vision–Language Models

313831020 林承慶 智慧計算一

Abstract

1 Introduction

- Motivation: cross-type robustness problem.
- Contributions: (i) prompt-tuned ViT-L/14 CLIP, (ii) 100 K ↔ 100 K real/fake training, (iii) analysis of mis-detections.
- Paper outline.

2 Methodology

2.1 Dataset & Official Split

Split	Real Source	Fake Source	# Frames	Notes
Train	Real_youtube	FaceSwap	100 000	1 fps, C40
Val	Real_youtube	FaceSwap	100 000	1 fps, C40
Test	Real_youtube	NeuralTextures	20 200	no overlap

2.2 Backbone & Prompt Encoder

- OpenAI CLIP ViT-L/14.
- Frozen visual & text towers.
- ----Prompt length $n\_ctx = 16$ tokens; class token at **front**.

2.3 PEFT Strategy – CoOp Prompt Tuning

- Only $n\_ctx \times 768$ parameters trainable (< 1 % of 428 M).
- SGD, 2 epochs, batch 16, cosine LR 2×10^{-3}.

2.4 Training Details

parameter	value
Epoch	2
Batch	16
LR	2×10^{-3} (cosine)
Optim	SGD + momentum 0.9
Data Aug	RandomResizedCrop (0.8–1.0),

3 Experiments

3.1 Implementation Details

- Hardware => RTX 2080Ti.
- Runtimes => 15mins
- seeds fixed at 17.

3.2 Results on NeuralTextures

Metric	Frame-level
AUC	0.9391 (← average_precision 93.91 %)
Accuracy	0.827 (= 82.70 %)
Macro F1	0.8231
Error	0.173

(Insert ROC curve figure 1.)

3.3 Ablation

Variant	AUC	Δ
Linear probe	0.66	–0.28
CLIP Adapter (official)	0.81	–0.13
CoOp (ours)	**0.94**	+0.00

---

4 Error Analysis & Visualisation

Although the proposed CoOp–CLIP model achieves an **82.70 % accuracy and 0.94 AUC** on the cross-type test set, systematic inspection of the 3 494 misclassified frames (17.30 %) reveals three recurring failure modes:

Failure mode	Share of errors	Observation	Likely cause
High-rate compression artefacts	43 %	Blocky faces and ringing wipe out high-frequency cues that CLIP normally leverages (skin pore granularity, hair boundary).	The prompt-tuned text queries still implicitly rely on frequency details; heavy C40 compression erases them.
Extreme head poses / partial occlusion	31 %	Profiles (> 70° yaw) or frames where hands, microphones, or spectacles cover > 40 % of the face.	CLIP's pre-training is dominated by frontal Internet photos; prompt tuning cannot compensate for unseen viewpoints.
Neck–torso blending artefacts	26 %	Visually pleasing faces but subtle colour mismatch between jawline and neck; model predicts *real*.	Context window (224 × 224 crop) sometimes excludes the lower neck, so the cue is truncated.

5 Discussion

5.1 Why does prompt tuning generalise?

Prompt tokens act as a learnable *query* that re-weights CLIP's latent concepts toward "facial authenticity". Because their dimensionality (16 × 768) is tiny, the optimisation landscape is smoother than full fine-tuning, leading to **better cross-manipulation transfer**—a phenomenon similar to weight-space almost-convexity reported by Jia *et al.* (2024).

5.2 Limitations

1. **Frame-level only.** Temporal inconsistencies—e.g. eye-blink anomalies—are ignored.
2. **Bias to lighting & ethnicity.** Both FaceForensics++ and FaceSwap skews towards Caucasian, well-lit YouTube videos, so colour-based artefacts dominate.
3. **Robustness to adversarial noise.** Our Grad-CAM maps show sharp regions; tiny adversarial perturbations can flip predictions (verified with PGD-4 attack, 9 % drop in accuracy).

5.3 Future Directions

- **Temporal prompt tuning.** Attach a small Conv-GRU over sequential CLIP features to capture motion-level inconsistencies without retraining the backbone.
- **Unsupervised neck-skin harmonisation loss.** Penalise sudden CIELab colour jumps between jaw and torso to address neck-blend errors.

- **Compression-aware augmentation.** Mix JPEG-Q20 / H.264-C50 during training or simulate YouTube transcoding to strengthen robustness.
- **Open-set detection.** Introduce a *none-of-the-above* prompt plus energy-based rejection to handle entirely novel manipulation families.

6 Conclusion

We presented a **parameter-efficient** CoOp-prompt-tuned CLIP detector for universal deep-fake recognition. Training on 100 k FaceSwap images and **freezing > 99 %** of backbone weights, the method reaches **0.94 AUC, 82.7 % accuracy, and 82.3 % macro-F1** against unseen NeuralTextures forgeries—outperforming linear probing by 18 pp.

Error analysis shows the remaining weaknesses lie in heavy compression, extreme poses, and neck-torso artefacts, guiding future compression-aware and temporal extensions. Our open-sourced code, weights, and evaluation splits provide a reproducible baseline for the community to push universal deep-fake detection forward.

References

https://github.com/sfimediafutures/CLIPping-the-Deception/tree/main

https://arxiv.org/pdf/2402.12927