

Quantifying the Efficacy of Foundational Transfer Learning in Digital Forensics: A Comparative Analysis of Parameter Isolated Adaptation vs. Traditional Architectures

Rooha Tanveer[†]

Computer Science

Air University, Multan, Pakistan

rooha9tanveer@gmail.com

Hashir Khan

Computer Science

Air University, Multan, Pakistan

Hashirbaloch.hk@gmail.com

Mamoona Rafique Khan

Computer Science

Air University, Multan, Pakistan

mamoona.rafiqee@aumc.edu.pk

ABSTRACT

Most deepfake videos circulating online are non-consensual sexual content, and women are the targets in nearly all of them. Building detection that works in the field is therefore as much a gender-equity problem as a forensic one. We evaluate the visual encoder of CLIP (ViT-B/32) for binary authenticity classification on FaceForensics++ (C23) using a two-stage parameter-efficient fine-tuning (PEFT) curriculum: linear probing of a classification head on a frozen backbone, followed by partial unfreezing of the last six transformer blocks. A third stage that unfreezes the entire backbone is included as an ablation. Phase 2 reaches AUC 0.9410, accuracy 87.64%, and F1 0.9163; full unfreezing then drops validation AUC to 0.9332 despite a lower training loss, which we read as direct evidence that holistic optimization erases the pre-trained visual priors the model relies on. The best checkpoint is shipped as Cockatoos, an ONNX-exported inference module that runs on CPU so that civil-society responders are not locked out of detection by GPU cost.

CCS CONCEPTS

• Security and privacy → Digital forensics; • Computing methodologies → Machine learning; Computer vision; • Social and professional topics → Gender Equality.

KEYWORDS

Deepfake detection, gender-based violence online, CLIP, Vision Transformers, parameter-efficient fine-tuning, digital forensics.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

GEC 2026, May 7–8, 2026, Patras, Greece

© 2026 Copyright held by the owner/author(s). ACM ISBN 978-1-4503-0000-0/26/05.

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

Deepfake technology does not affect men and women equally. Industry audits of synthetic video circulating online have found that most of it is non-consensual intimate imagery, and that the people in those videos are almost always women [1], [2]. The harm reaches beyond celebrities into the lives of journalists, political activists, and ordinary users, and it has been linked to losses in mental health, employment, and political voice. Building detectors that work in the wild is, for that reason, part of the larger problem of online safety for women.

The field has a generalization problem. Detectors trained from scratch on a single benchmark, such as MesoNet or head-pose models, tend to do well on the test split they were tuned on but lose accuracy when the codec, manipulation family, or camera changes [3], [4]. What they have learned is mostly the noise signature of one dataset, which is exactly the kind of cue that disappears the moment a survivor uploads a re-encoded clip from a phone.

Our position is that a detector should inherit a general-purpose visual prior rather than try to learn one from a small forensic corpus. We adapt the visual encoder of CLIP (ViT-B/32) [5] to binary authenticity classification using a parameter-efficient fine-tuning (PEFT) curriculum that keeps most of the backbone fixed. The empirical study on FaceForensics++ (C23) shows that partial unfreezing beats both linear probing and full fine-tuning, with full fine-tuning measurably worsening validation AUC. The resulting model is shipped as Cockatoos, an ONNX inference package that runs on CPU so civil-society organizations supporting deepfake survivors are not gated by GPU access.

2 METHODOLOGY

2.1 Data and Preprocessing

The C23 (lightly compressed) split of FaceForensics++ [4] is used, which contains original videos and six manipulation families: Deepfakes, Face2Face, FaceSwap, FaceShifter, NeuralTextures, and DeepFakeDetection. From each of 7,000 videos, K=10 frames are sampled at evenly spaced positions. Faces are detected and aligned with RetinaFace through the InsightFace buffalo_1 pack and cropped to 224×224. The resulting corpus has 9,999 real and 28,786 fake aligned crops, split at the video level so no subject appears in more than one partition. Augmentation is designed to reflect what tends to happen to videos in the wild: random JPEG re-encoding ($q \in [70,95]$, $p=0.35$), small rotations, horizontal flips, and a light Gaussian blur.

2.2 Architecture

The vision encoder of CLIP ViT-B/32 has 12 transformer blocks and produces a 768-dimensional output. A small classification head is attached on top: LayerNorm, then Linear (768→256), GELU, Dropout (0.2), and Linear (256→2). The text encoder is discarded. Inputs are normalized with the original CLIP statistics.

2.3 PEFT Curriculum

End to end fine tuning of a foundation model on a small forensic dataset tends to overwrite the broad visual priors that made the model useful to begin with. This failure mode is referred to here as manifold collapse. To avoid it, the curriculum below opens up trainable capacity in stages and keeps most of the backbone fixed throughout.

Phase 1: linear probing. The backbone is frozen and only the head is trained for 3 epochs ($lr=1e-3$, AdamW, weight decay $1e-4$). This gives a near-zero-shot baseline showing how well real and fake faces are already separated in CLIP's embedding space.

Phase 2: partial unfreezing. The last K=6 transformer blocks and the final LayerNorm are unfrozen and trained alongside the head with discriminative learning rates (backbone $5e-5$, head $1e-3$) for 6 epochs. Earlier blocks stay frozen so low- and mid-level structure is preserved, while later blocks shift toward forensic cues such as boundary inconsistencies and lighting mismatches.

Ablation: full unfreezing. To test the manifold-collapse hypothesis directly, the entire encoder is further unfrozen and trained at $lr=1e-6$ for 2 epochs.

All training uses AMP, cross-entropy loss, batch size 64, and selection on validation AUC.

3 RESULTS

Each phase of the curriculum is evaluated on the held-out validation split using AUC, accuracy, precision, recall, and F1. All five matter because deepfake detection is class-imbalanced: a model can post a high single-number score by mirroring the majority class. Table 1 summarizes the three phases.

Configuration	AUC	Acc	Prec	Rec	F1
Phase 1: Linear probe	0.7849	0.7446	—	—	0.8225
Phase 2: Partial unfreeze (adopted)	0.9410	0.8764	0.9388	0.8736	0.9163
Phase 3 (ablation): Full unfreeze	0.9332	0.8633	0.9388	0.8736	0.9050

Table 1: Validation performance across the three training phases on FaceForensics++ (C23). Phase 2 is the configuration we adopt; Phase 3 is reported as an ablation.

Phase 1 already reaches AUC 0.7849 with a frozen backbone and only a trained linear head. Real and fake faces are partially separable in CLIP’s pre-trained embedding space without any forensic supervision, which is the first piece of evidence that the prior is worth keeping. Phase 2 then lifts AUC to 0.9410 and F1 to 0.9163 by unfreezing only the last six transformer blocks. The confusion matrix [[824, 165], [366, 2529]] shows balanced behavior across both classes: precision of 0.9388 keeps false accusations of authentic media low, and recall of 0.8736 keeps coverage of manipulated content reasonable.

The interesting result is what happens when the entire backbone is unfrozen in Phase 3. Training loss falls from 0.0863 to 0.0429, but validation AUC moves the wrong way, from 0.9410 down to 0.9332 the classic signature of overfitting. Many published deepfake detectors that report very high accuracy on a single FaceForensics++ split exhibit the same pattern at a larger scale: they memorize the codec and recurring scene priors of one dataset and then degrade when those statistics shift. The Phase 3 ablation reproduces that failure inside the curriculum and explains why Phase 2 is the configuration adopted: a model that learns less of the dataset keeps more of what it knew before.

Taken together, the three phases trace a clear arc. The pre-trained CLIP backbone already encodes useful structure for authenticity discrimination; partial unfreezing converts that structure into a strong forensic detector with around 30M trainable parameters; full unfreezing degrades the result. PEFT favors cross-distribution robustness over peak in-distribution accuracy. A model overfit to the C23 codec can post higher single-dataset numbers, but it tends to break on social-media re-encodes, screen recordings, and unseen manipulation methods precisely the content that matters in real abuse cases. Comparable transfer-learning baselines on FaceForensics++ such as Xception [6] sit in the same accuracy range while training many more parameters.

4 DEPLOYMENT: THE COCKATOOS MODULE

Most organizations that work with survivors of image-based sexual abuse do not have access to GPU clusters. To make detection usable in those settings, the Phase 2 checkpoint is packaged as Cockatoos, a small inference module exported to ONNX (fixed 224×224 input, opset 17) and run on the ONNX Runtime CPU provider. The package includes a JSON label map and a single-call API that returns a calibrated probability for a face crop. Because the runtime is platform-independent, the same artifact runs on Linux, Windows, macOS, and mobile CPUs without recompilation.

5 CONCLUSION

Building deepfake detection that works in the field is as much a gender-equity problem as a forensic one, because nearly all of the people harmed by non-consensual synthetic media are women. The route taken here is to inherit a general visual prior from CLIP rather than learn one from a small forensic dataset, and to release training capacity in stages so that prior survives. On FaceForensics++ (C23) this produces AUC 0.9410 and F1 0.9163 with only the last six transformer blocks unfrozen, while fully unfreezing the backbone reduces validation AUC despite a lower training loss the empirical evidence behind the manifold-collapse argument and the reason Phase 2 is the configuration adopted.

The Phase 2 checkpoint is shipped as Cockatoos, a CPU only ONNX inference module so survivor-support organizations are not locked out of detection by GPU cost. Decisions about generalization, deployment cost, and how clearly failure modes are reported are read here as gender-equity decisions rather than separate engineering questions. Future work will extend the curriculum to C40 compression and adversarial perturbations, and evaluate Cockatoos on real reports from survivor-support organizations.

ACKNOWLEDGMENTS

We thank our supervisor, for her guidance and patience throughout this project, and for trusting us with the room to figure things out. We also thank our parents, whose support made it possible for us to do this work in the first place.

REFERENCES

- [1] H. Ajder, G. Patrini, F. Cavalli, and L. Cullen, "The state of deepfakes: Landscape, threats, and impact," Deeptrace, Tech. Rep., 2019.
- [2] Sensity AI, "The state of deepfakes 2020: Updated landscape and threats," Sensity AI, Tech. Rep., 2020.
- [3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in Proc. IEEE Int. Workshop Information Forensics and Security (WIFS), 2018, pp. 1–7.
- [4] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2019, pp. 1–11.
- [5] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Machine Learning (ICML), 2021, pp. 8748–8763.
- [6] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1251–1258.