Abstract - Benchmarking the Robustness of Fake-Video Detectors Under Adversarial Perturbations

The rapid progress of diffusion-based generators has enabled the creation of realistic and temporally coherent synthetic videos, pushing multimedia forensics toward adversarial settings.The core challenge is whether detection models remain stable under imperceptible perturbations capable of flipping predictions and undermining authentication and misinformation monitoring.

While adversarial robustness has been extensively studied for images, video detectors introduce temporal and multimodal dependencies that alter vulnerability patterns. Perturbations may propagate across frames, disrupt motion cues, or interfere with optical-flow estimation, making robustness harder to assess.

This work introduces a unified benchmark covering 2D Convolutional Neural Network (CNN)-based detectors derived from ResNet50-like backbones, 3D CNN-based architectures modeling spatiotemporal patterns, Vision Transformer (ViT)-based models using self-attention over video sequences, vision–language-inspired detectors, and multimodal systems combining RGB appearance with optical-flow motion cues. The benchmark evaluates these architectures under FGSM, PGD, DeepFool, and a Universal Attack, with matched perceptual distortion and consistent video-level decision rules.

Experimental results show that 2D CNN-based detectors fail under all perturbations, while 3D CNN-based models offer only limited robustness. ViT-based detectors achieve the strongest stability, resisting all but the most aggressive attacks. In multimodal systems, perturbing RGB alone is sufficient to compromise the motion branch because optical flow is recomputed from altered frames, causing severe PGD vulnerability. Transferability is strongest within architectural families and, across families, attacks crafted on transformer-based models transfer more effectively to CNNs.

Ablation studies reveal that prototype-based heads and frozen encoders stabilize Vision Transformer embeddings, while extensive temporal aggregation strengthens robustness by enforcing motion consistency. Flow-based features prove naturally more stable than RGB, but this advantage disappears when optical flow is derived from adversarially perturbed frames.

Overall, the benchmark provides a reproducible foundation for evaluating fake-video detectors under adversarial conditions and clarifies the architectural and temporal factors that shape robustness in modern video forensics.

Note: This work is derived from my Master's thesis and is being developed into a letter that we plan to send to IEEE Signal Processing Letters by the end of December.