Multi-Domain Deepfake Detection Framework Using Transformer-Based Cross-Attention Fusion

Sanjaya Shankar Tripathy *
Assistant Professor (Supervisor), Department of Electronics & Communication Engineering, Birla Institute of Technology, Mesra, Ranchi, India

Tarush Anand
Department of Electronics & Communication Engineering, Birla Institute of Technology, Mesra, Ranchi, India

Chetan Singh
Department of Electronics & Communication Engineering, Birla Institute of Technology, Mesra, Ranchi, India

Rounit Kumar Sinha
Department of Electronics & Communication Engineering, Birla Institute of Technology, Mesra, Ranchi, India

Abstract

Modern generative models such as style-based Generative Adversarial Networks (StyleGAN) synthesize facial images realistic enough that single-cue deepfake detectors are no longer reliable. This paper presents a quad-stream deepfake detection framework that jointly analyses a facial image from spatial, frequency, noise-residual, and statistical forensic perspectives. Stream 1 uses a ViT-Base/16 or Swin-Transformer backbone with an auxiliary mask-supervision branch for global context and forgery localisation. Stream 2 uses a modern CNN backbone (EfficientNet-V2-S / ConvNeXt) for spatial texture cues. Stream 3 forms a five-channel Learned Noise Pattern (LNP) fingerprint from deep denoising residuals fused with FFT amplitude and phase spectra. Stream 4 fits a Generalised Gaussian Distribution to the 63 AC coefficients of an 8×8 block DCT to build a 126-D statistical fingerprint. All four 256-D stream embeddings are combined through a Gated Cross-Attention Fusion module that adaptively re-weights each stream. Trained and evaluated on the 140k Real and Fake Faces dataset (70,000 real FFHQ and 70,000 StyleGAN images), the proposed system attains a test AUC-ROC of 0.9371 and an overall accuracy of 89.18% on 20,000 unseen samples, with validation AUC reaching 0.9905 after 15 epochs.

Keywords

Deepfake detection, StyleGAN, Vision Transformer, frequency-domain analysis, noise residual fingerprinting, multi-stream fusion, gated cross-attention.

Acknowledgment

The authors thank the Department of ECE, BIT Mesra for academic support, and Dept. of CSE, BIT Mesra HPC facility for access to the NVIDIA L40 GPU cluster used for all training runs.