Adversarial Robustness of Deepfake Detectors: Does Input Representation Matter?

Alice, Sean, Meiqi, Yuantong

Visual Computing — PBL Final Presentation

Problem Statement

Real Face

Unmodified facial image used as baseline for detection models.

Deepfake

Synthetically generated face — the primary target for detection.

Adversarial Attack

Imperceptible pixel changes that cause detectors to misclassify.

Research Question

Does modifying the input representation (pixel vs. frequency domain) improve the robustness of deepfake detectors against adversarial attacks?

Approach: Three Classifiers

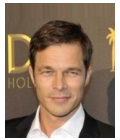
Same CNN architecture, same data. Only the input representation changes.

Pixel Classifier

Input: Raw RGB Image

CNN Architecture

Spectrum Classifier



Input: Raw RGB Image

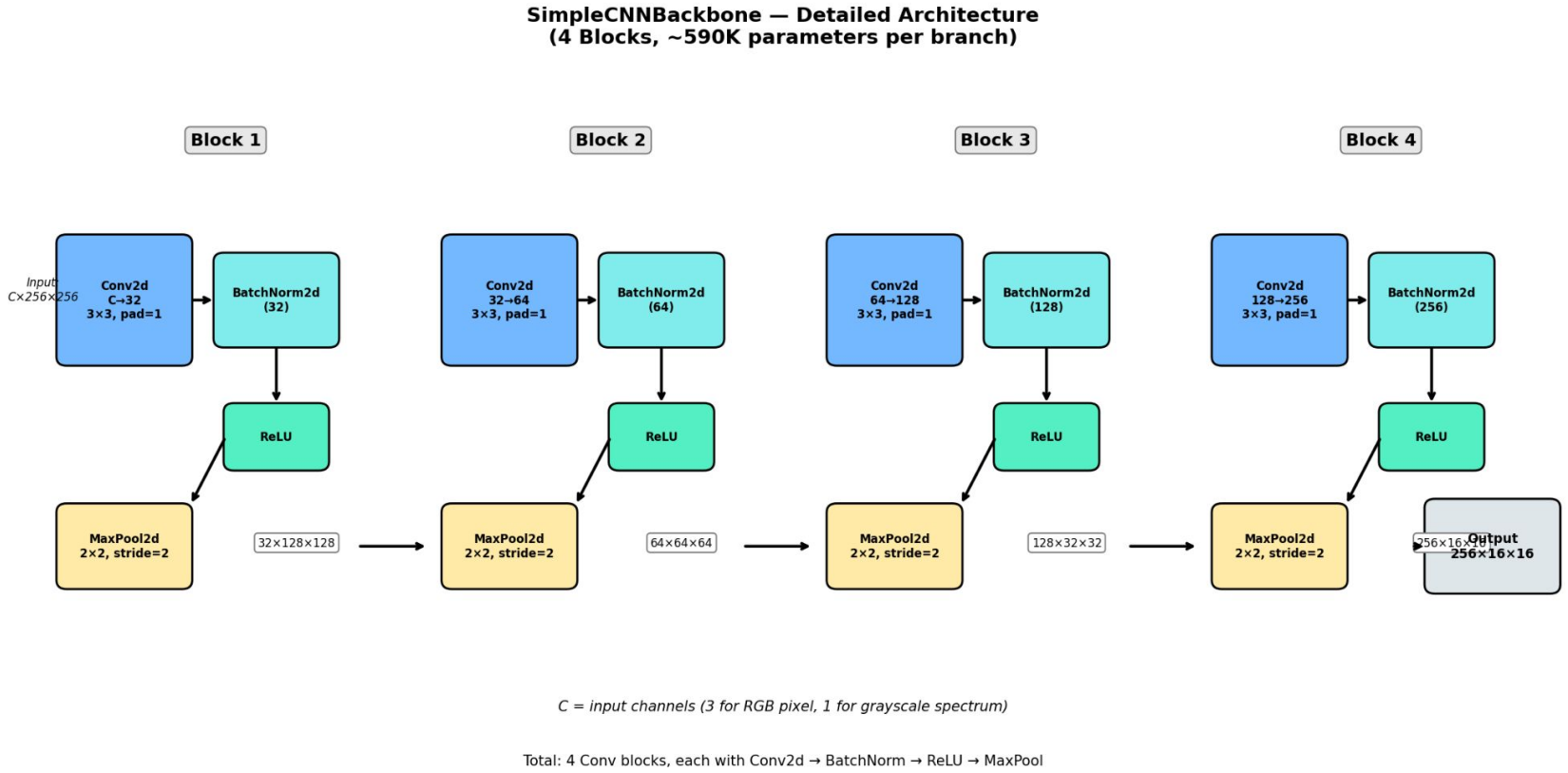
2D FFT (Log-Magnitude)

Dual-Branch Classifier

Input: Raw RGB Image

Concatenated Feature Maps

CNN Architecture



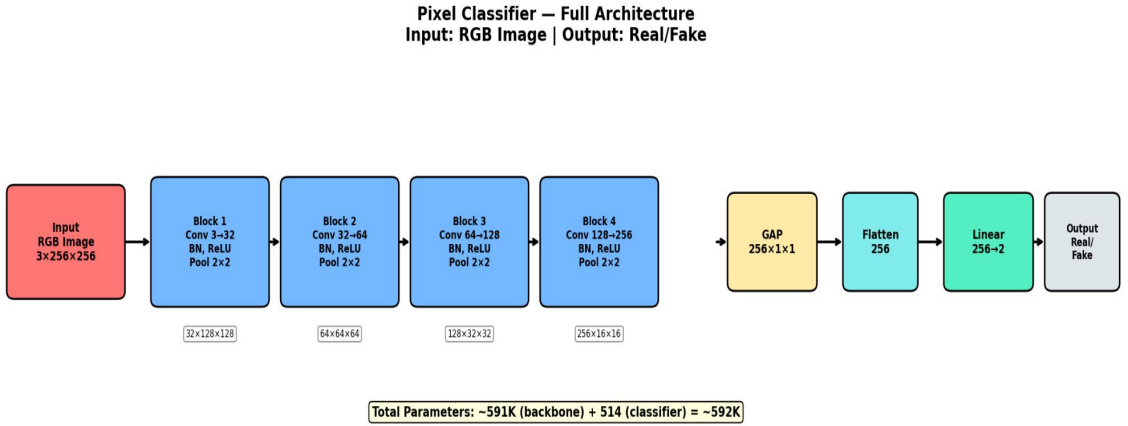
Approach: Three Classifiers

Same CNN architecture, same data. Only the input representation changes.

Pixel Classifier

Input: Raw RGB Image

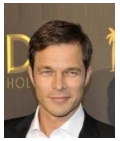
CNN Architecture



Approach: Three Classifiers

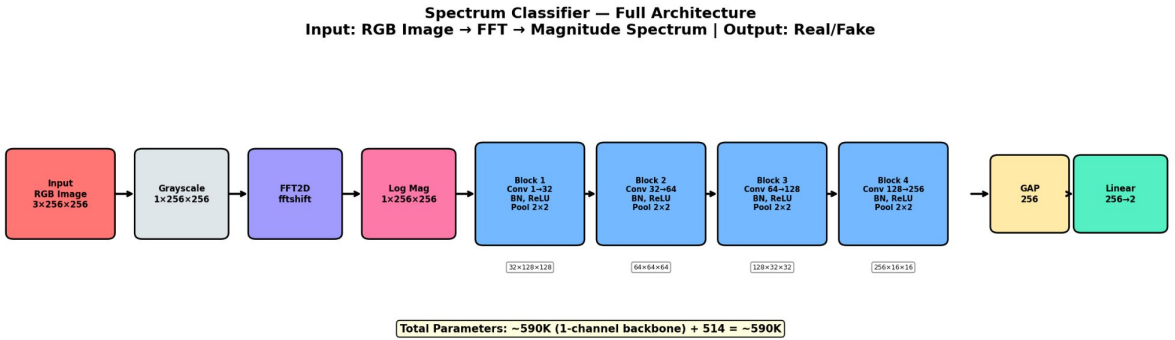
Same CNN architecture, same data. Only the input representation changes.

Spectrum Classifier



Input: Raw RGB Image

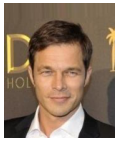
2D FFT (Log-Magnitude)



Approach: Three Classifiers

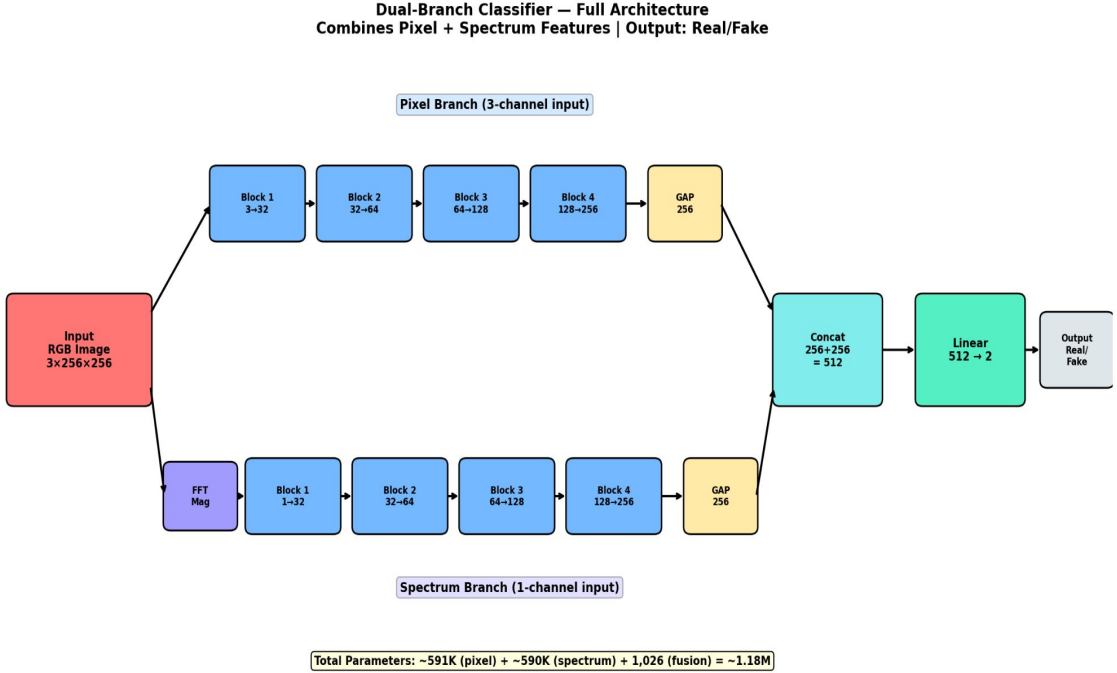
Same CNN architecture, same data. Only the input representation changes.

Dual-Branch Classifier



Input: Raw RGB Image

Concatenated Feature Maps



Model Architecture Comparison

Component	Pixel Classifier	Spectrum Classifier	Dual-Branch Classifier
Input	3×256×256 RGB	1×256×256 Spectrum	Both
Preprocessing	None	Grayscale → FFT → Log Mag	Both paths
Backbone	4-block CNN 32→64→128→256	4-block CNN 32→64→128→256	2× 4-block CNN
Feature Dim	256	256	512 (concat)
Classifier	Linear 256→2	Linear 256→2	Linear 512→2
Parameters	~592K	~590K	~1.18M
Output	16×16 feature map	16×16 feature map	2× 16×16 → concat

- **Identical Backbone across all 3 classifiers**
- **Controlled experiment: isolated effect of input representation**
- **Dual-branch has 2× parameters but combines both feature spaces**
- **Same training setup: Adam optimizer, 15 epochs, batch size 32**

Dataset Overview

Training / Validation / Test

5,000 real faces
4,630 AI-generated fakes

Split: 80% train / 10% val / 10% test
All images resized to 256×256
Deterministic split (seed 42)

Cross-Generator Test Set

Never seen during training

SD 1.5 — 500 images (512→256)
SD 2.1 — 500 images (768→256)
SDXL — 500 images (1024→256)

250 male + 250 female per version

Training data tests detection capability. Cross-generator data tests generalization to unseen generators.

Spectral Analysis

High-Frequency Energy

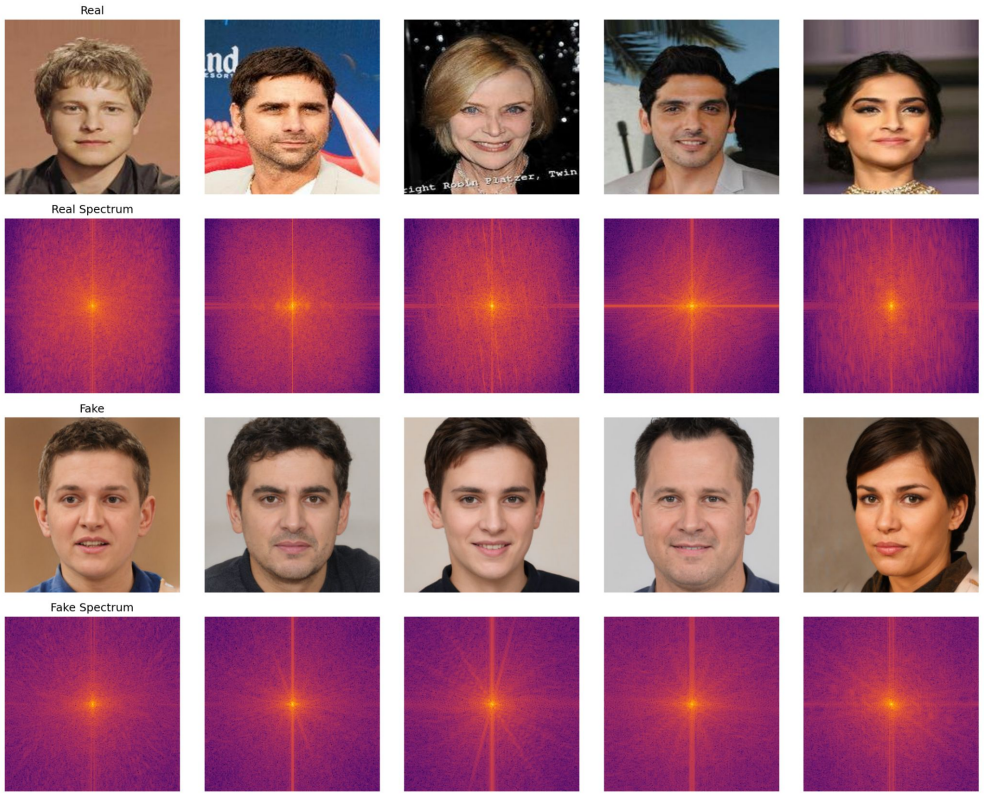
Fake images exhibit significantly more high-frequency energy in their power spectrum, indicating structural artifacts from the generation process.

Azimuthal Profiles

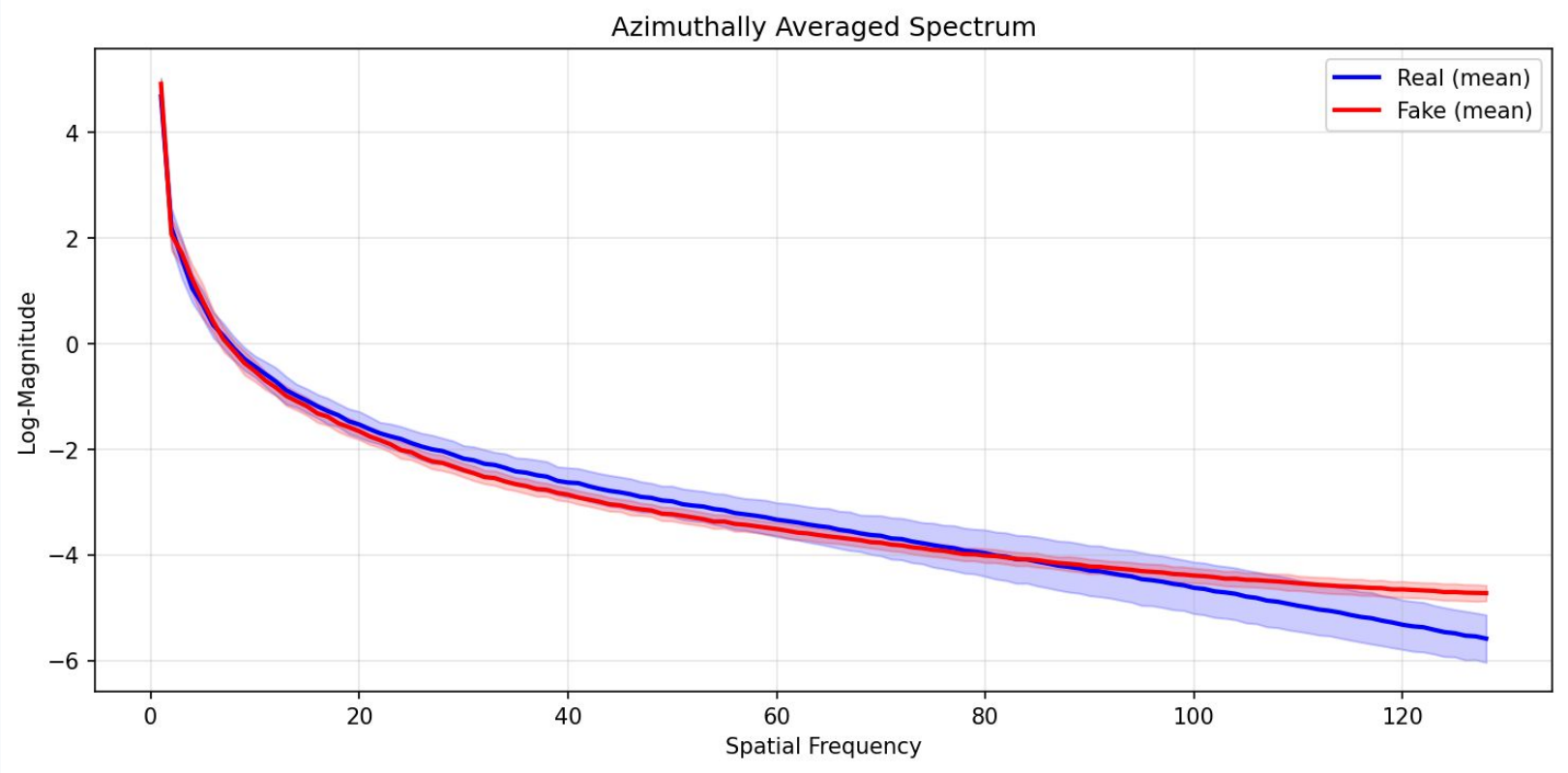
Radial average of the spectrum shows clear separation between real and fake profiles. Grayscale spectrum is more discriminative than per-channel RGB.

Key Insight: Frequency domain features are discriminative — fake images leave distinct spectral fingerprints. This motivates testing whether spectral representations also provide adversarial robustness.

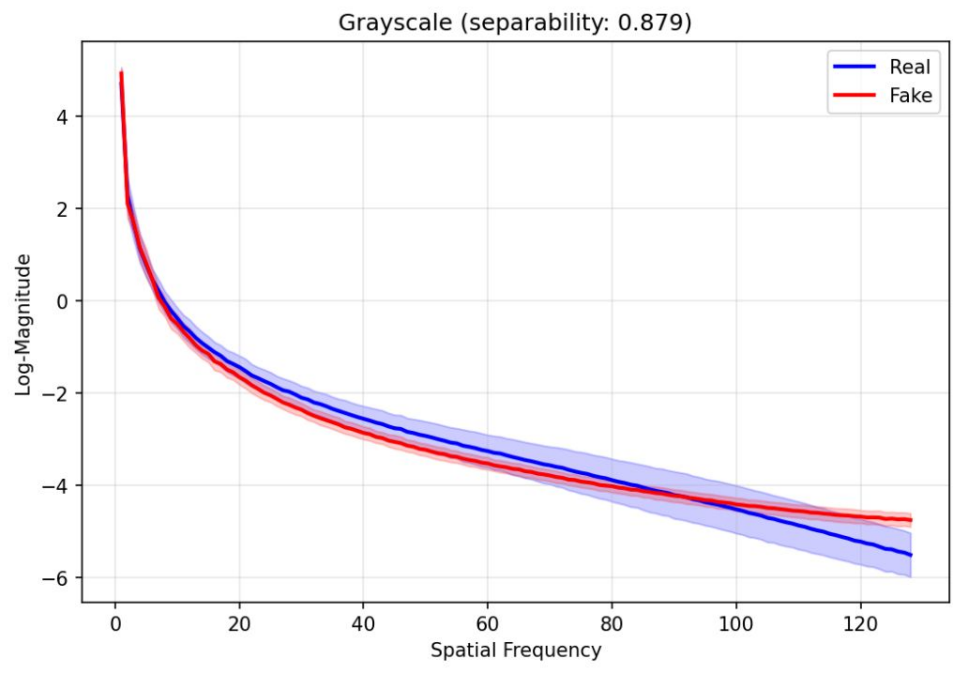
Spectral Analysis

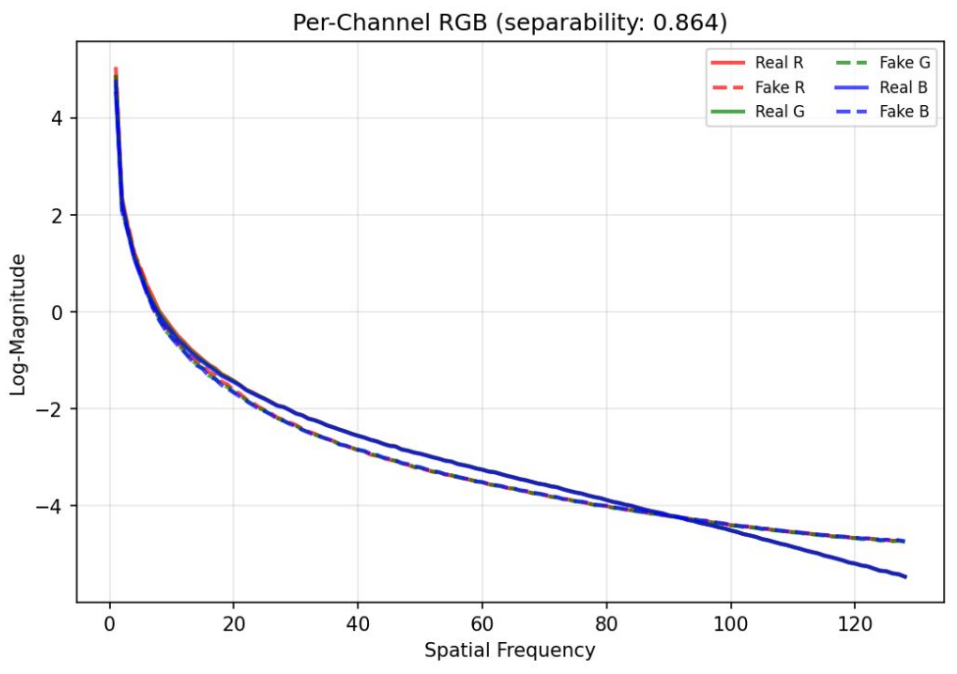


Spectral Analysis: Why Frequency Matters

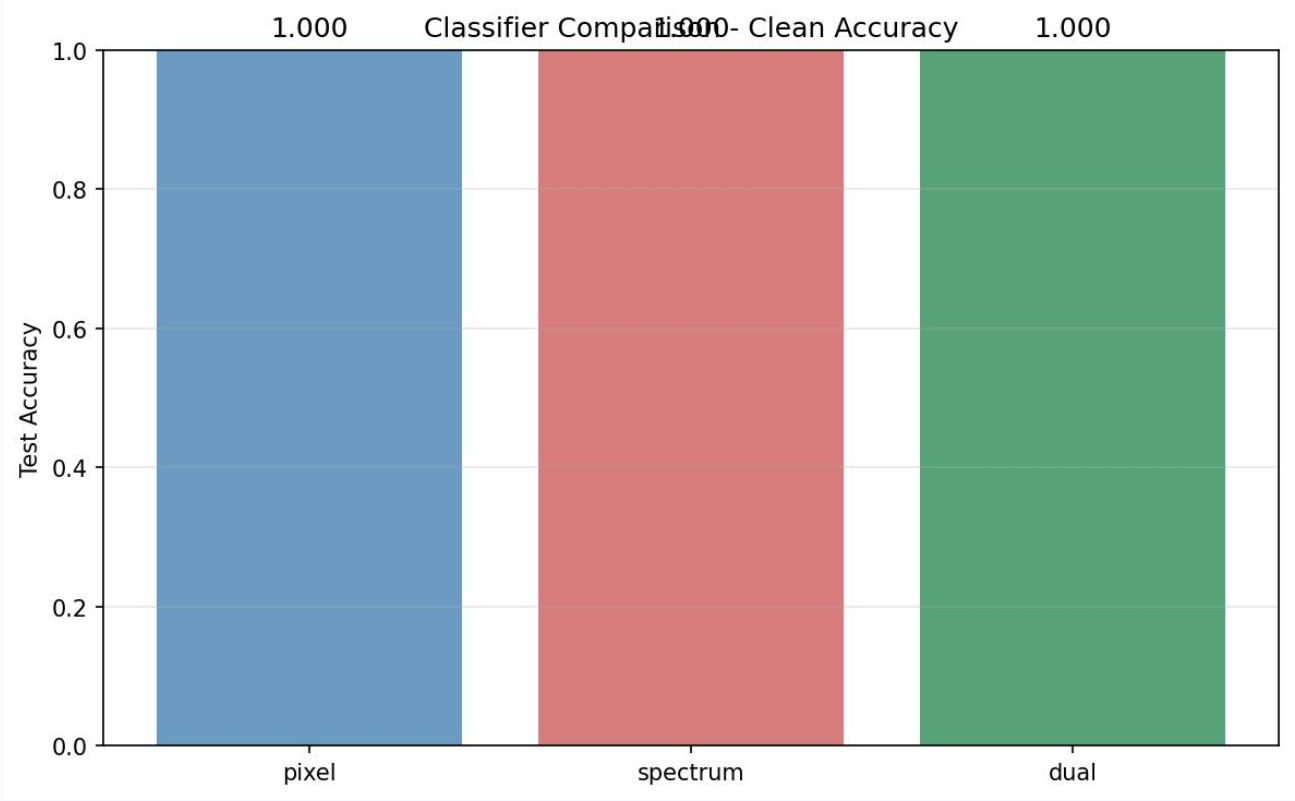


Spectral Analysis

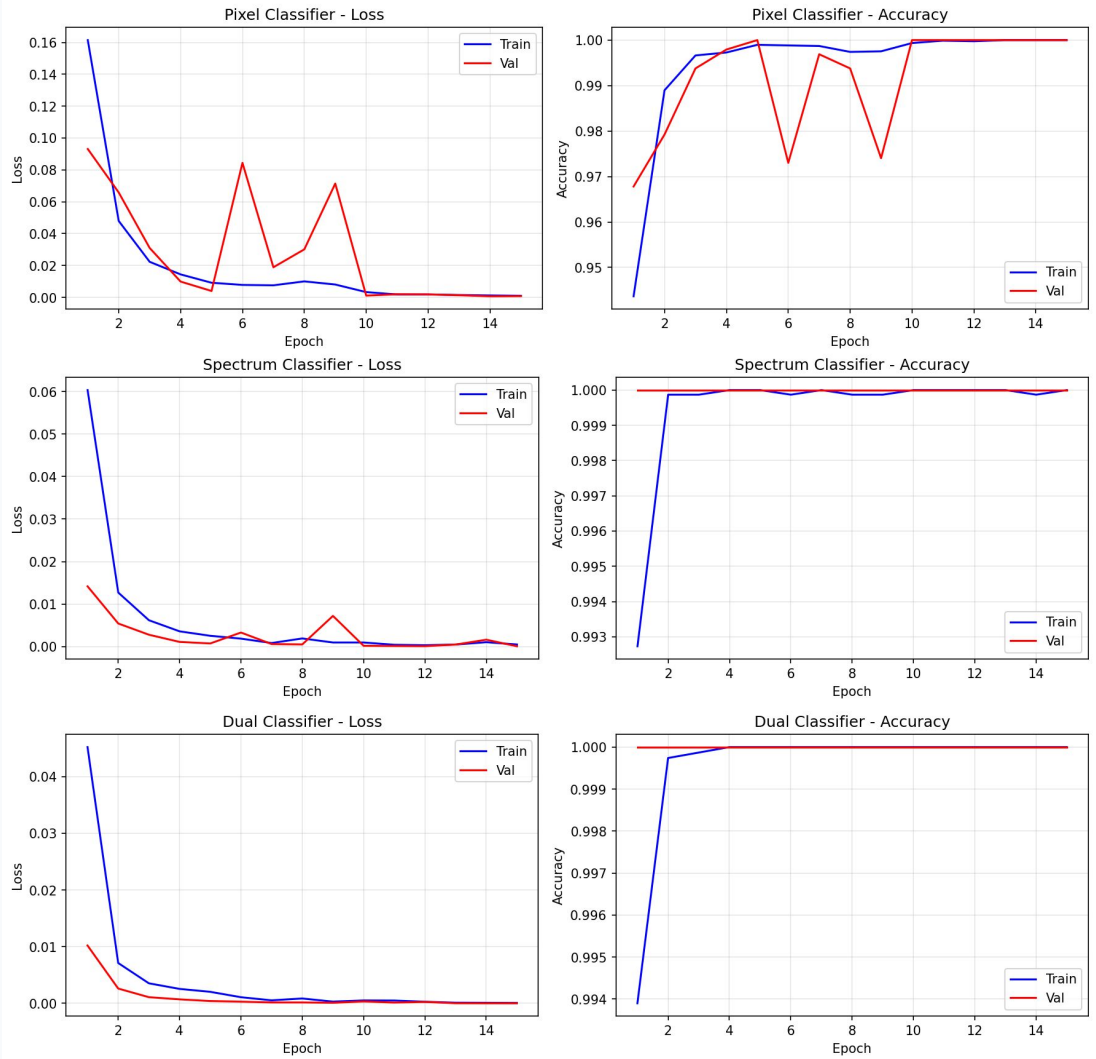




Training Curves

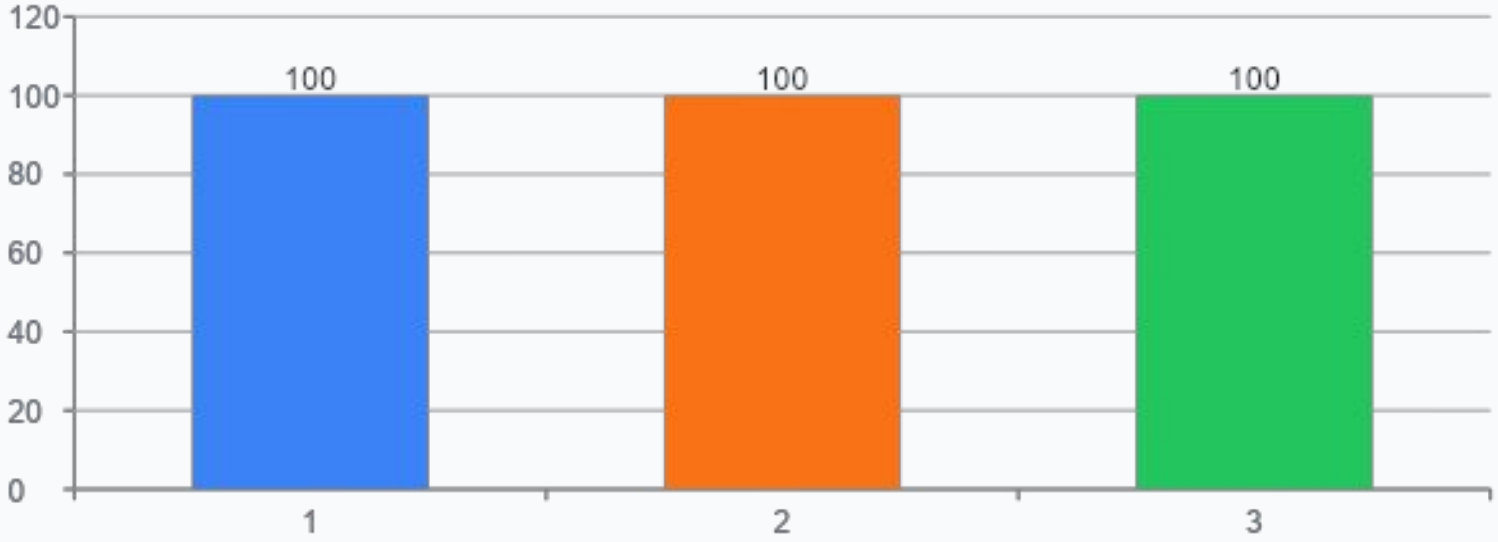


Training Curves



Clean Accuracy (No Attacks)

All three classifiers achieve perfect accuracy on the held-out test set.



All three classifiers achieve perfect accuracy on in-distribution test data, establishing a baseline for adversarial evaluation.

Adversarial Attacks: FGSM & PGD

Fast Gradient Sign Method (Single-step)

One gradient step. Fast but weaker.
Perturbation = ε × sign(gradient).

Projected Gradient Descent with 20 iterations (Iterative)

20 gradient steps with α = ε/4. Stronger, more reliable attack.
L∞ projected.

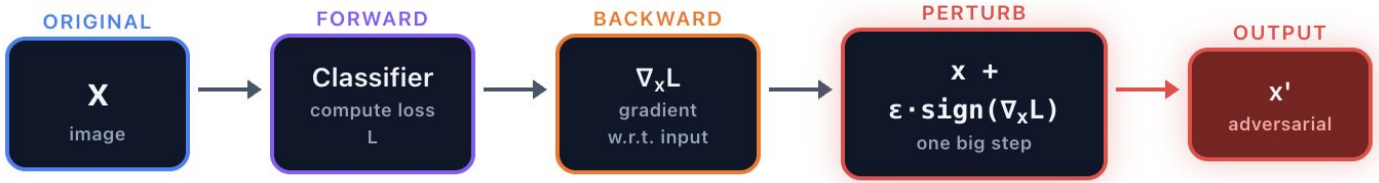
Attacking the Spectrum Classifier

Perturbations are added in pixel space, then the FFT is computed.
Gradients backpropagate through the FFT to optimize perturbations for the frequency domain.
The FFT is linear and differentiable — the attacker can craft pixel-space perturbations that produce targeted effects in frequency space.

Adversarial Attacks: FGSM & PGD

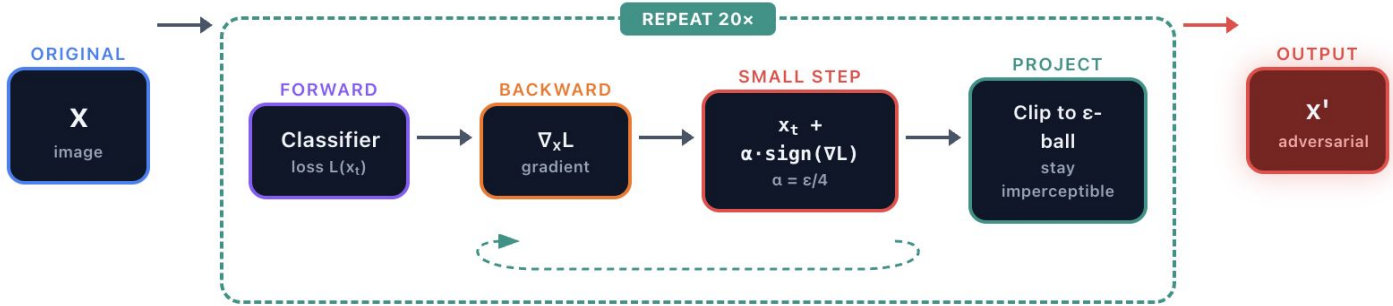
FGSM Fast Gradient Sign Method

1 STEP

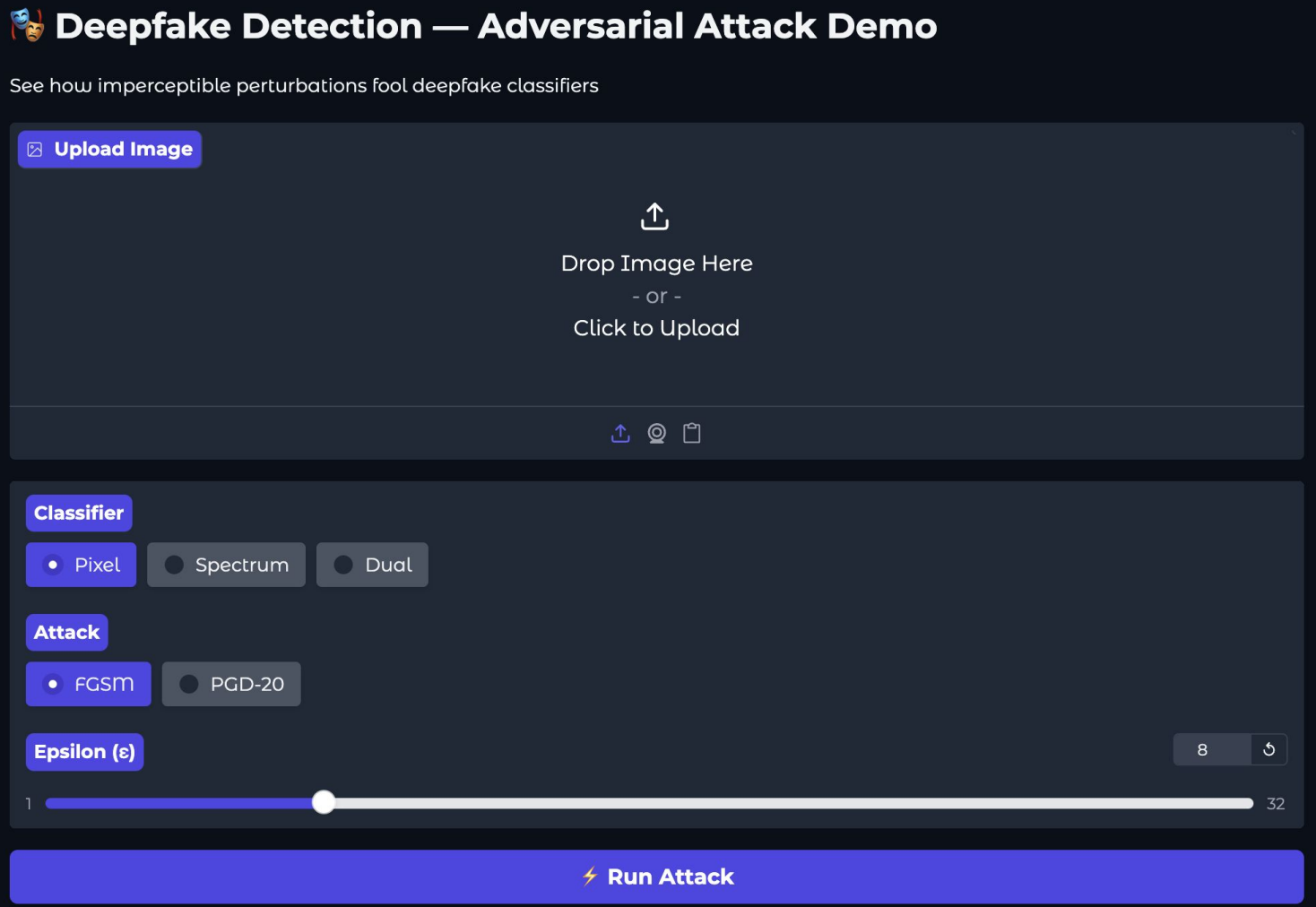


PGD-20 Projected Gradient Descent

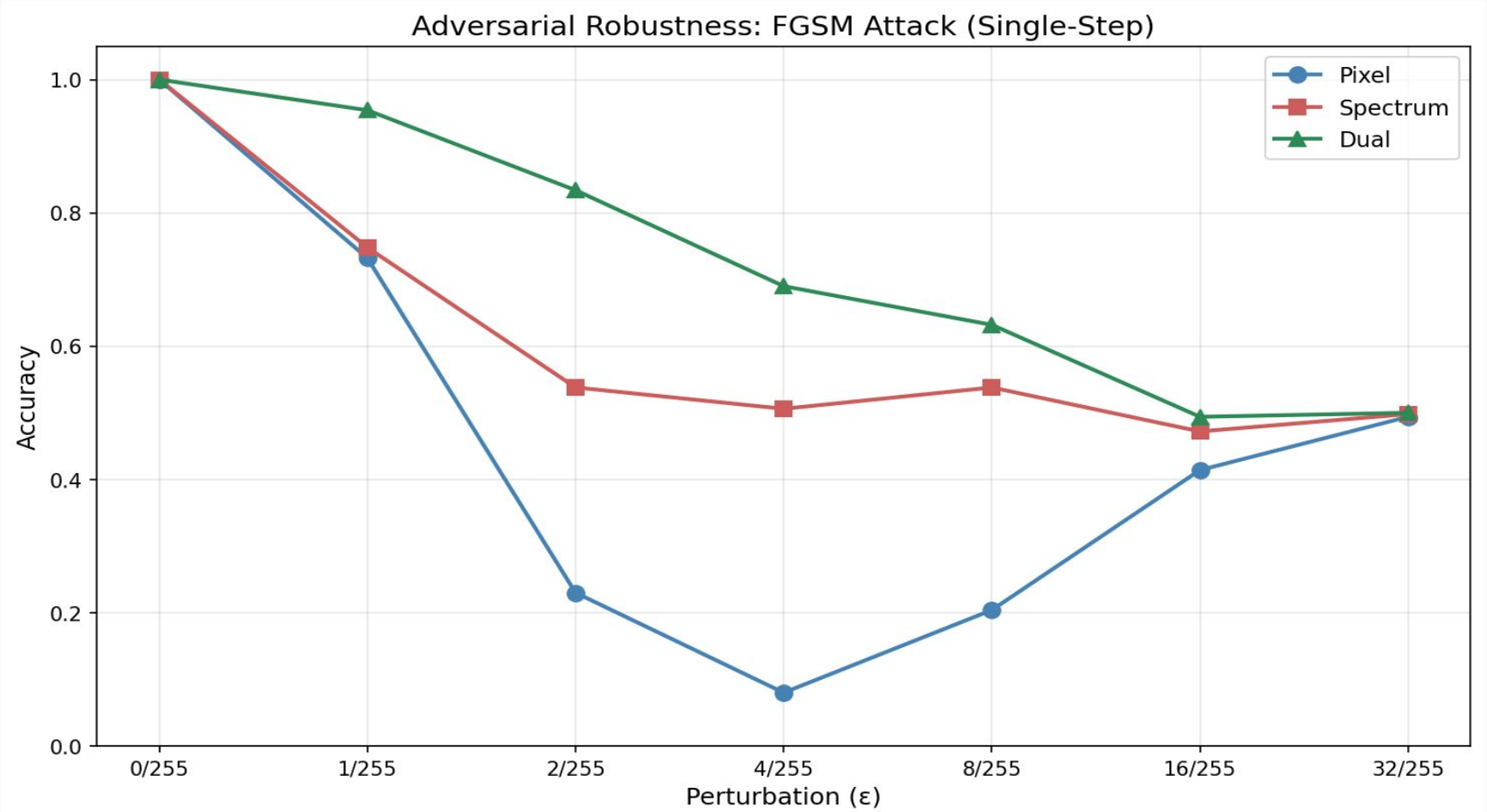
20 STEPS



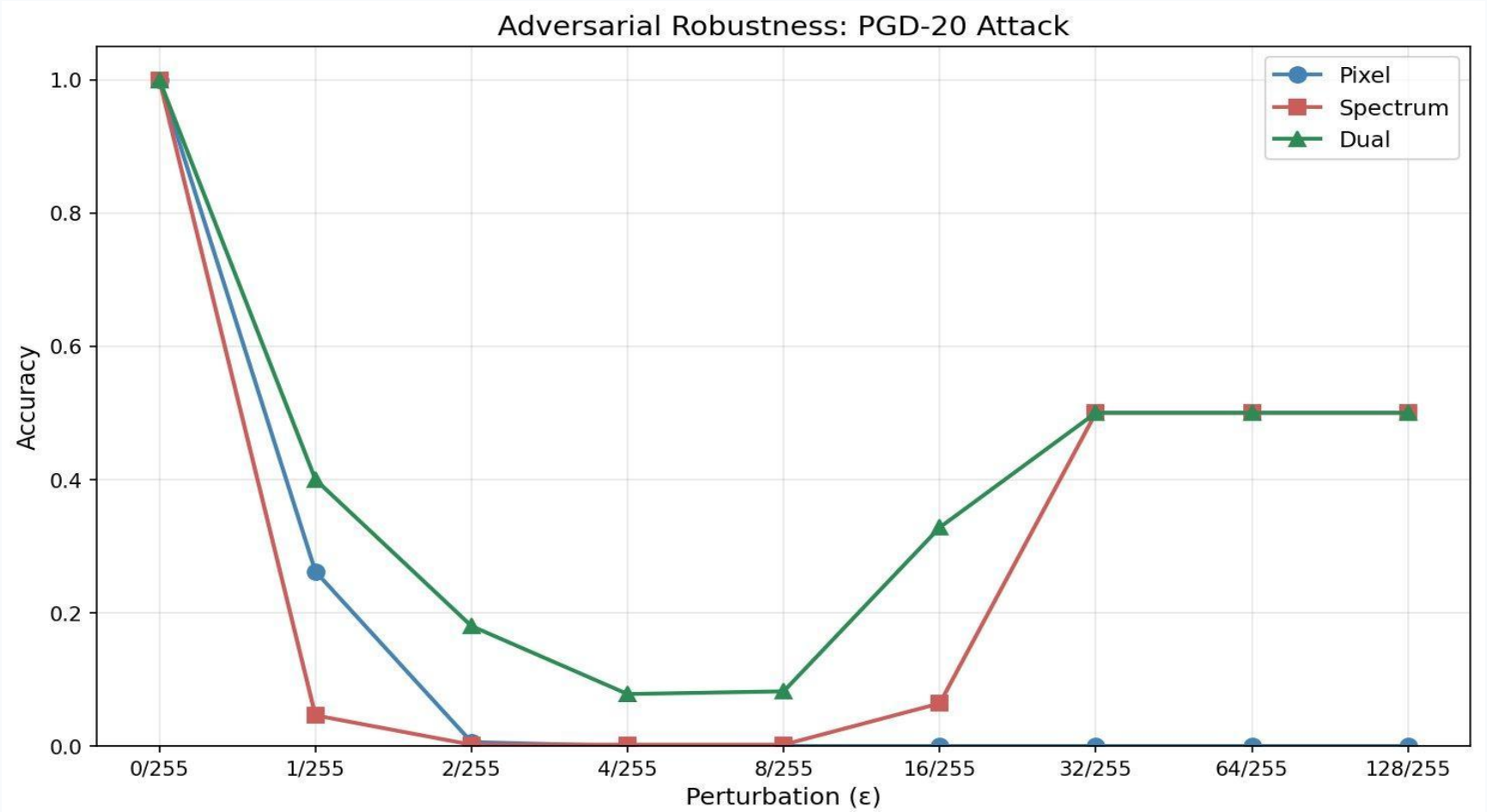
Demo


Deepfake Detection — Adversarial Attack Demo
See how imperceptible perturbations fool deepfake classifiers
Upload Image
Drop Image Here
- or -
Click to Upload
Classifier
Pixel
Spectrum
Dual
Attack
FGSM
PGD-20
Epsilon (ε)
8
1
32
Run Attack

Adversarial Robustness: FGSM Results



Adversarial Robustness: PGD-20 Results



Interpreting the Adversarial Results

We focus on practical perturbation range: ε = 1/255 to 8/255

ε	Pixel	Spectrum	Dual
0 (clean)	100%	100%	100%
1/255	26%	4%	40%
2/255	1%	1%	18%
4/255	0%	1%	8%
8/255	0%	1%	9%

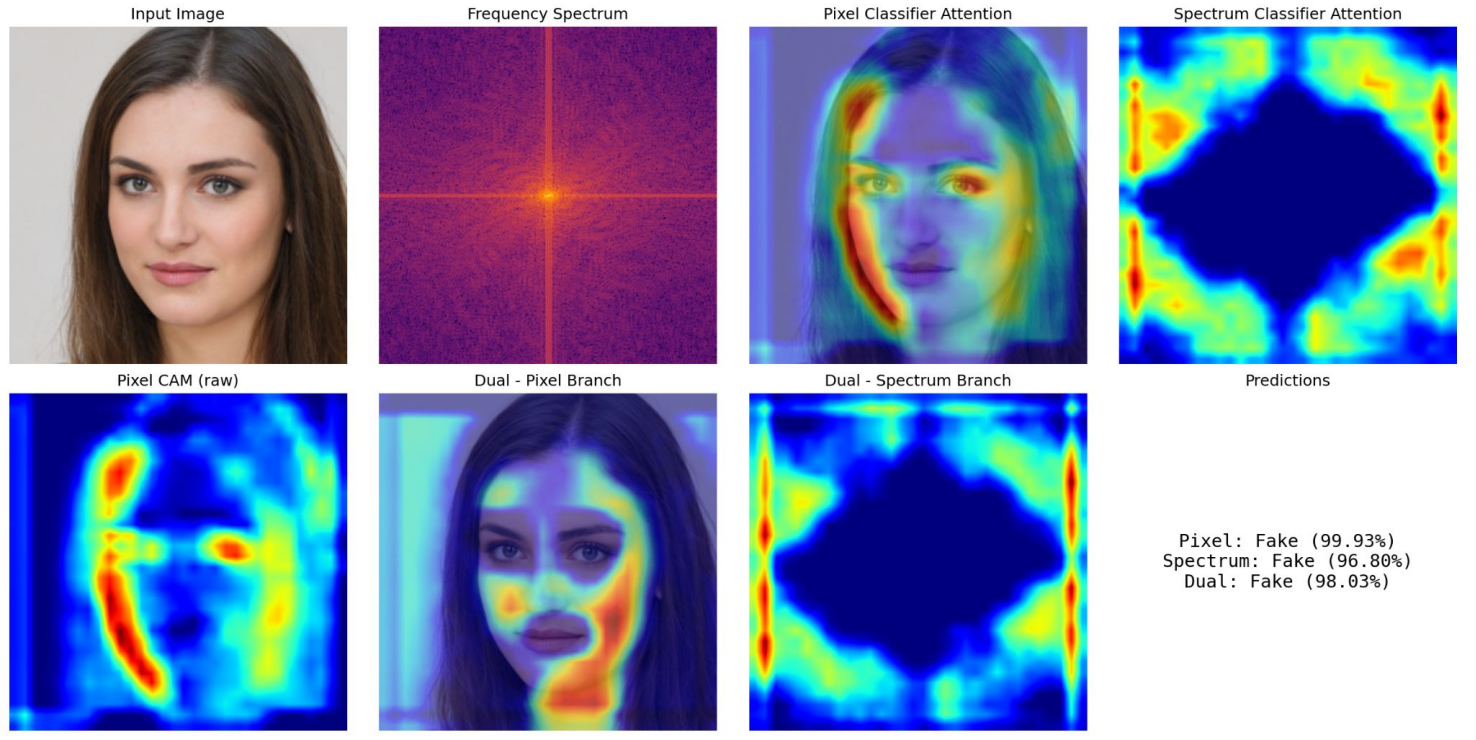
Key Findings

Dual-branch is most robust
Degrades slowest across all ε values

Spectrum is most fragile
Collapses to 4% at ε = 1/255

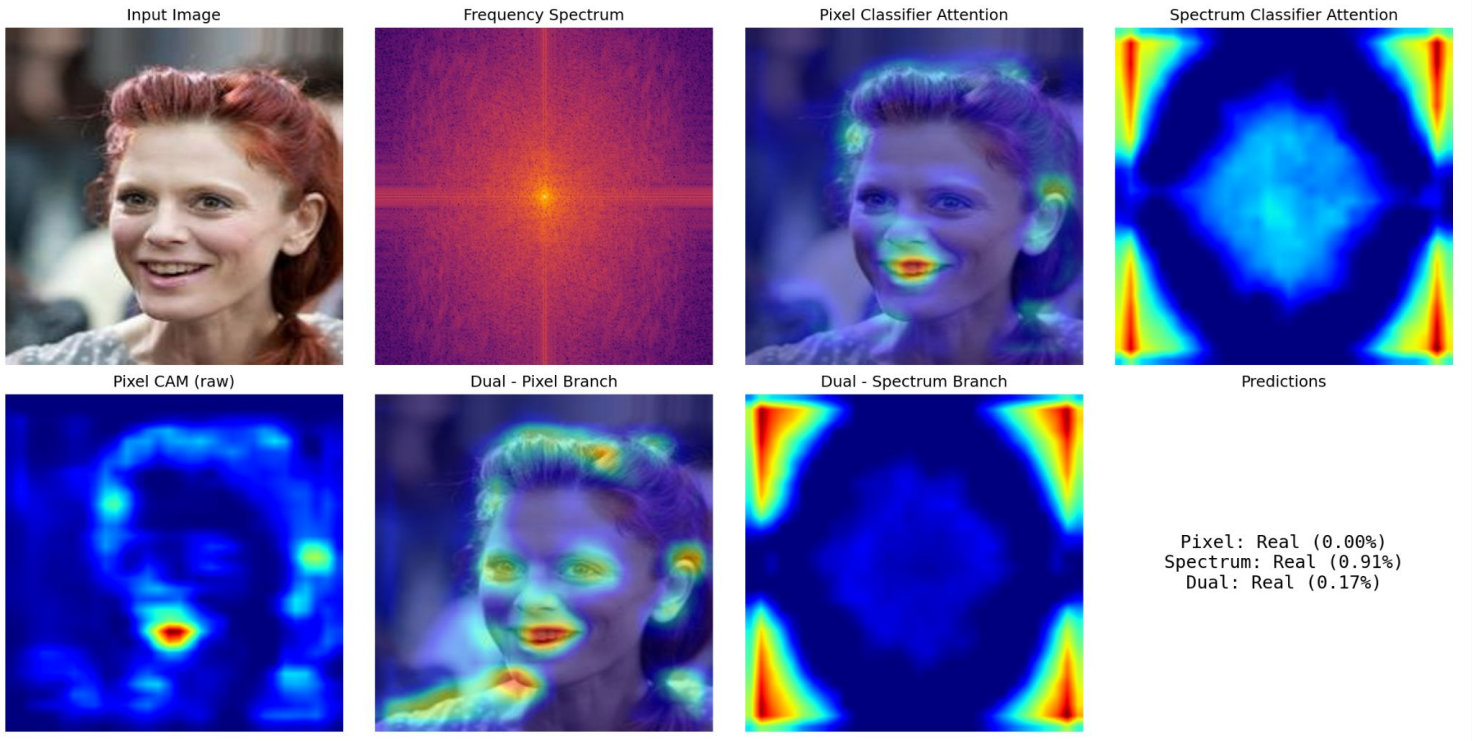
50% plateau ≠ robustness
At high ε, models output random guesses — the attack destroys all signal

Interpretability / Grad-CAM



- Pixel CAM: weak, diffuse activation
- Spectrum CAM: activation at edges/corners but low intensity
- Dual pixel branch: scattered, no focal point

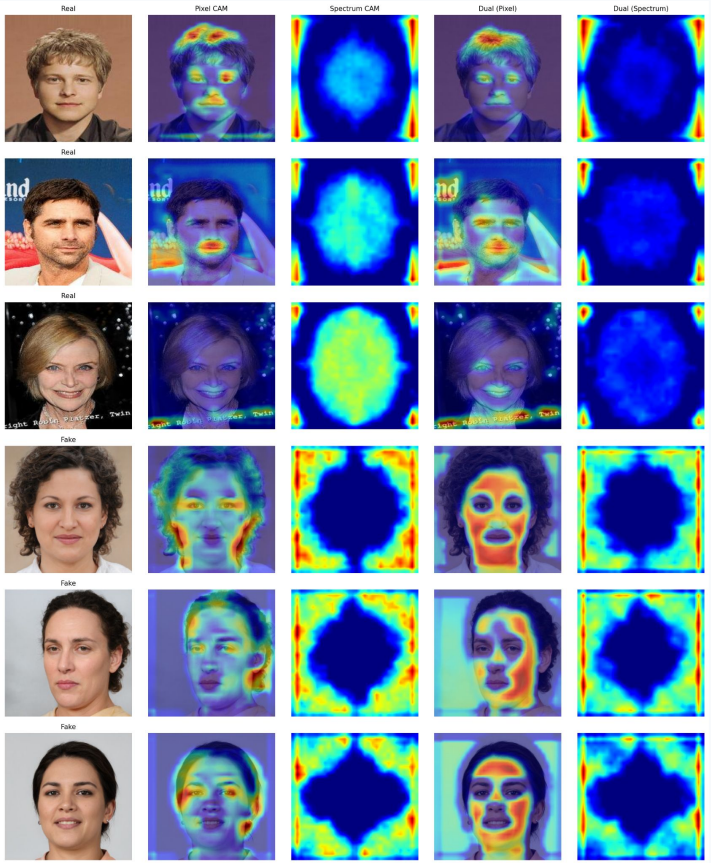
Interpretability / Grad-CAM



- Pixel CAM: strong activation on eyes, nose, jawline
- Spectrum CAM: hot edges/corners (high-frequency artifacts)
- Dual pixel branch: even more concentrated than standalone pixel

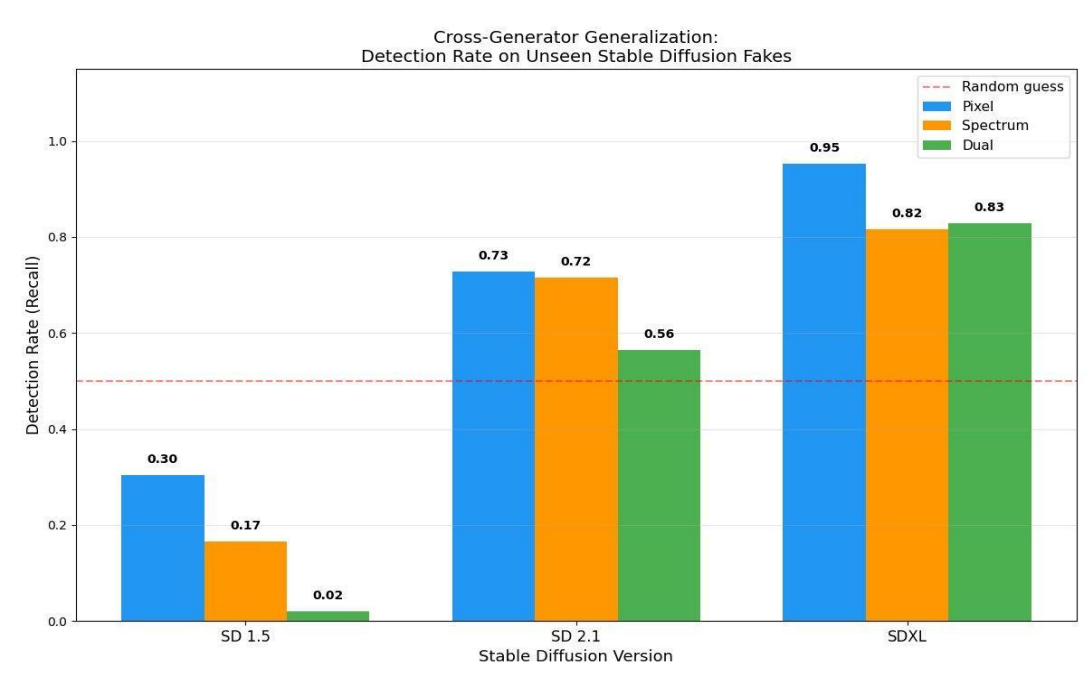
Interpretability / Grad-CAM

- Spectrum attention pattern is nearly identical regardless of image (rigid, predictable)
- This rigidity helps explain why spectrum classifier is easiest to attack



Cross-Generator Generalization

Can classifiers trained on one generator detect fakes from Stable Diffusion (never seen in training)?



Findings

All classifiers fail on SD 1.5
Below random chance — learned artifacts don't transfer

Pixel generalizes best overall
Outperforms spectrum and dual on every SD version

Detection improves SD 1.5 →2.1→SDXL
Higher-res generators share more artifacts with training data after downsampling

Dual-branch overfits worst
More inputs ≠ better generalization

Conclusions

Answer to Research Question

Input representation significantly affects adversarial robustness — but not in the expected direction. The frequency-domain classifier is the most fragile, not the most robust.

Adversarial Robustness

Dual-branch is most robust. Spectrum collapses at ε=1/255. Redundant representations help more than any single domain.

Cross-Generator

Pixel classifier generalizes best. Spectral fingerprints are partially generator-specific, not universal.

Practical Implication

For real-world deployment: combine representations (dual-branch) for robustness, prefer pixel features for generalization.

Future Directions

1. **Adversarial Training**

   Train classifiers on adversarial examples to improve robustness across all representations.

2. **More Generators**

   Test on Midjourney, DALL-E, and newer diffusion models to validate cross-generator findings.

3. **Hybrid Architectures**

   Explore attention-based fusion of pixel and spectrum branches instead of simple concatenation.

4. **Real-time Detection**

   Benchmark inference speed of spectrum computation for practical deployment scenarios.

Thank you!