UNMASKING THE ARTIFICIAL

Forensic Defense Against Deepfake Imagery

Kenneth G. Hartman

GSE, CISSP, GCFA, GCFE, GASF

Lucid Truth Technologies | SANS Certified Instructor

AGENDA

01 **The Democratization of Deception**
Threat landscape & the challenge

02 **Hygiene & Metadata Analysis**
Exif, ELA, and their limits

03 **Camera Ballistics — The "Gold Standard"**
PRNU sensor fingerprinting

04 **Detecting the Ghost in the Machine**
Frequency analysis & model fingerprinting

05 **The Future of Authentication (C2PA)**
Cryptographic provenance

06 **Conclusion & Practical Takeaways**
Defense in depth

SECTION I

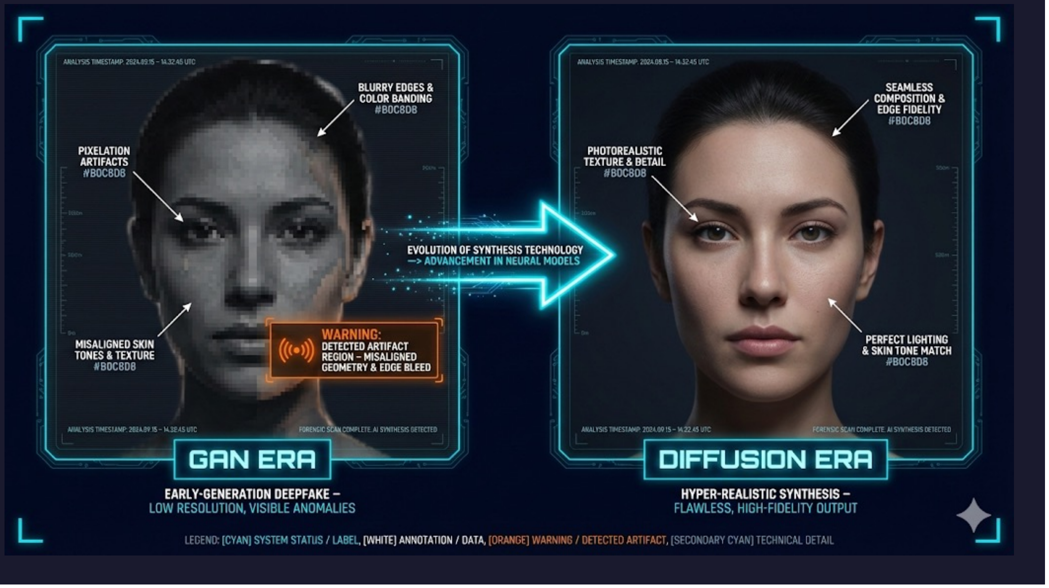
The Democratization of Deception

How generative AI redefined the threat landscape

The Current Landscape

- **Early Deepfakes:** Face-swaps and crude manipulations gave way to sophisticated neural synthesis
- **Generative Adversarial Networks (GANs):** Generator vs. discriminator produces photorealistic output
- **Diffusion Models:** DALL·E, Midjourney, Stable Diffusion — text-to-image at unprecedented quality

Each generation of models makes previous detection methods less reliable


ANALYSIS TIMESTAMP: 2024.09.15 – 14.32.45 UTC
BLURRY EDGES & COLOR BANDING #B0C8D8
PIXELATION ARTIFACTS #B0C8D8
MISALIGNED SKIN TONES & TEXTURE #B0C8D8
WARNING: DETECTED ARTIFACT REGION – MISALIGNED GEOMETRY & EDGE BLEED
ANALYSIS TIMESTAMP: 2024.09.15 – 14.32.45 UTC
FORENSIC SCAN COMPLETE. AI SYNTHESIS DETECTED
GAN ERA
EARLY-GENERATION DEEPFAKE – LOW RESOLUTION, VISIBLE ANOMALIES
EVOLUTION OF SYNTHESIS TECHNOLOGY
—> ADVANCEMENT IN NEURAL MODELS
ANALYSIS TIMESTAMP: 2024.08.15 – 14.32.45 UTC
SEAMLESS COMPOSITION & EDGE FIDELITY #B0C8D8
PHOTOREALISTIC TEXTURE & DETAIL #B0C8D8
PERFECT LIGHTING & SKIN TONE MATCH #B0C8D8
ANALYSIS TIMESTAMP: 2024.09.15 – 14.22.45 UTC
FORENSIC SCAN COMPLETE. AI SYNTHESIS DETECTED
DIFFUSION ERA
HYPER-REALISTIC SYNTHESIS – FLAWLESS, HIGH-FIDELITY OUTPUT
LEGEND: [CYAN] SYSTEM STATUS / LABEL, [WHITE] ANNOTATION / DATA, [ORANGE] WARNING / DETECTED ARTIFACT, [SECONDARY CYAN] TECHNICAL DETAIL

The Threat Vector

- **Accessibility:** Open-source tools lower the barrier — anyone with a GPU can generate convincing fakes
- **Fraud & Evidence Tampering:** Synthetic images used in insurance fraud, legal disputes, identity theft
- **Disinformation:** Fabricated images of events, people, and documents spread at scale
- **The Liar's Dividend:** Real evidence dismissed as "probably AI-generated" — doubt weaponized

https://www.etsy.com/listing/1667241073/realistic-silicone-sixth-finger-prop

The Liar’s Dividend

- **The Concept —** Coined by law professors Chesney & Citron (2019): deepfakes don’t just enable fabrication — they give bad actors a plausible excuse to deny authentic evidence
- **The Inversion —** The danger is no longer just fake evidence being believed; it’s real evidence being dismissed
- **In the Courtroom —** A defendant can claim any genuine photo, video, or audio was AI-generated — and create reasonable doubt without proving it
- **The Erosion Effect —** As deepfake awareness grows among juries, so does generalized skepticism toward all digital evidence
- **Already Happening —** Defense attorneys and defendants have invoked “that could be AI” in response to legitimate evidence in real cases
- **The Paradox —** The better deepfake detection becomes, the more confidently bad actors deny authentic evidence knowing detection is imperfect
- **The Stakes —** Victims of real crimes may see their genuine evidence dismissed; guilty parties may escape accountability behind manufactured doubt

Visual Inspection

- **Unusual Details** — Asymmetrical features, odd finger placement, strange proportions
- **Texture/Pattern Repetition** — Unnatural repetition in hair, skin, clothing, or backgrounds
- **Lighting & Shadows** — Inconsistent light sources or shadows that don't match the scene
- **Background Anomalies** — Overly simple/complex backgrounds or out-of-place elements
- **Facial Features** — Odd eyes, ears, or hair; uncanny symmetry or asymmetry
- **Contextual Errors** — Objects out of place or mismatched scale within the scene
- **Garbled Text** — Jumbled, misspelled, or nonsensical text on signs or labels
- **Digital Artifacts** — Pixelation, strange color patterns, or illogical blur
- **Emotional Inconsistency** — Facial expressions that don't match the mood or context

Test Your Skill

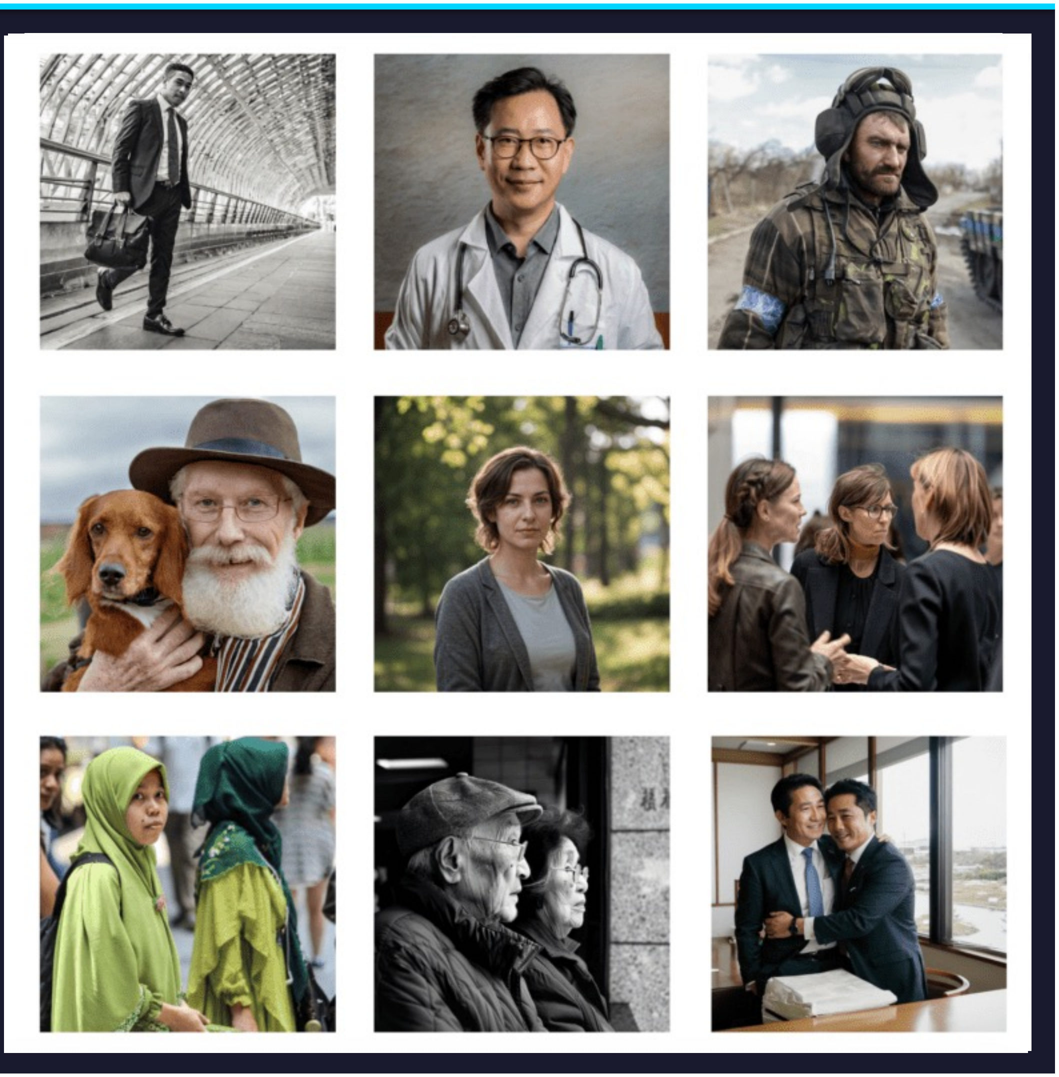


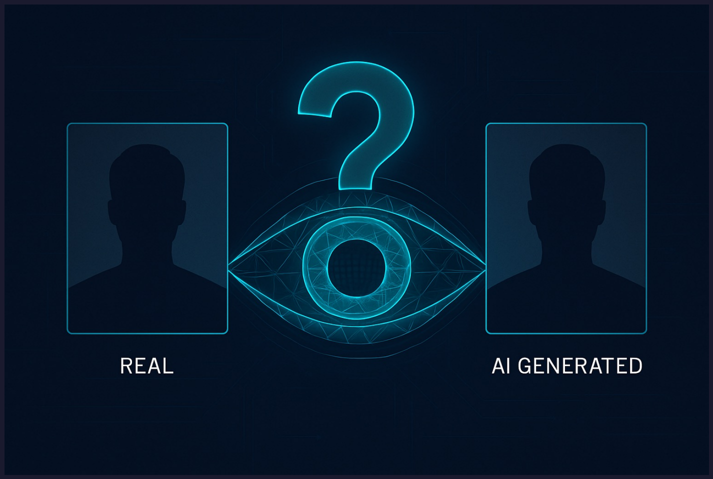
Image Grid from:
How to Distinguish AI-Generated Images from Authentic Photographs
Negar Kamali, Karyn Nakamura, Angelos Chatzimparmpas, Jessica Hullman, Matthew Groh

https://arxiv.org/abs/2406.08651

The Challenge

- State-of-the-art generators pass casual human review >90% of the time
- A layered forensic approach is required — no single technique is enough
- We need tools that examine what humans cannot see

Human visual inspection is no longer sufficient.



SECTION II

Hygiene & Metadata Analysis

The first line of defense: what the file tells you before you look at the pixels

Exif & File Signatures

Layer 1: What the metadata reveals

- **Software Tags:** Generative tools often leave telltale "Software" metadata (e.g., "DALL·E", "Stable Diffusion")
- **File Header Anomalies:** Synthetic images may have unusual compression profiles or missing camera-specific fields
- **Tool Spotlight — ExifTool:** Industry-standard CLI for extracting and analyzing all embedded metadata
  - **Quick wins:** Check Make, Model, Software, GPS, and date/time fields for inconsistencies
  - **Red flags:** Missing EXIF entirely, or camera fields present but no lens data

DEMONSTRATION

ExifTool Demonstration

DEMONSTRATION

Nanobanana Demonstration

Create a prompt for an LLM to create a realistic image that looks like it was shot by a Canon EOS Rebel T100 camera. The image should be of a sports car.

A sports car photographed with a Canon EOS Rebel T100 DSLR camera, 18-55mm kit lens at 35mm, f/5.6 aperture, ISO 400, 1/500s shutter speed. Natural daylight, slight lens vignetting, shallow depth of field with soft background bokeh. Slight chromatic aberration at edges, realistic sensor noise grain, JPEG compression artifact typical of consumer DSLR. The car is parked on a suburban street, three-quarter front angle, slightly overexposed highlights on the hood. Photorealistic, candid amateur photography style, not studio, not professional lighting. Shot by a casual photographer. 8K, RAW-style rendering.

Visual Forensics

Layer 1: Pixel-level inspection techniques

- **Lighting Consistency:** Check if light sources, highlights, and shadows are physically coherent across the scene
- **Error Level Analysis (ELA):** Re-compress at known quality and compare — manipulated regions show different error levels
- **Compression Artifacts:** JPEG quantization patterns differ between original captures and composites
- **Perspective & Geometry:** Vanishing points, reflections, and proportions that violate physical reality

Limitations of Metadata Analysis

Metadata can be stripped, forged, or simply absent.

- Social media platforms strip EXIF data on upload — most shared images have no metadata
- Sophisticated actors deliberately craft fake metadata to mislead investigators
- Metadata presence proves nothing by itself — it must be corroborated
- We need techniques that analyze the image content itself, not just the wrapper

https://lucidtruthtechnologies.com/photo-metadata/

Error Level Analysis

Detecting Compression Anomalies in Digital Images

What Is ELA?

- Re-saves the image at a known JPEG quality level
- Computes the pixel-level *difference* between re-saved and original
- Amplifies that difference into a visible map
- Uniform regions = consistent compression history (authentic)
- Bright/hot regions = higher error = different origin or recent modification

What to Look For:

🔴 **Hot edges around objects** → composite insertion

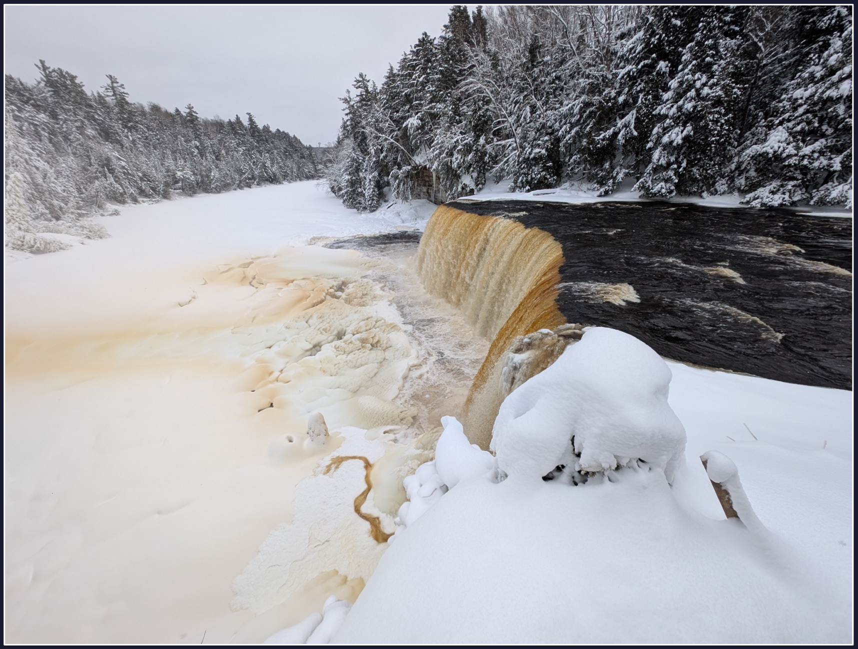
🔴 **Inconsistent error levels** between foreground and background → different source materials

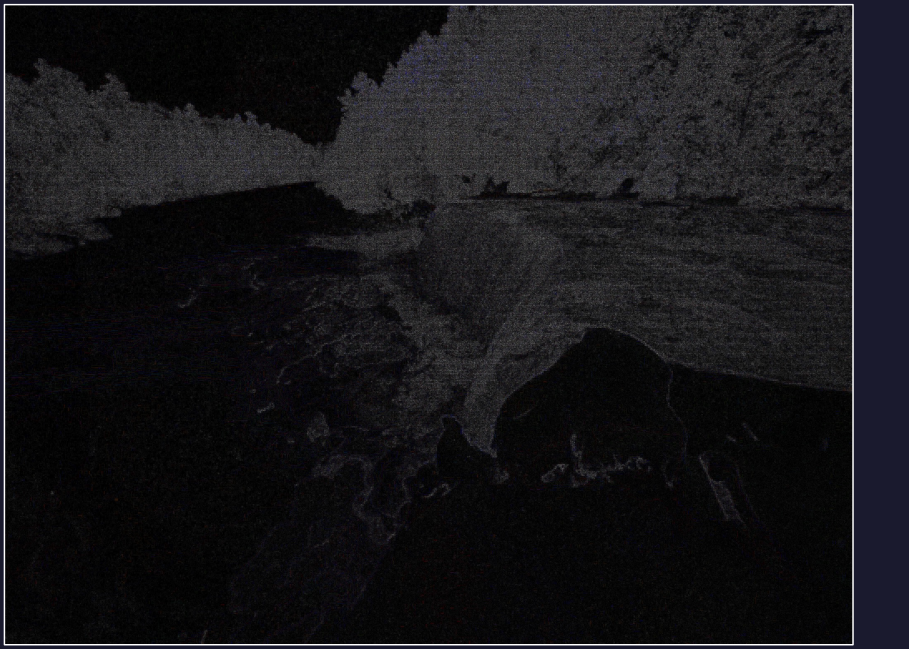
🔴 **Uniform darkness across entire image** → heavily recompressed (provenance concern)

🟢 **Consistent, low-level texture throughout** → authentic single-capture image

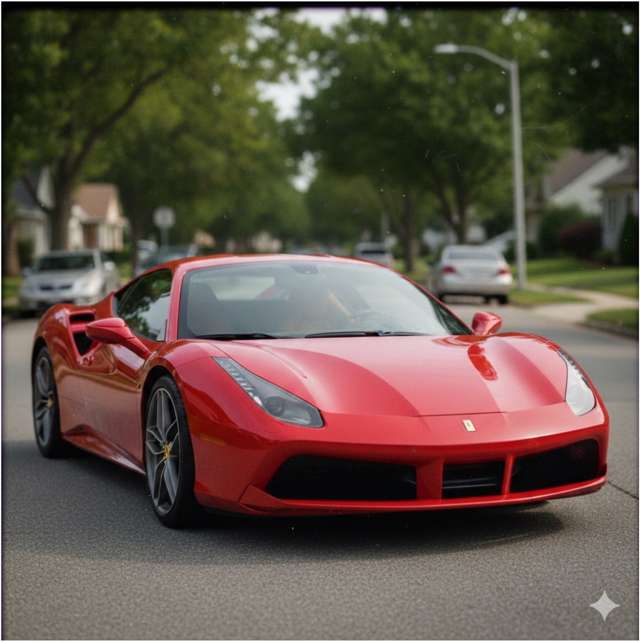
https://29a.ch/photo-forensics/#error-level-analysis

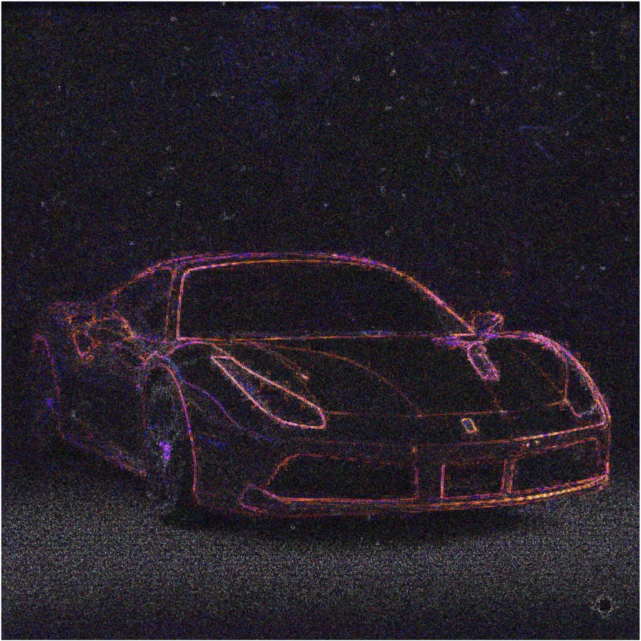
ELA – Known Good Image (Pixel)





ELA – Nano Banana Image





Limitations of Error Level Analysis

ELA is a triage tool, not a definitive test — understand where it helps and where it falls short.

Sources of False Positives

🔴 **Social media recompression** — platforms like Instagram and Facebook re-encode on upload, flattening authentic images and creating misleading results

🔴 **Multiple legitimate saves** — a real photo edited in Photoshop and saved several times will show elevated error levels even if unmanipulated

🔴 **High-contrast content** — edges, text overlays, and fine textures naturally produce higher ELA activity in authentic images

🔴 **Format conversion** — PNG-to-JPEG conversion resets compression history, defeating ELA entirely

Practical Guidance

✅ Always request **original, unmodified files** in discovery — never rely on social media downloads or screenshots

✅ ELA works best on **JPEG images** — largely ineffective on PNG (lossless compression)

✅ Use as **triage, not conclusion** — ELA raises questions; PRNU and expert analysis answer them

✅ **Tools:** 29a.ch/photo-forensics (free triage) · FotoForensics & Amped Authenticate (court-grade documentation)

ELA can open a door for discovery. It cannot close a case on its own.

AI Watermarking: Invisible Signatures from the Source

When the Generator Marks Its Own Output

- **Google SynthID —** Imperceptible pixel-level watermark; survives cropping, resizing, screenshots, and print/re-photograph cycles
- **Google Visible Watermark —** Four-pointed star icon in the bottom-right corner of Gemini-generated images; trivially removed with any photo editor
- **Meta Invisible Watermarking —** Frequency-domain embedding on Imagine-generated images; detectable after heavy editing
- **Meta Stable Signature —** Watermark baked into the model decoder itself; every output pre-watermarked at the architecture level
- **OpenAI / DALL·E 3 —** No pixel-level watermark; relies on metadata disclosure only
- **The Critical Gap —** Open-source models (Stable Diffusion, Flux, ComfyUI) watermark nothing — and that’s exactly where evidentiary risk lives

SECTION III

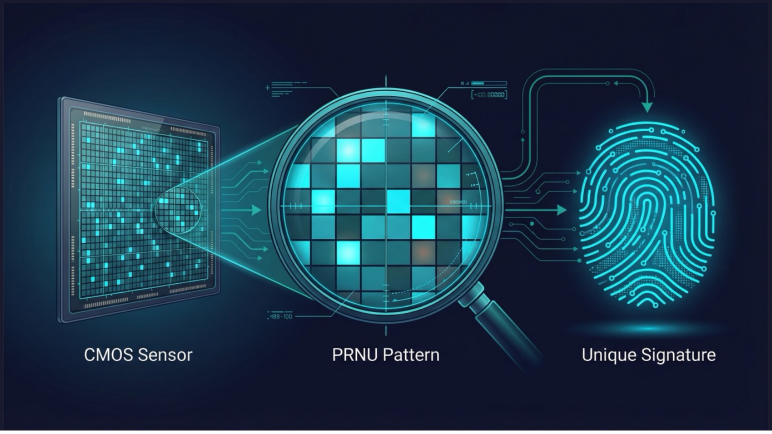
Camera Ballistics — The "Gold Standard"

Using sensor fingerprints to authenticate the origin of an image

Sensor Fingerprinting: PRNU

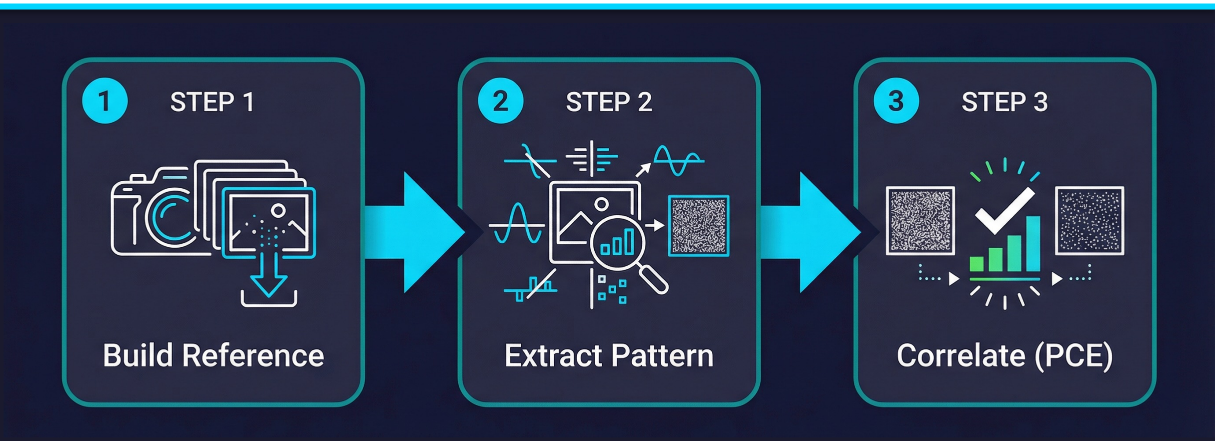
Layer 2: The "digital fingerprint" of a camera sensor

- **Photo Response Non-Uniformity:** Unique "digital bullet scratches" from manufacturing imperfections in the sensor
- **Every Camera Is Unique:** Even identical make/model cameras produce different PRNU patterns
- **Persistent & Involuntary:** The pattern is embedded in every photo the sensor captures — it cannot be turned off
- **Analogy:** Like matching a bullet to a specific gun barrel — hence "camera ballistics"



PRNU Methodology

Layer 2: Extraction, correlation, and scoring



1 STEP 1

Build Reference

2 STEP 2

Extract Pattern

3 STEP 3

Correlate (PCE)

- **Step 1 — Build Reference Fingerprint:** Average the noise residuals from multiple flat-field images taken by the suspect camera
- **Step 2 — Extract Questioned Pattern:** Isolate the noise residual from the image under investigation
- **Step 3 — Correlate:** Compute Peak-to-Correlation Energy (PCE) between reference and questioned patterns
- **High PCE Score:** Strong evidence the image came from that specific camera
- **Low/No PCE Score:** Image did NOT come from that camera — potentially synthetic

Forensic Validity of PRNU

Layer 2: Why courts accept this evidence

- **Peer-Reviewed Science:** Decades of published research supporting PRNU-based identification
- **Daubert Standard:** Meets the legal threshold for admissibility of scientific evidence in U.S. courts
- **Court-Tested:** Successfully used in criminal cases for source identification
- **Key Application:** Proving an image did NOT come from a specific camera flags it as potentially synthetic
- **Limitation:** Requires access to the suspected source camera (or images from it) to build the reference fingerprint

DEMONSTRATION

PRNU Demonstration

SECTION IV

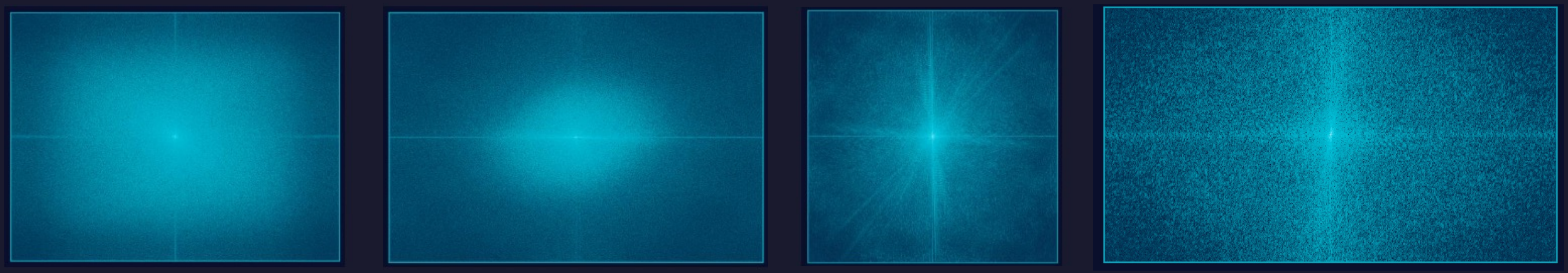
Detecting the Ghost in the Machine

AI-specific forensic techniques that identify synthetic origins

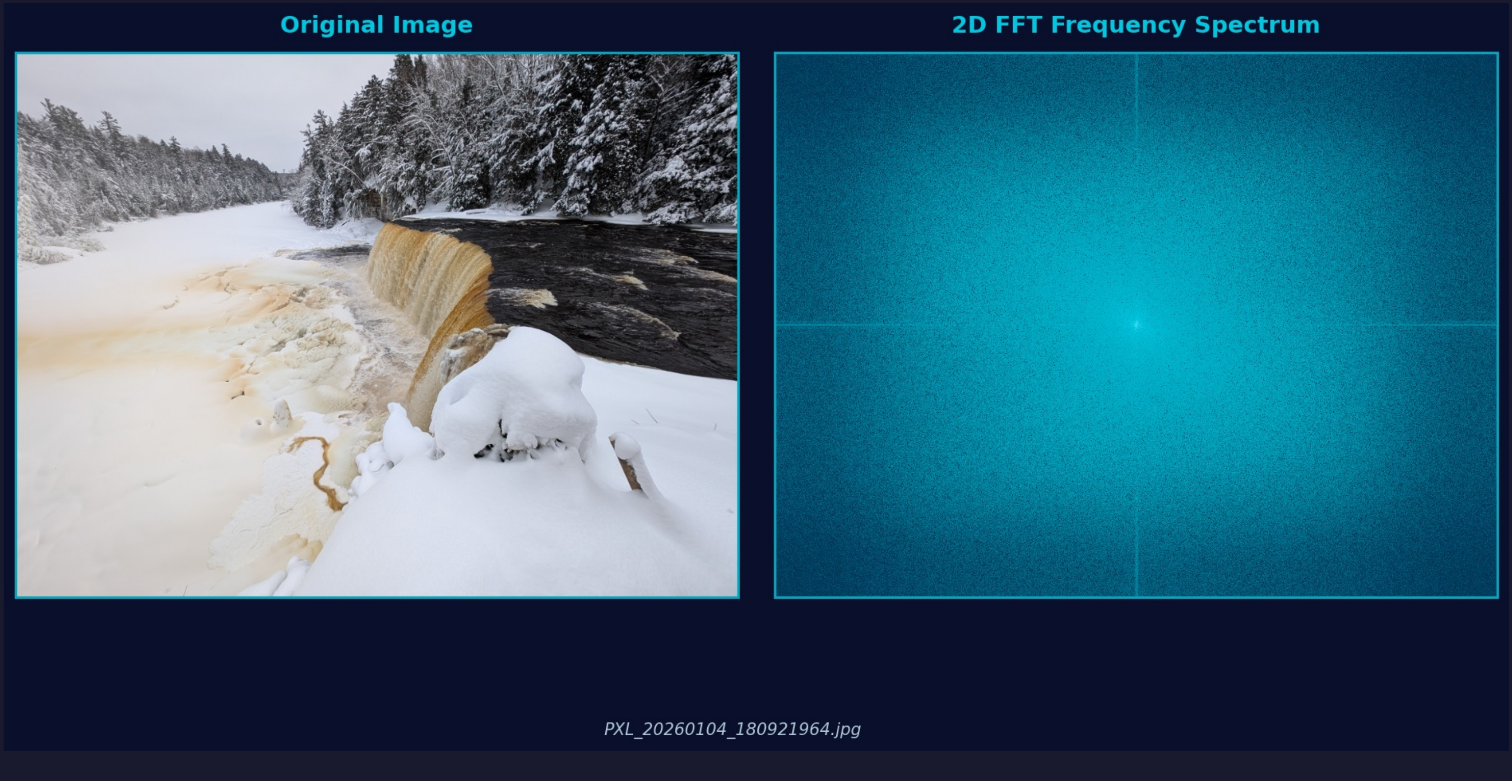
Frequency Domain Analysis

Layer 3: What the frequency spectrum reveals

- **Fourier Transform:** Convert spatial image data to frequency domain to reveal hidden patterns
- **GAN Artifacts:** Generators often produce unnatural grid patterns or periodic artifacts invisible to the naked eye
- **High-Frequency Signatures:** Real cameras produce characteristic noise profiles — synthetic images do not match
- **Spectral Anomalies:** Checkerboard patterns in the frequency domain are a strong indicator of GAN-generated content



Frequency Domain Analysis – Pixel 8

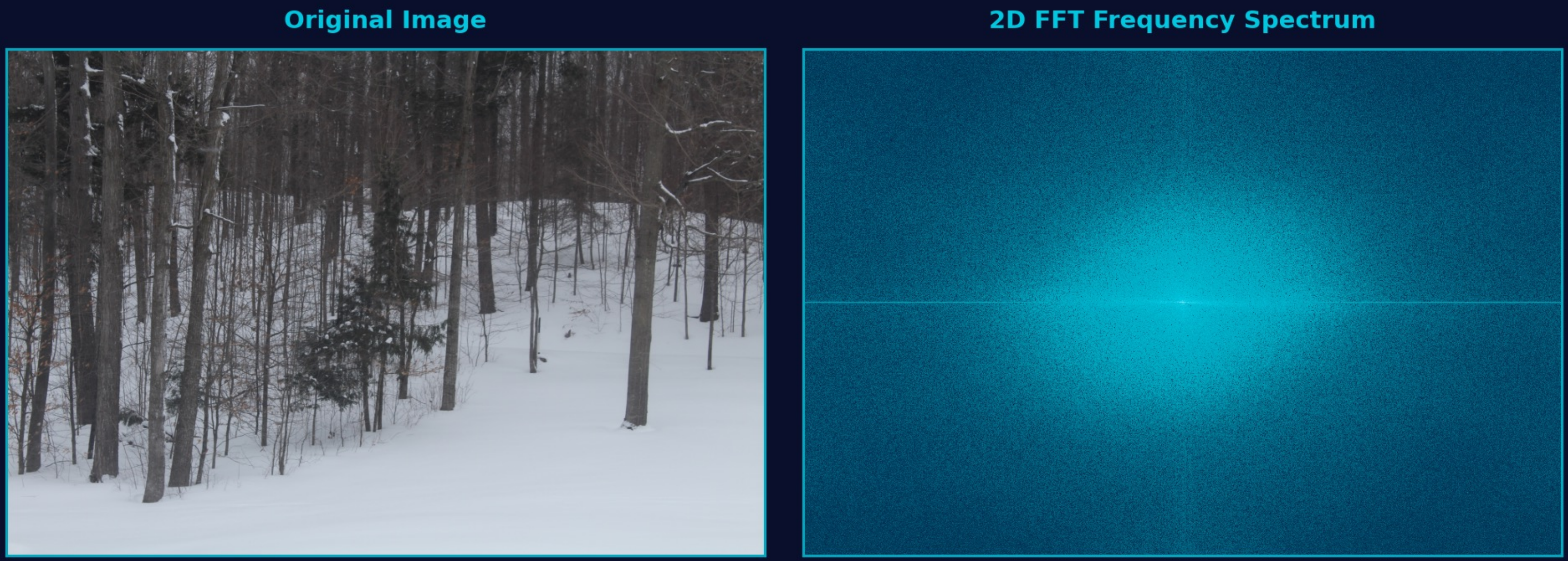


Original Image

2D FFT Frequency Spectrum

PXL_20260104_180921964.jpg

Frequency Domain Analysis – Canon EOS T100

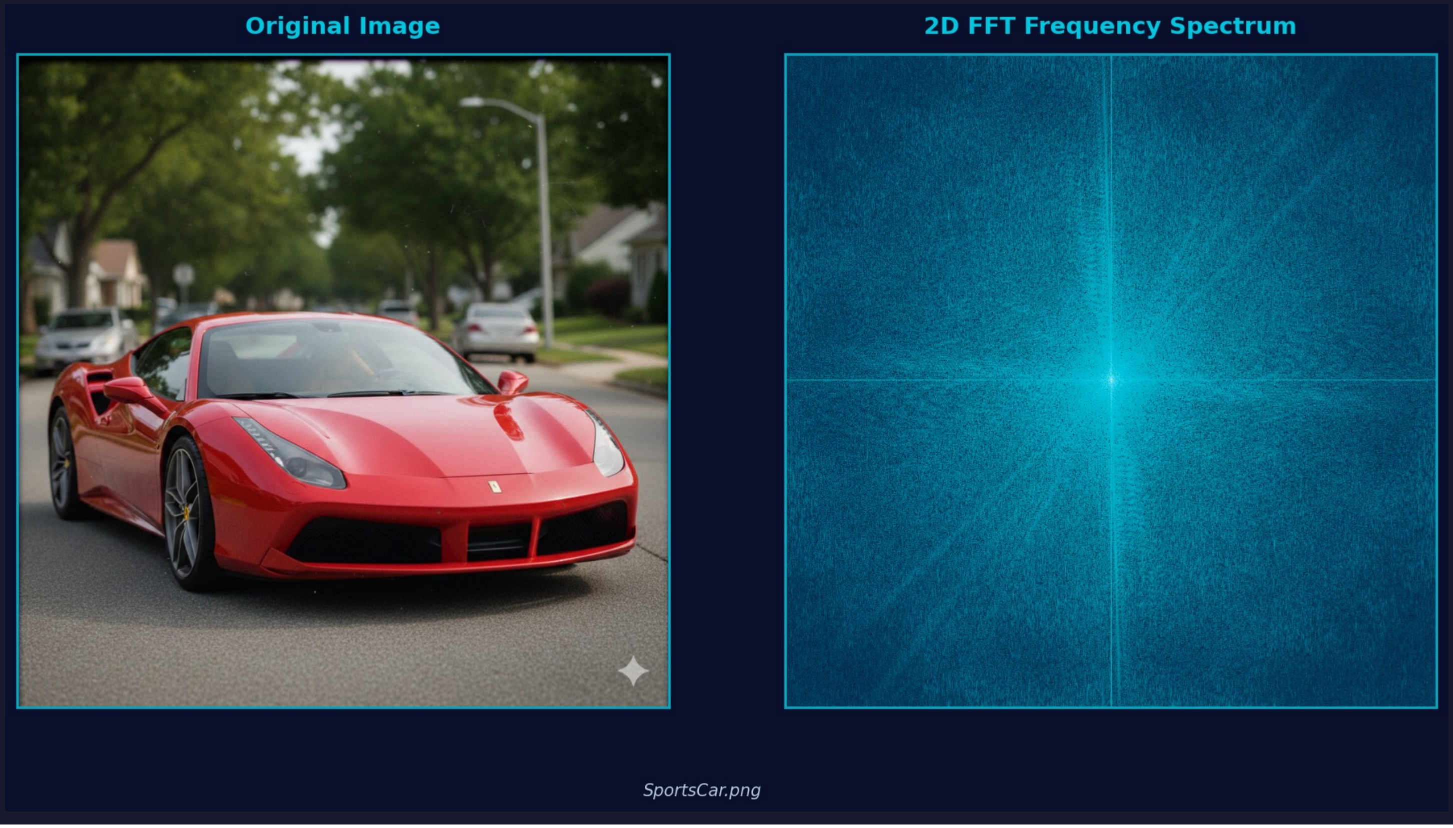


Original Image

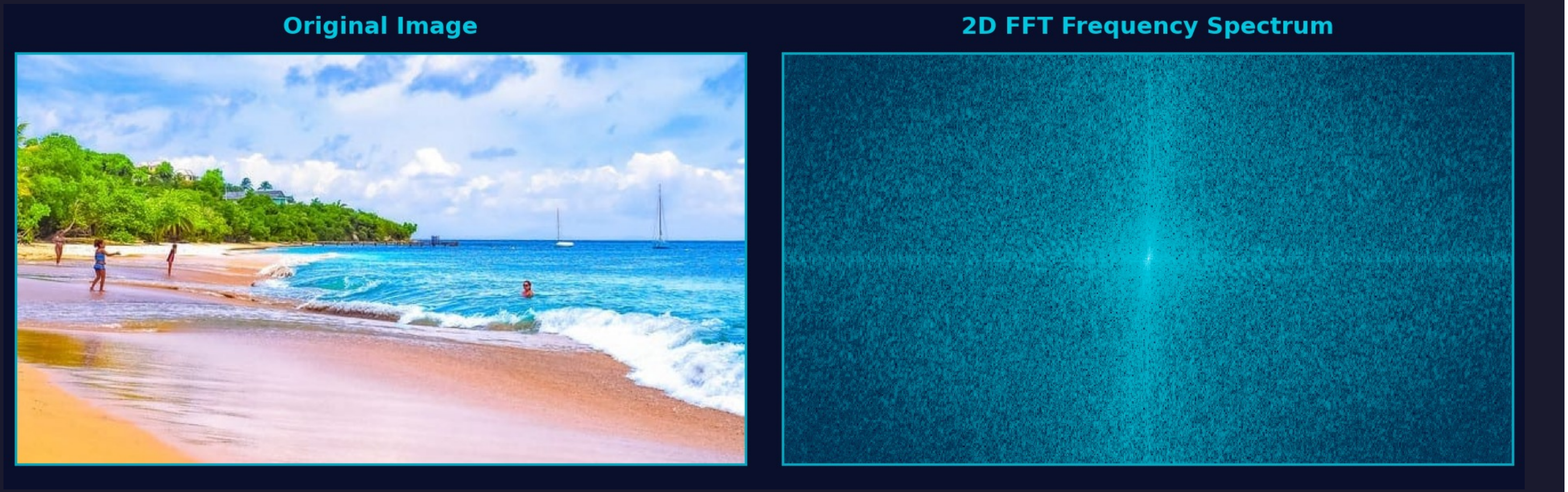
2D FFT Frequency Spectrum

IMG_0014.JPG

Frequency Domain Analysis – Diffusion Model (Nano Banana)



Frequency Domain Analysis – GAN Model (StyleGAN)



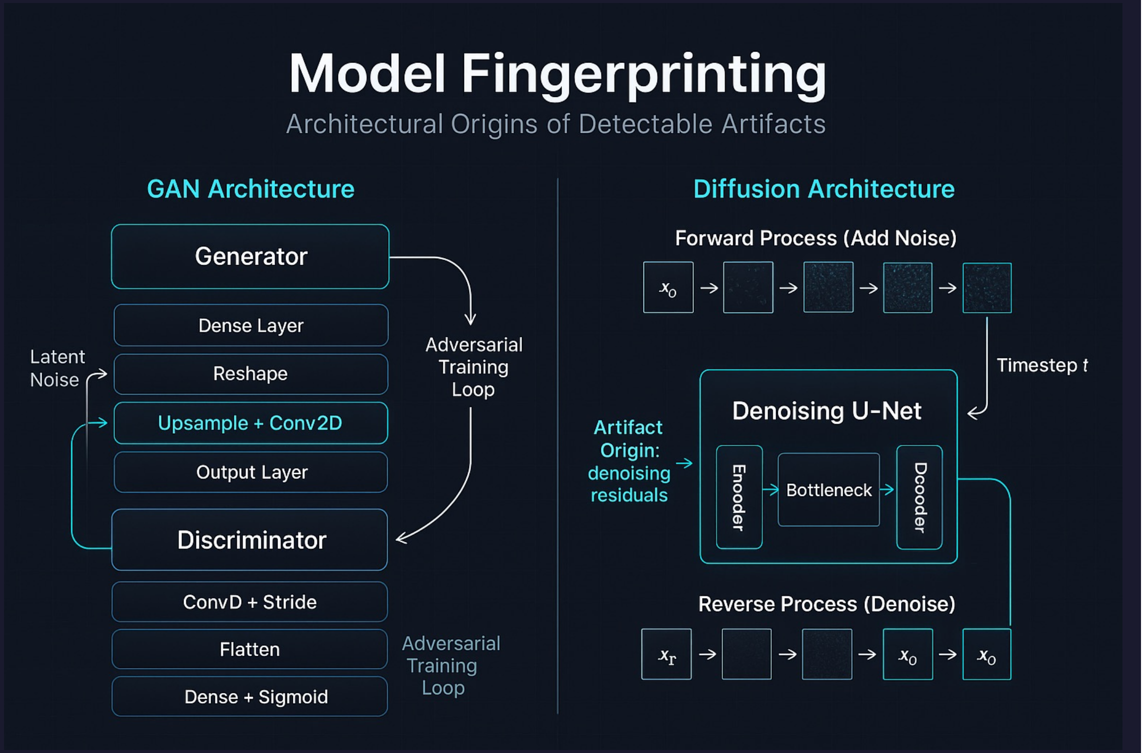
Model Fingerprinting

GANs

- Statistical residuals from generator /discriminator training
- Characteristic upsampling artifacts
- Mode-specific spectral signatures

Diffusion Models

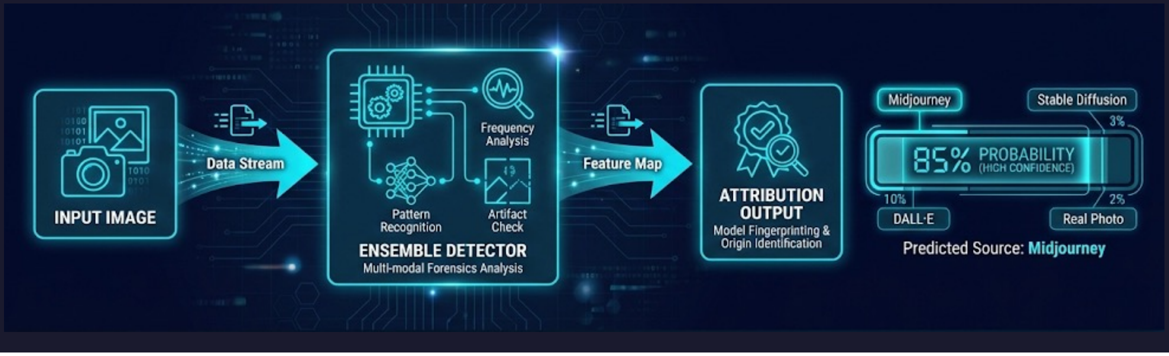
- Distinct denoising step artifacts
- Different statistical distributions in noise residuals
- Scheduler-dependent patterns



Automated Detectors & Attribution

Layer 3: Scaling detection with machine learning

- **CLIP-Based Models:** Vision-language approaches that generalize across unseen generators
- **Model Attribution:** Classifying images to specific tools (Midjourney v5 vs. DALL·E 3 vs. Stable Diffusion)
- **Transfer Learning:** Detectors trained on known generators can partially generalize to new ones
- **Ensemble Approaches:** Combining multiple detection signals improves accuracy and reduces false positives
- **Challenge:** New model architectures require continuous retraining of detectors



The Black Box Problem

When AI Detects AI — Legal Admissibility Challenges

- **Opaque Decision-Making —** AI detectors produce a probability score, not an explanation; even developers cannot fully articulate why a specific image triggered a result
- **Daubert Challenge —** Without transparent methodology, AI-based detection evidence is vulnerable to admissibility challenges under Rule 702
- **Known Error Rate Problem —** Published accuracy figures reflect controlled benchmarks, not real-world compressed, screenshot, or re-shared images
- **Non-Determinism —** Some AI systems produce different outputs on identical inputs depending on hardware, software version, or random seed
- **Lack of Reproducibility —** If opposing counsel cannot independently replicate the result, the methodology fails a basic forensic reliability test
- **The Defense Strategy —** Demand the model version, training dataset, validation methodology, and real-world false positive rate before any AI detector result is admitted

Ask a Large Language Model

Now that you know the caveats…

Analyze this image and determine whether it is AI-generated or captured by a real camera. Examine the following forensic indicators and provide a detailed rationale for your conclusion:

- Texture and surface quality — Are textures organic and naturally varied, or do they show signs of over-smoothing, repetition, or synthetic uniformity?
- Lighting and shadows — Are light sources physically consistent? Do shadows fall correctly relative to the apparent light direction?
- Fine detail rendering — Examine edges, hair, fur, fabric, water, or other complex surfaces. Do they show natural irregularity or synthetic perfection?
- Frequency characteristics — Does the image appear to have the natural noise grain of a camera sensor, or does it have the smooth, noiseless quality typical of diffusion model output?
- Contextual coherence — Do all elements in the scene make physical and spatial sense?
- Facial features — If faces are present, examine eyes, ears, teeth, and hairlines for uncanny symmetry, impossible reflections, or subtle anatomical errors.
- Text and labels — If any text appears in the image, is it legible, correctly spelled, and contextually appropriate?
- Background integrity — Does the background show natural complexity, or does it appear artificially simple?
- Compression and artifact signature — Does the image show natural JPEG compression patterns consistent with a camera?
- Watermarking indicators — Examine the image for any visible AI disclosure markers.

Based on your analysis, state your conclusion: Real photograph or AI-generated, along with a confidence level (Low / Medium / High) and a summary of the two or three most compelling indicators.

SECTION V

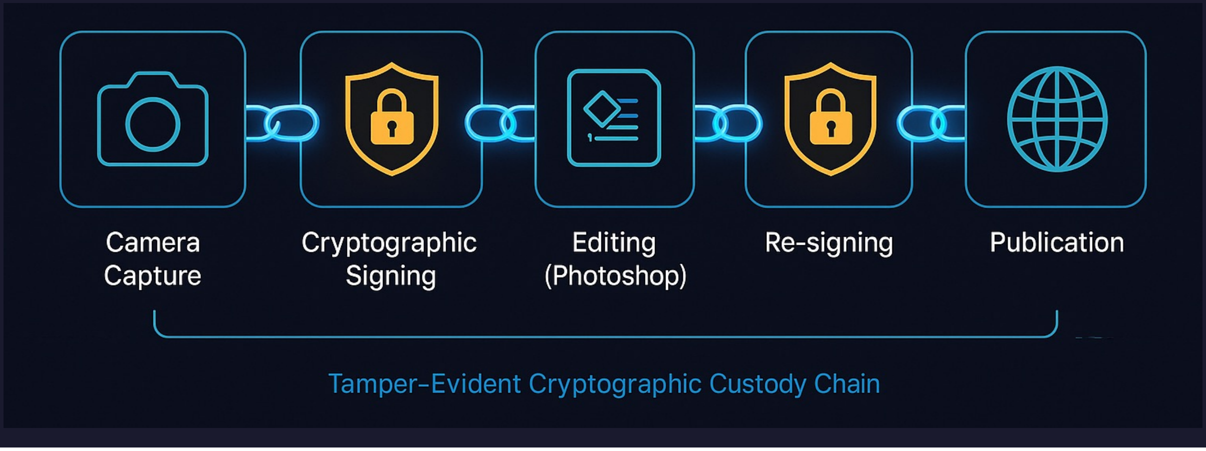
The Future of Authentication (C2PA)

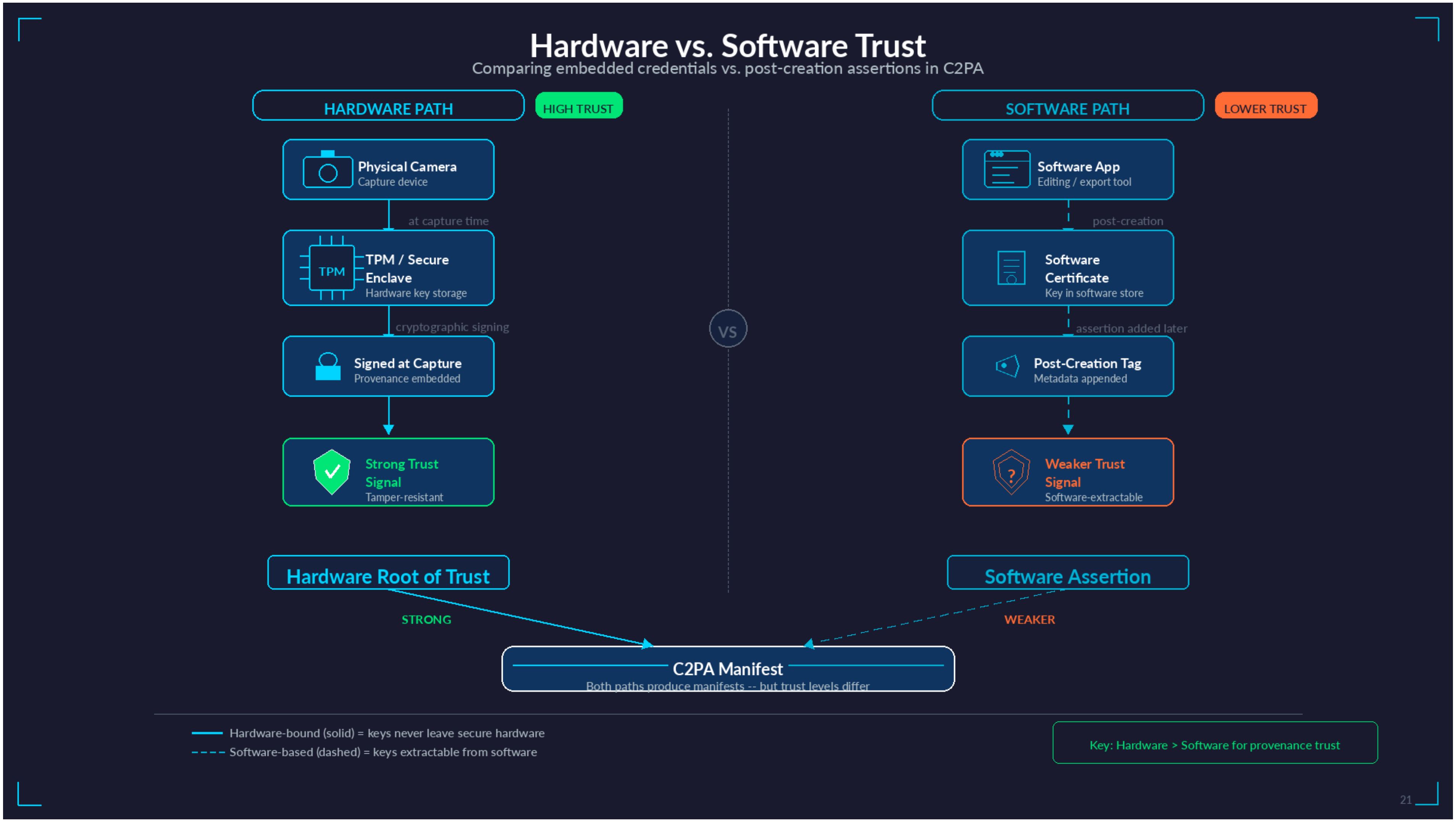
From detecting fakes to proving reality

Content Credentials (C2PA)

Shift the paradigm: prove an image is REAL rather than trying to prove it is FAKE.

- **Coalition for Content Provenance and Authenticity:** Industry standard for signed metadata
- **Cryptographic Signatures:** Tamper-evident chain of custody from capture to publication
- **Manifest System:** Records every edit, export, and transformation with cryptographic proof





Hardware vs. Software Trust

Comparing embedded credentials vs. post-creation assertions in C2PA

HARDWARE PATH

HIGH TRUST

Physical Camera
Capture device

at capture time

TPM / Secure Enclave
Hardware key storage

cryptographic signing

Signed at Capture
Provenance embedded

Strong Trust Signal
Tamper-resistant

SOFTWARE PATH

LOWER TRUST

Software App
Editing / export tool

post-creation

Software Certificate
Key in software store

assertion added later

Post-Creation Tag
Metadata appended

Weaker Trust Signal
Software-extractable

Hardware Root of Trust

STRONG

Software Assertion

WEAKER

C2PA Manifest
Both paths produce manifests -- but trust levels differ

— Hardware-bound (solid) = keys never leave secure hardware
- - - Software-based (dashed) = keys extractable from software

Key: Hardware > Software for provenance trust

Hardware vs. Software Trust

📷 Camera-Embedded

- Credentials signed at moment of capture
- Hardware-rooted trust (TPM/secure enclave)
- Unbroken chain of custody from sensor to file
- **Manufacturers:** Leica, Sony, Nikon
- Highest security assurance

💻 Software-Added

- "Digital labels" added post-creation
- **Applications:** Photoshop, Firefly, DALL·E
- Offers transparency about edits and AI use
- Lower security assurance — no hardware root of trust
- Still valuable for disclosure and chain of edits

C2PA Tool: Validating Content Credentials

Layer 4: Practical tools for cryptographic verification

- **C2PA Tool CLI:** Open-source command-line tool for inspecting and validating manifests
- **Validate Signatures:** Verify that credentials have not been tampered with
- **Inspect Trust Chain:** Trace the full provenance from creation through every transformation
- **Check Assertions:** Review what the signer claims about the content (camera capture, AI-generated, edited)
- **Integration:** Can be incorporated into automated verification workflows and content management systems

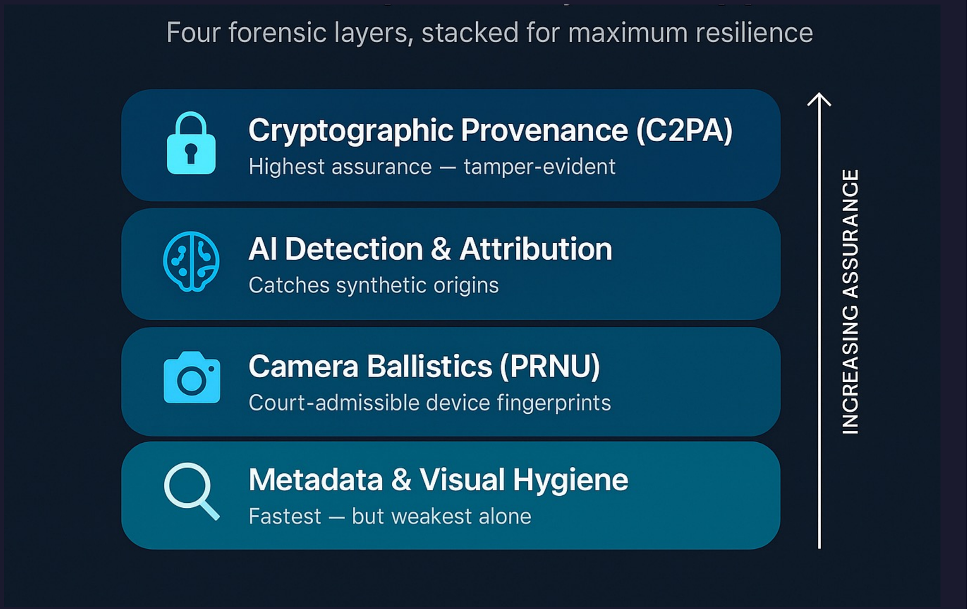
SECTION VI

Conclusion & Practical Takeaways

Assembling your forensic toolkit

Defense in Depth: A Layered Approach

- **Layer 1 — Metadata & Visual Hygiene:** Quick triage with ExifTool and ELA (fast but bypassable)
- **Layer 2 — Camera Ballistics (PRNU):** Court-admissible source identification (requires reference camera)
- **Layer 3 — AI Detection & Attribution:** Frequency analysis, model fingerprinting, CLIP classifiers
- **Layer 4 — Cryptographic Provenance (C2PA):** Hardware-signed credentials prove authenticity at capture
- **No Single Layer Is Sufficient:** Combine multiple techniques for robust verification



Four forensic layers, stacked for maximum resilience

Cryptographic Provenance (C2PA)
Highest assurance — tamper-evident

AI Detection & Attribution
Catches synthetic origins

Camera Ballistics (PRNU)
Court-admissible device fingerprints

Metadata & Visual Hygiene
Fastest — but weakest alone

Key Takeaways

- **The Cat-and-Mouse Game:** As detectors improve, generative models will adapt to hide forensic traces
- **Layered Defense Is Non-Negotiable:** No single tool catches everything — combine metadata, PRNU, AI detection, and C2PA
- **Legal Readiness:** PRNU meets Daubert; C2PA creates auditable chains of custody — both are court-relevant
- **Invest in C2PA Infrastructure:** Hardware-signed credentials are the strongest long-term defense
- **Stay Current:** Detection tools and generative models evolve rapidly — continuous learning is essential

QUESTIONS & DISCUSSION

Legal Admissibility • Enterprise Security • Implementation

ABOUT THE SPEAKER

LUCID TRUTH TECHNOLOGIES®

Kenneth G. Hartman

GSE, CISSP, GCFA, GCFE, GASF

- **Founder, Lucid Truth Technologies:** Digital forensics consultancy specializing in contested digital evidence for criminal defense attorneys throughout Michigan
- **Licensed Private Investigator:** State of Michigan License No. 3701207402; legally authorized to conduct forensic investigations and generate court-admissible work product
- **GIAC Security Expert (GSE #198):** Fewer than 250 holders worldwide; the most rigorous certification in information security
- **SANS Certified Instructor:** Currently teaching SEC502: Cloud Security Tactical Defense; prior security leadership at Google, Illumina, and SAP Ariba
- **Published Researcher:** Three-part blog series on AI image authentication; GIAC Gold papers on Amazon EC2 forensics and BitTorrent investigations; SANS Reading Room contributor since 2014
- **Criminal Defense Specialist:** CDITC certified; Associate member of CDAM, Michigan Council of Professional Investigators, and National Association of Legal Investigators