arXiv:2602.03423v2 [cs.CR] 29 Oct 2026

Origin Lens: A Privacy-First Mobile Framework for Cryptographic Image Provenance and AI Detection

Alexander Loth
Frankfurt University of Applied Sciences
Frankfurt am Main, Germany
Doctoral Center for Applied Computer Science (PZAI)
Wiesbaden, Germany
alexander.loth@stud.fra-uas.de

Dominique Conceicao Rosario
Peter Ebinger
Martin Kappes
dominique.conceicaorosario@stud.fra-uas.de
peter.ebinger@fra-uas.de
kappes@fra-uas.de
Frankfurt University of Applied Sciences
Frankfurt am Main, Germany

Marc-Oliver Pahl
IMT Atlantique, UMR IRISA, Chaire Cyber CNI
Rennes, France
marc-oliver.pahl@imt-atlantique.fr

Abstract

The proliferation of generative AI poses challenges for information integrity assurance, requiring systems that connect model governance with end-user verification. We present **Origin Lens**, a privacy-first mobile framework that targets visual disinformation through a layered verification architecture. Unlike server-side detection systems, Origin Lens performs cryptographic image provenance verification and metadata-based AI detection locally on the device via a Rust/Flutter hybrid architecture. Our system integrates multiple signals—including cryptographic provenance, generative-model metadata and watermarks, and optional retrieval-augmented verification—to provide users with graded confidence indicators at the point of consumption. We discuss the framework's alignment with regulatory requirements (EU AI Act, DSA) and its role in verification infrastructure that complements platform-level mechanisms.

CCS Concepts

• **Security and privacy** → **Digital rights management**; • **Computing methodologies** → *Computer vision*; • **Information systems** → *Multimedia information systems.*

Keywords

Information Integrity, Image Provenance, C2PA, AI Detection, Misinformation Detection, Privacy-First, Mobile Framework, Generative AI, AI Safety

1 Introduction

Generative AI has introduced a *verification asymmetry* in digital societies: while AI models can generate synthetic media at scale, individual users lack tools to assess content authenticity at the point of consumption. Our prior survey on generative AI and fake news documents this dual-use nature of LLMs—enabling new disinformation capabilities while offering potential detection solutions [20]. As noted by Régis et al., the intersection of AI and democratic processes requires technical and governance interventions [33]. This challenge spans multiple dimensions of AI safety research—from *Model Design* (transparency mechanisms like watermarking) to *Model Ecosystem* governance (regulatory compliance and infrastructure).

Existing approaches to misinformation mitigation predominantly rely on centralized, platform-level interventions [12]. However, this creates dependencies on opaque moderation systems and raises concerns about surveillance and data harvesting [1].

This paper introduces **Origin Lens**, an open-source, privacy-first mobile framework for cryptographic image provenance and AI detection that builds on the C2PA (Coalition for Content Provenance and Authenticity) standard [3] alongside heuristic AI detection.[1] The framework aims to make provenance verification accessible to non-expert users. Our contribution is threefold: (1) a privacy-preserving architecture that performs cryptographic and metadata verification entirely on-device, (2) a defense-in-depth verification pipeline combining cryptographic, heuristic, and contextual signals with graded confidence indicators, and (3) a discussion of how client-side verification tools can complement platform governance.

2 Related Work

Recent work in information integrity has focused on robustness evaluation and scalable detection. The *OpenFake* dataset [17] provides deepfake benchmarks, while *Veracity* [6] demonstrates retrieval-augmented fact-checking. Thibault et al. audit misinformation detection datasets and evaluation practices [37]. Technical detection must be paired with effective human-AI interaction. Wang et al. show that psychological inoculation has limited effectiveness in a realistic social media feed [39]. Our JudgeGPT/RogueGPT studies[2] reveal a perception-accuracy gap [21, 22], and empirical analysis shows cognitive fatigue degrades fake content detection by 10.2 percentage points [22]. Puelma Touzel et al. [30] and Ghafouri et al. [10] observe that information flow complexity and uncertainty

This research is funded by the European Union project CyberSecDome under the grant agreement 101120779. This paper was peer-reviewed and accepted at the Future of Information Integrity Research (FIIR) Workshop at The Web Conference 2026 (https://sites.google.com/view/fiirworkshop); the workshop was subsequently cancelled and the paper was not included in the proceedings. A separate, independent paper on Origin Lens appeared at ACM WebSci '26 [25]. Origin Lens is part of the first author's doctoral research on AI-generated disinformation [19].



This work is licensed under a Creative Commons Attribution 4.0 International License.

[1] Source code: https://github.com/aloth/origin-lens; App Store: https://apps.apple.com/app/id6756628121; other platforms on the roadmap.
[2] https://github.com/aloth/JudgeGPT; https://github.com/aloth/RogueGPT

communication affect trust calibration. Article 50(2) of the EU AI Act [9] mandates machine-readable marking of AI-generated content, while the Digital Services Act (DSA) [7] requires platforms to implement content moderation transparency. Human judges rate text from current LLMs as nearly human-written (mean HumanMachineScores of 0.46–0.50) [23]. The C2PA standard [3] enables decentralized provenance verification, but most tools remain server-dependent. Origin Lens provides a privacy-first, client-side alternative.

3 System Architecture

Origin Lens employs a hybrid mobile architecture optimized for performance and memory safety (see Figure 1). The core logic is implemented in **Rust** [28] for its memory safety guarantees [15] when parsing complex binary formats, while the user interface is built with **Flutter** [11] for cross-platform mobile deployment.

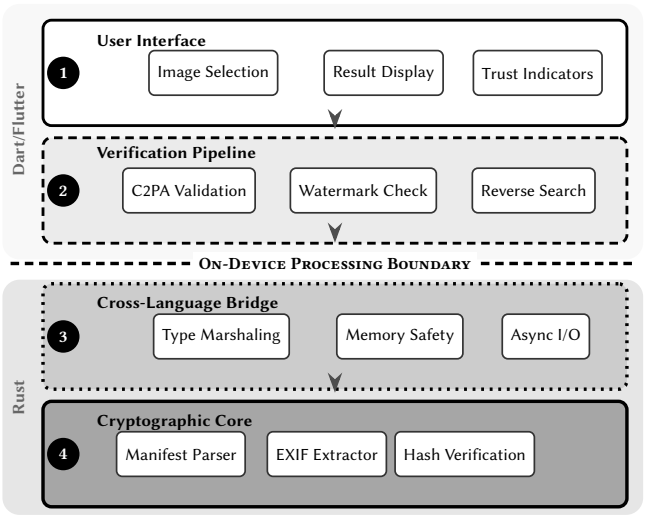


Figure 1: Origin Lens architecture: Flutter UI, verification pipeline, FFI bridge, and Rust core. Cryptographic operations execute on-device below the processing boundary.

Defense-in-Depth Strategy. Given epistemic uncertainty in modern media [10], we implement a four-layer pipeline: (1) *Cryptographic Provenance*—parse JUMBF (JPEG Universal Metadata Box Format) boxes to validate C2PA manifests, verify X.509 trust chains, and compute SHA-256 hashes ensuring hard binding [3, 16, 41]; (2) *Heuristic Metadata*—analyze EXIF and IPTC photo metadata for generative model artifacts (e.g., Stable Diffusion parameters); (3) *Watermark Detection*—detect imperceptible watermarks (e.g., SynthID) via API [4, 13, 14]; (4) *Contextual Verification*—opt-in reverse image search for prior attributions.

Privacy and Uncertainty Communication. In contrast to cloud-based detection services, Origin Lens processes C2PA and metadata entirely on-device via FFI (Foreign Function Interface) bindings, following Privacy by Design principles [2] with privacy as the default. Cryptographic provenance and heuristic metadata analysis require no network access. Watermark detection (e.g., SynthID) sends the image to an external API and runs automatically when no C2PA manifest is found. Contextual verification (reverse image search) transmits image data to external services, leaving digital traces; this feature is therefore strictly opt-in and disabled by default, aligning with GDPR data protection by design and by default (Article 25) and data minimization (Article 5(1)(c)) [8].

Communicating uncertainty is a known challenge [10]. Origin Lens implements a hierarchical confidence model: *high* (valid C2PA with trusted root), *medium* (EXIF patterns, watermarks), and *low* (opt-in reverse image search requiring user interpretation).

4 Security & Threat Modeling

We applied STRIDE threat modeling [34] to the Origin Lens verification pipeline, analyzing each architectural layer (Figure 1) for potential attack vectors.

Identified Threats. At the *Verification Pipeline* layer, we identify two primary risks: (1) *Manifest Stripping* (Tampering)—adversaries remove C2PA metadata during redistribution, a known limitation of content credentials [3]; and (2) *Certificate Spoofing* (Spoofing)—attackers forge X.509 certificates to inject false provenance claims. At the *Cross-Language Bridge*, malformed input could trigger memory corruption [29]; Rust's ownership model mitigates this at the *Cryptographic Core*. The *User Interface* faces Information Disclosure risks if verification results are cached insecurely.

Mitigations. Origin Lens enforces hard binding validation [3]: SHA-256 hashes bind the manifest to image content—any modification invalidates the credential. Against spoofing, signer certificate chains are validated through the c2pa-rs reference implementation.

Open Challenges. The analog hole—screen capture—circumvents cryptographic provenance [32]. We address this through defense-in-depth: heuristic metadata and watermark detection provide secondary signals when manifests are absent. Ecosystem adoption remains critical for verification coverage [3].

5 Evaluation & Discussion

On iOS (iPhone 15 Pro), C2PA validation completes in under 500 ms for 12 MP images, with EXIF parsing under 50 ms. This latency is acceptable for interactive use. For result communication, Origin Lens uses a traffic-light UI [36, 40]: *green* (valid C2PA from trusted root), *purple* (C2PA/EXIF indicating generative origin), *red* (hash mismatch or broken chain), and *orange* (no manifest found, expired certificate, or parsing issue).

Limitations. C2PA effectiveness depends on ecosystem adoption [3]. Adversarial actors may employ manifest stripping or analog-hole attacks. Heuristic detection faces distribution shift as models evolve [5, 26, 27, 35, 37, 38], and demographic predictors show weaker effects for AI-generated content [22]. Ultimately, technical tools require a complementary culture of verification and data literacy to achieve societal impact [18, 31].

6 Conclusion and Future Directions

Origin Lens provides an open-source, privacy-first implementation for on-device image provenance verification. By performing verification locally with graded uncertainty signals, the framework complements platform-level mechanisms. Future work includes

lightweight neural networks for pixel-based detection, privacy-preserving federated aggregation, and cross-jurisdictional provenance standards. As our research continues, we invite experts to participate in our survey on verification practices [24]: https://github.com/aloth/verification-crisis.

References

[1] Hannah Bloch-Wehba. 2022. Content Moderation as Surveillance. *Berkeley Technology Law Journal* 36, 3 (2022), 1297–1340. doi:10.15779/Z389C6S202

[2] Ann Cavoukian. 2011. *Privacy by Design: The 7 Foundational Principles.* Technical Report. Information and Privacy Commissioner of Ontario, Toronto, ON, Canada. Originally published August 2009; revised January 2011.

[3] Coalition for Content Provenance and Authenticity. 2025. *C2PA Technical Specification v2.3.* Technical Report. C2PA. https://spec.c2pa.org/specifications/specifications/2.3/specs/C2PA_Specification.html

[4] I.J. Cox, J. Kilian, F.T. Leighton, and T. Shamoon. 1997. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing* 6, 12 (1997), 1673–1687. doi:10.1109/83.650120

[5] Davide Cozzolino and Luisa Verdoliva. 2020. Noiseprint: A CNN-Based Camera Model Fingerprint. *IEEE Transactions on Information Forensics and Security* 15 (2020), 144–159. doi:10.1109/TIFS.2019.2916364

[6] Taylor Lynn Curtis, Maximilian Puelma Touzel, William Garneau, Manon Gruaz, Mike Pinder, Li Wei Wang, Sukanya Krishna, Luda Cohen, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2025. Veracity: An Open-Source AI Fact-Checking System. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25).* IJCAI Organization, 11021–11024. doi:10.24963/ijcai.2025/1254

[7] European Parliament and Council of the European Union. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). *Official Journal of the European Union* L 277 (Oct. 2022), 1–102. http://data.europa.eu/eli/reg/2022/2065/oj

[8] European Parliament and Council. 2016. *Regulation (EU) 2016/679 (General Data Protection Regulation).* Regulation. Official Journal of the European Union. Article 25: Data protection by design and by default.

[9] European Parliament and Council. 2024. *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act).* Regulation. Official Journal of the European Union. http://data.europa.eu/eli/reg/2024/1689/oj Article 50: transparency obligations for providers and deployers of certain AI systems.

[10] Bijean Ghafouri, Shahrad Mohammadzadeh, James Zhou, Pratheeksha Nair, Jacob-Junqi Tian, Hikaru Tsujimura, Mayank Goel, Sukanya Krishna, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. 2024. Epistemic Integrity in Large Language Models. arXiv:2411.06528

[11] Google. 2024. Flutter: Build apps for any screen. https://flutter.dev. Cross-platform UI toolkit.

[12] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020). doi:10.1177/2053951719897945

[13] Sven Gowal et al. 2025. SynthID-Image: Image watermarking at internet scale. arXiv:2510.09263 [cs.CR] https://arxiv.org/abs/2510.09263

[14] Zhengyuan Jiang, Jinghuai Zhang, and Neil Zhenqiang Gong. 2023. Evading Watermark based Detection of AI-Generated Content. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23).* ACM, New York, NY, USA, 1168–1181. doi:10.1145/3576915.3623189

[15] Ralf Jung, Jacques-Henri Jourdan, Robbert Krebbers, and Derek Dreyer. 2017. RustBelt: Securing the Foundations of the Rust Programming Language. *Proc. ACM Program. Lang.* 2, POPL, Article 66 (2017), 34 pages. doi:10.1145/3158154

[16] Daniel Kang, Tatsunori Hashimoto, Ion Stoica, and Yi Sun. 2022. ZK-IMG: Attested Images via Zero-Knowledge Proofs to Fight Disinformation. arXiv:2211.04775 [cs.CR] https://arxiv.org/abs/2211.04775

[17] Victor Livernoche, Akshatha Arodi, Andreea Musulan, Zachary Yang, Adam Salvail, Gaétan Marceau Caron, Jean-François Godbout, and Reihaneh Rabbany. 2025. OpenFake: An Open Dataset and Platform Toward Real-World Deepfake Detection. *arXiv preprint arXiv:2509.09495* (2025).

[18] Alexander Loth. 2021. *Decisively Digital: From Creating a Culture to Designing Strategy.* John Wiley & Sons, Inc., Hoboken, NJ, USA. https://www.wiley.com/en-us/Decisively+Digital%3A+From+Creating+a+Culture+to+Designing+Strategy-p-9781119737285

[19] Alexander Loth. 2026. The Indistinguishability Threshold: Measuring Cognitive Vulnerabilities to AI-Generated Disinformation. In *Companion Publication of the 2026 18th ACM Web Science Conference.* ACM, New York, NY, USA, 79–82. doi:10.1145/3795513.3807421 PhD Symposium.

[20] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2024. Blessing or Curse? A Survey on the Impact of Generative AI on Fake News. doi:10.48550/arXiv.2404.03021 arXiv:2404.03021 [cs.CL]

[21] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. Eroding the Truth-Default: A Causal Analysis of Human Susceptibility to Foundation Model Hallucinations and Disinformation in the Wild. In *Companion Proceedings of the ACM Web Conference 2026 (WWW '26 Companion)* (Dubai, United Arab Emirates). ACM, New York, NY, USA, 1125–1132. doi:10.1145/3774905.3795832 Also available as arXiv:2601.22871.

[22] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. Industrialized Deception: The Collateral Effects of LLM-Generated Misinformation on Digital Ecosystems. In *Companion Proceedings of the ACM Web Conference 2026 (WWW '26 Companion)* (Dubai, United Arab Emirates). ACM, New York, NY, USA, 295–302. doi:10.1145/3774905.3795471 Also available as arXiv:2601.21963.

[23] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. Interrupting the Chain: Human Perception of AI-Generated Disinformation Through a Kill Chain Lens. In *INFORMATIK 2026: Digitale Resilienz (Lecture Notes in Informatics (LNI), Vol. P-384)*, Julian Dehne, Daniel Stollberg, and Florian Tschorsch (Eds.). Gesellschaft für Informatik, Bonn, Germany, 321–330. doi:10.18420/inf2026_22 arXiv:2608.21389 [cs.CY]

[24] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. The Verification Crisis: Expert Perceptions of GenAI Disinformation and the Case for Reproducible Provenance. In *Companion Proceedings of the ACM Web Conference 2026 (WWW '26 Companion)* (Dubai, United Arab Emirates). ACM, New York, NY, USA, 980–988. doi:10.1145/3774905.3795484 Also available as arXiv:2602.02100.

[25] Alexander Loth, Dominique Conceicao Rosario, Peter Ebinger, Martin Kappes, and Marc-Oliver Pahl. 2026. Origin Lens: Reclaiming Trust on the AI-Mediated Web Through On-Device Image Provenance Verification. In *Companion Publication of the 2026 18th ACM Web Science Conference.* ACM, New York, NY, USA, 1–5. doi:10.1145/3795513.3806658

[26] J. Lukáš, J. Fridrich, and M. Goljan. 2006. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security* 1, 2 (2006), 205–214. doi:10.1109/TIFS.2006.873602

[27] Hannes Mareen, Dante Vanden Bussche, Fabrizio Guillaro, Davide Cozzolino, Glenn Van Wallendael, Peter Lambert, and Luisa Verdoliva. 2023. Comprint: Image Forgery Detection and Localization Using Compression Fingerprints. In *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*, Jean-Jacques Rousseau and Bill Kapralos (Eds.). Springer Nature Switzerland, Cham, 281–299. doi:10.1007/978-3-031-37742-6_23

[28] Nicholas D Matsakis and Felix S Klock. 2014. The Rust Language. *ACM SIGAda Ada Letters* 34, 3 (2014), 103–104. doi:10.1145/2692956.2663188

[29] Ian McCormack, Joshua Sunshine, and Jonathan Aldrich. 2025. A Study of Undefined Behavior Across Foreign Function Boundaries in Rust Libraries. In *Proceedings of the 2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE).* IEEE, Ottawa, ON, Canada, 2075–2086. doi:10.1109/ICSE55347.2025.00167

[30] Maximilian Puelma Touzel, Sneheel Sarangi, Gayatri Krishnakumar, Busra Tugce Gurbuz, Austin Welch, Zachary Yang, Andreea Musulan, Hao Yu, Ethan Kosak-Hine, Tom Gibbs, Camille Thibault, Reihaneh Rabbany, Jean-François Godbout, Dan Zhao, and Kellin Pelrine. 2025. SandboxSocial: A Sandbox for Social Media Using Multimodal AI Agents. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25), Demo Track.* IJCAI Organization, 11100–11103. doi:10.24963/ijcai.2025/1271

[31] Sijia Qian, Cuihua Shen, and Jingwen Zhang. 2022. Fighting cheapfakes: using a digital media literacy intervention to motivate reverse search of out-of-context visual misinformation. *Journal of Computer-Mediated Communication* 28, 1 (2022), zmac024. doi:10.1093/jcmc/zmac024

[32] Bhaktipriya Radharapu and Harish Krishna. 2024. RealSeal: Revolutionizing Media Authentication with Real-Time Realism Scoring. In *Proceedings of the 26th International Conference on Multimodal Interaction (ICMI '24).* ACM, New York, NY, USA, 585–590. doi:10.1145/3678957.3678960

[33] Catherine Régis, Florian Martin-Bariteau, Jake Okechukwu Effoduh, Juan David Gutiérrez, Gina Neff, Carlos Affonso Souza, and Célia Zolynski. 2025. *AI in the Ballot Box: Four Actions to Safeguard Election Integrity and Uphold Democracy.* Technical Report. IVADO Global Policy Briefs. https://doi.org/10.32920/28382087

[34] Adam Shostack. 2014. *Threat Modeling: Designing for Security.* John Wiley & Sons.

[35] Saud Sohail, Syed Muhammad Sajjad, Adeel Zafar, Zafar Iqbal, Zia Muhammad, and Muhammad Kazim. 2025. Deepfake Image Forensics for Privacy Protection and Authenticity Using Deep Learning. *Information* 16, 4 (2025). doi:10.3390/info16040270

[36] Borce Stojkovski, Gabriele Lenzini, and Vincent Koenig. 2021. "I Personally Relate It to the Traffic Light": A User Study on Security & Privacy Indicators in a Secure Email System Committed to Privacy by Default. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing (SAC '21).* ACM, New York, NY, USA, 1235–1246. doi:10.1145/3412841.3441998

[37] Camille Thibault, Jacob-Junqi Tian, Gabrielle Péloquin-Skulski, Taylor Lynn Curtis, James Zhou, Florence Laflamme, Luke Yuxiang Guan, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. 2025. A Guide to Misinformation Detection Data and Evaluation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25).* ACM, New York, NY,

USA, 5801–5809. doi:10.1145/3711896.3737437

[38] Luisa Verdoliva. 2020. Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing* 14, 5 (2020), 910–932. doi:10.1109/JSTSP.2020.3002101

[39] Sze Yuh Nina Wang, Samantha C. Phillips, Kathleen M. Carley, Hause Lin, and Gordon Pennycook. 2025. Limited effectiveness of psychological inoculation against misinformation in a social media feed. *PNAS Nexus* 4, 6 (2025), pgaf172. doi:10.1093/pnasnexus/pgaf172

[40] Christopher D. Wickens and Anthony D. Andre. 1990. Proximity Compatibility and Information Display: Effects of Color, Space, and Objectness on Information Integration. *Human Factors* 32, 1 (1990), 61–77. doi:10.1177/001872089003200105

[41] Mingyang Xie, Manav Kulshrestha, Shaojie Wang, Jinghan Yang, Ayan Chakrabarti, Ning Zhang, and Yevgeniy Vorobeychik. 2022. PROVES: Establishing Image Provenance using Semantic Signatures. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 3017–3026. doi:10.1109/WACV51458.2022.00307

Supplementary Materials

This supplement provides additional figures, architectural details, user interface screenshots, and regulatory context that support the main text.

A Defense-in-Depth Verification Pipeline

Origin Lens implements a four-layer defense-in-depth strategy for image verification, where each layer provides independent verification signals with decreasing confidence levels:

Layer 1: C2PA Provenance. The primary verification layer parses JUMBF-embedded C2PA manifests, validates X.509 certificate chains, and verifies SHA-256 hard bindings between the manifest and image content.

Layer 2: EXIF/IPTC Metadata. When C2PA manifests are absent, the system analyzes EXIF and IPTC metadata for generative AI signatures, including Stable Diffusion parameters, DALL-E identifiers, and Midjourney tags.

Layer 3: Watermark Detection. Detection of imperceptible watermarks (e.g., Google SynthID) via an external API, run automatically when no C2PA manifest is found, provides additional signals for AI-generated content identification.

Layer 4: Contextual Verification. As a final fallback, users may opt in to reverse image search for prior attributions and contextual information.

Figure S1 illustrates this layered approach.

B C2PA Manifest Structure

The C2PA standard defines a hierarchical manifest structure embedded within image files, consisting of four core components:

Assertions contain metadata about the content's creation, including timestamps, software used, and editing actions performed.

Claims aggregate assertions and establish relationships with ingredient manifests (for composite images).

Signatures use ECDSA or RSA algorithms with X.509 certificates to cryptographically bind claims to the signing entity.

Hard Bindings compute SHA-256 hashes over the asset's bytes (excluding the embedded manifest), ensuring any modification invalidates the manifest.

Figure S2 illustrates these relationships.



Figure S1: Defense-in-depth verification layers. Primary layers (top) provide highest confidence; lower layers offer fallback signals when cryptographic provenance is unavailable.

C System Architecture Details

Origin Lens employs a hybrid mobile architecture optimized for performance and memory safety, separating the Dart/Flutter presentation layer from the Rust cryptographic core. The architecture provides several benefits:

- **Memory Safety:** Rust's ownership model prevents buffer overflows and use-after-free vulnerabilities when parsing complex binary formats.
- **Performance:** Native code execution for computationally intensive cryptographic operations.
- **Cross-Platform:** Flutter enables deployment to iOS, Android, and desktop platforms from a single codebase.

Figure S3 presents the detailed layered architecture.

D Analysis Workflow

The image analysis workflow follows a decision tree that first checks for C2PA manifests and falls back to heuristic analysis when cryptographic provenance is unavailable. Images with C2PA manifests undergo cryptographic verification (signature and hash validation), while images without manifests are analyzed for AI generation signatures in EXIF metadata. The workflow produces four possible outcomes: Verified, Invalid, AI Generated, or Warning. Figure S4 illustrates this complete workflow.

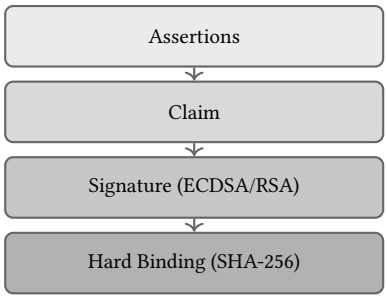


Figure S2: C2PA manifest structure showing the relationship between assertions, claims, cryptographic signatures, and content hash bindings.

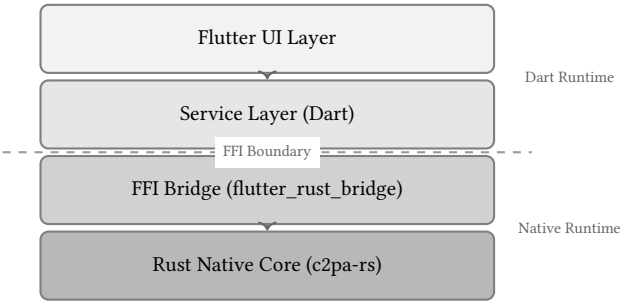


Figure S3: Origin Lens hybrid architecture. The FFI boundary separates managed Dart code from native Rust, enabling memory-safe cryptographic operations.

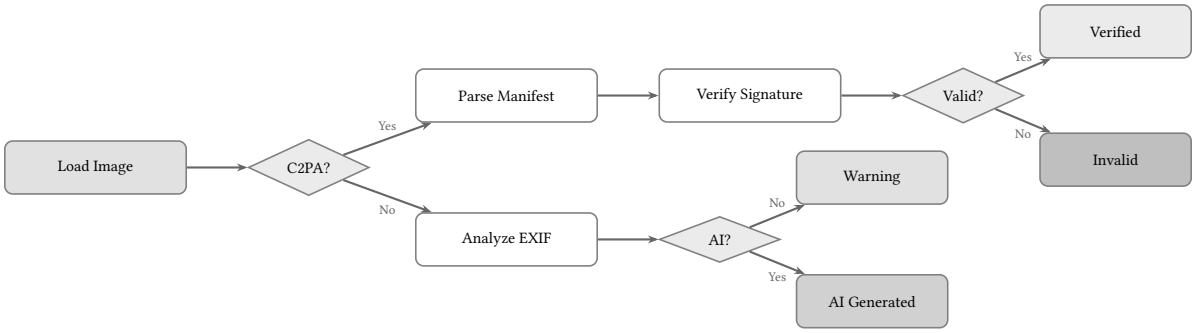


Figure S4: Image analysis workflow showing decision points and four possible verification outcomes.

E Verification Status Indicators

Origin Lens communicates verification results using a traffic-light-inspired visual system. Table S1 describes each status indicator and its meaning.

Table S1: Verification status indicators and their meanings.

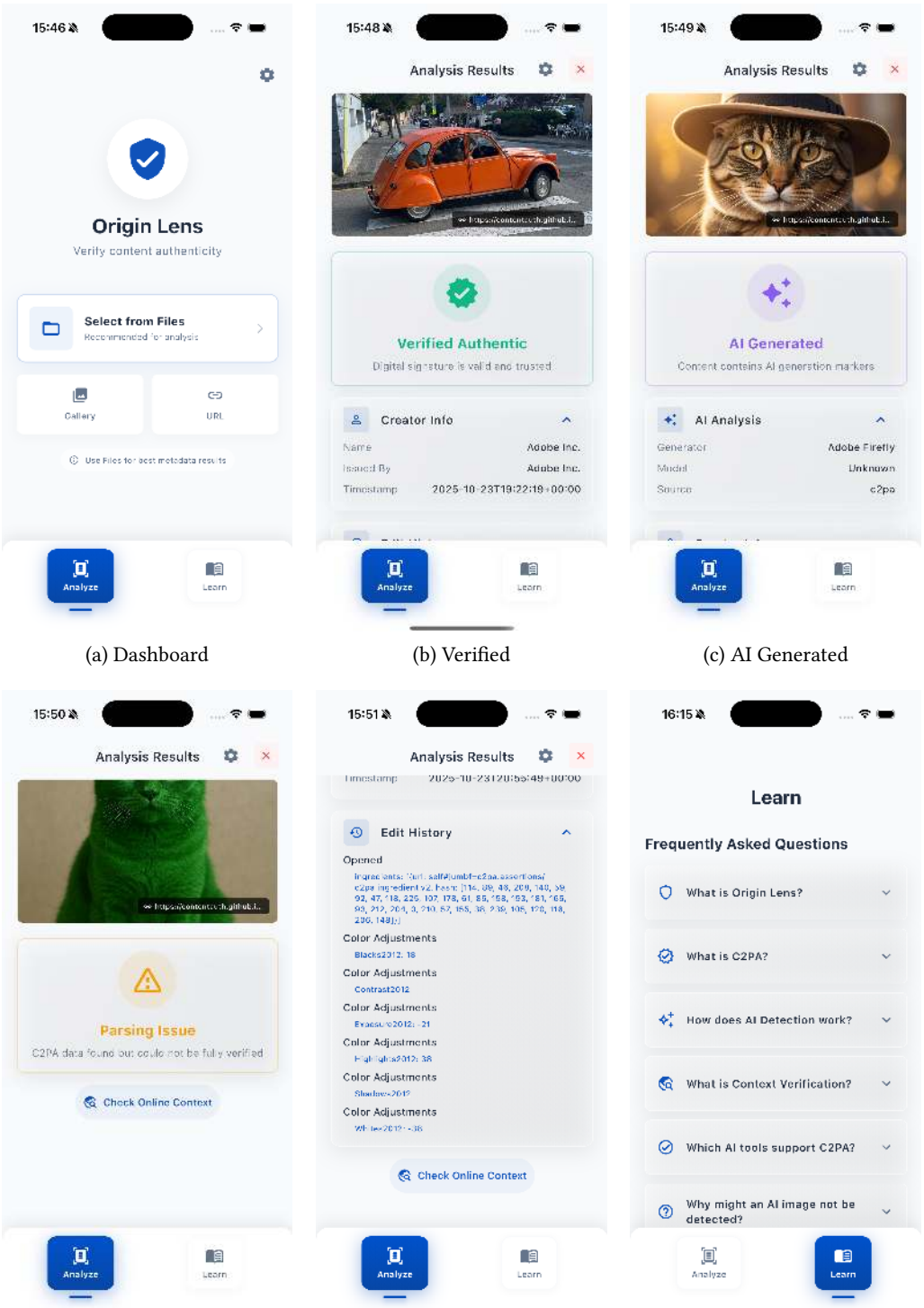
Status	Color	Description
Verified	Green	Valid C2PA manifest with trusted certificate chain
AI Generated	Purple	Content identified as AI-generated via C2PA or EXIF markers
Warning	Orange	No manifest found, expired certificate, or parsing issue
Invalid	Red	Hash mismatch, broken certificate chain, or detected manipulation

F Regulatory Alignment

Origin Lens aligns with emerging EU regulations on AI transparency and cybersecurity. Table S2 summarizes how the framework addresses relevant regulatory requirements.

G User Interface Screenshots

This section presents the Origin Lens user interface across different verification scenarios. Figure S5 shows: (a) the main dashboard with upload options, (b) a verified result with valid C2PA manifest, (c) AI-generated content detection, (d) a parsing issue warning, (e) the detailed manifest history view, and (f) the educational FAQ section.



(a) Dashboard (b) Verified (c) AI Generated

(d) Parsing Issue (e) Edit History (f) FAQ / Learn

Figure S5: Origin Lens user interface screenshots: (a) Dashboard with upload options (Files, Gallery, URL), (b) Verified result with valid C2PA manifest and trusted signature chain, (c) AI-generated content detected via Adobe Firefly metadata, (d) Parsing issue warning when signature chain is incomplete, (e) Detail view showing C2PA manifest history with timestamps and ingredient hashes, (f) Learn section with expandable FAQ on C2PA and AI detection.

Table S2: Regulatory alignment of Origin Lens features.

Regulation	Relevant Requirements	Origin Lens Alignment
EU AI Act (2024/1689)	Article 50(2): Machine-readable marking of AI-generated content	C2PA manifest parsing and AI generation detection via metadata
GDPR (2016/679)	Article 25: Privacy by Design; Article 5(1)(c): Data minimization	On-device processing; no server transmission for core verification
Cyber Resilience Act (2024/2847)	Security-by-design for digital products	Rust memory safety; X.509 certificate validation
NIS2 Directive (2022/2555)	Supply chain security requirements	Local trust store; certificate chain verification