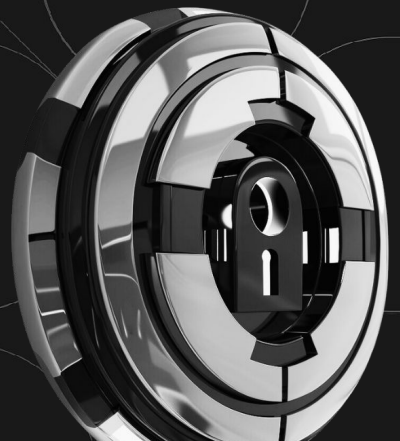
Secublox

DeepShield: Where trust begins

At the end of October 2024 we presented the DeepShield idea to Sprind. A concept that goes beyond where others have yet to begin, making deepfake prevention both simple and effective.

A multi-layered approach - integrating TEEs, cryptographic techniques, Blockchain and DLT - providing end-to-end authenticity verification and manipulation detection. Our idea to advanced technology and thus ensuring the integrity of media, protecting it from tampering and securing trust, preventing deepfakes and safeguarding content authenticity seamlessly brought new insights, some turns and unexpectancies but is now setting new standards for digital content authenticity and integrity.



Secublox

I. Stage 1

DeepShield goes beyond where others have yet to begin, making deepfake prevention both simple and effective. By leveraging a multi-layered approach—integrating TEEs, cryptographic techniques, Blockchain and DLT—we provide end-to-end authenticity verification and manipulation detection. Our advanced technology ensures the integrity of your media, protecting it from tampering and securing trust across platforms.

With DeepShield, preventing deepfakes and safeguarding content authenticity has never been more seamless.

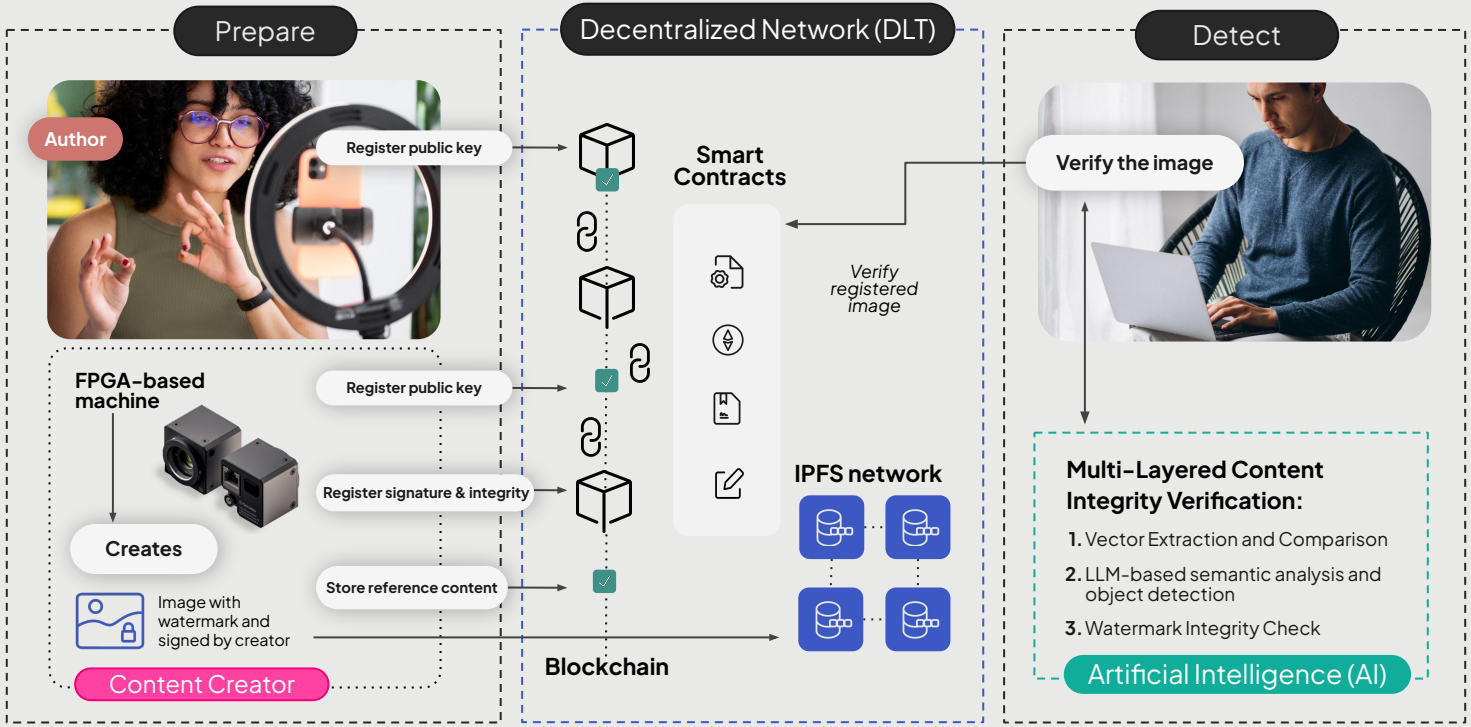
The idea: A Distributed and Cryptographic Approach for Authenticating Digital Content, based on Trusted Environments, Establishing Ownership, and Detecting Unwanted Manipulations.

DeepShield goes beyond where others have yet to begin, making deepfake prevention and detection simple and effective.

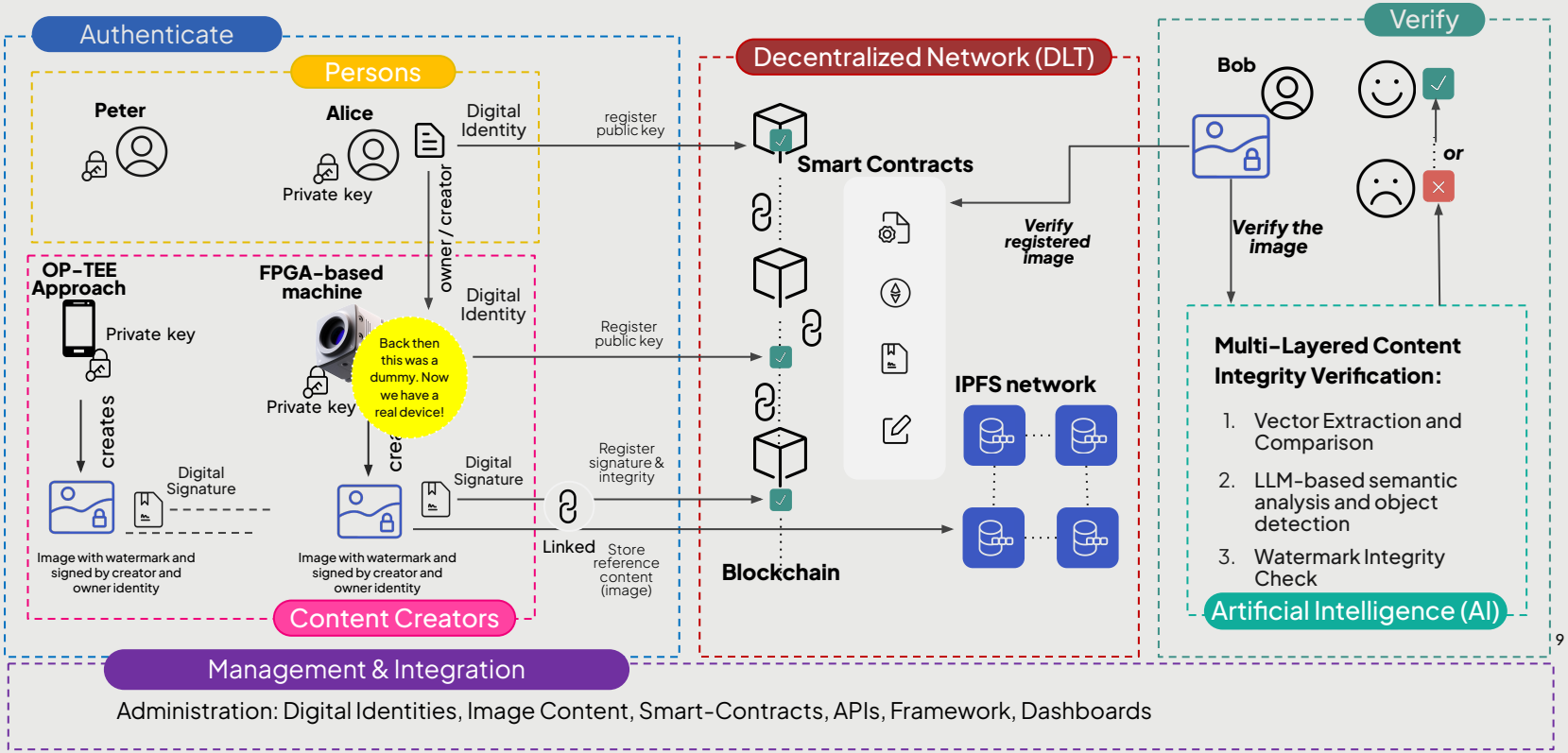
Our approach focuses on preventing deepfakes through a comprehensive, multi-layered content integrity verification. At the point of content creation, trusted accounts or devices with trusted execution environments, embed unique cryptographic watermarks into digital media. These watermarks are linked to device and/ or account public keys stored on a blockchain, enabling transparent verification. Content integrity is ensured through vector comparisons, AI-powered semantic analysis, and object detection to identify unauthorized alterations.

The system also performs watermark integrity checks to detect tampering attempts.
By leveraging a multi-layered approach—integrating Artificial Intelligence (AI), Blockchain/DLT, Cloud Technology, Cryptographic Techniques, Field Programmable Gate Array (FPGA), Platform Technologies, Smart-Contracts and Trusted Execution Environment (TEE) — we design, implement, provide and demonstrate a radically new end-to-end Proof-of-Concept for authenticity verification and manipulation detection.

Content authenticity and deepfake prevention powered by a blockchain framework with a multi-layered approach to ensure media authenticity and detect manipulation.



DeepShield enabled by a Blockchain framework simplifies deepfake prevention with our multi-layered approach, combining FPGAs & TEEs, cryptographic techniques, and DLT to secure media authenticity and manipulation detection

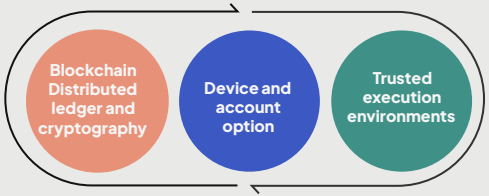


The iphone-approach for anti-deep fake: Best of breed of technologies for the best solution

Just like the iPhone transformed the tech world by combining existing technologies like touchscreens, GPS, and cameras into a single, innovative device, DeepShield merges cryptographic watermarking, AI analysis, and blockchain verification into a unified solution.

This novel integration offers flexibility, allowing for hardware enhancements such as Trusted Execution Environments (TEEs), and sets a new benchmark for digital media security. By blending the best-in-class components, DeepShield disrupts the market, creating a scalable, secure platform that reshapes content authenticity and integrity standards.

UNIQUE COMBINATION ✓



What we promised and what we achieved in stage 1

- Trusted Execution Environments (TEEs) ✓
- Decentralized Verification ✓
- blind spread spectrum watermarking ✓
- Immutable Record ✓
- Distributed Ledger Technology (DLT) ✓
- cryptographic techniques ✓
- FPGA-based machine vision sensors for ultra-fast processing and customizable encryption ✓
- Integration & Interoperability ✓
- Multi-layered Semantic Integrity ✓

Combining AI with Blockchain brings many advantages for anti-deep fake: Heidari, A., Navimipour, N. J., Dag, H., Talebi, S., & Unal, M. (2024). A novel blockchain-based deepfake detection method using federated and deep learning models. Cognitive Computation, 16(3), 1073–1091. https://doi.org/10.1007/s12559-024-10255-7 and Yun, J., et al. (2024). DRPChain: A new blockchain-based trusted DRM scheme for image content protection. PLOS ONE, 19(9), e0309743. And Madushanka, T., Kumara, D.S., & Rathnaweera, A.A. (2024). SecureRights: A Blockchain-Powered Trusted DRM Framework for Robust Protection and Asserting Digital Rights.

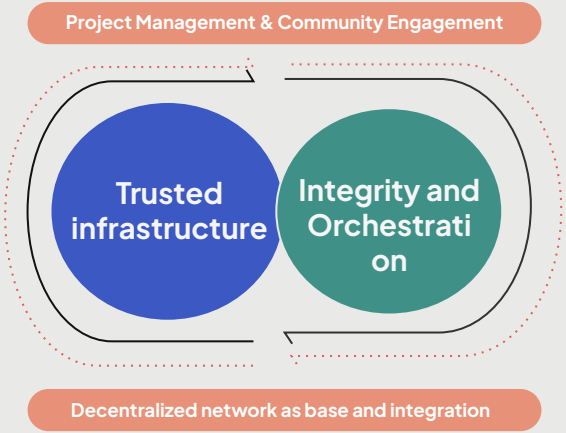
Dual-Stream Execution for High-Impact Innovation in Deepfake prevention and detection

How can we achieve the most innovation in the shortest amount of time?
That was the key question we asked ourselves. The solution: To maximize innovation and accelerate delivery, we structured DeepShield as a modular system, distributed across four dedicated development tracks (A–D). Each module focused on a distinct capability—from secure content creation to intelligent verification—while a unifying integration stream ensured seamless convergence. We didn't just talk about decentralization—we structured our work around it.

Trusted Infrastructure Layer (*Projects A & B*)

Development of the foundational technology for secure and verifiable media authentication.

- **Secure Content Creation**: Trusted execution via OP-TEE and FPGA-based systems ensuring hardware-rooted authenticity and digital signatures.
- **Decentralized Trust Architecture**: Use of blockchain, smart contracts, and IPFS to anchor identities and register immutable content records.



Cognitive Integrity & Orchestration Layer (Projects C & D)

Advanced content verification paired with seamless management and orchestration.

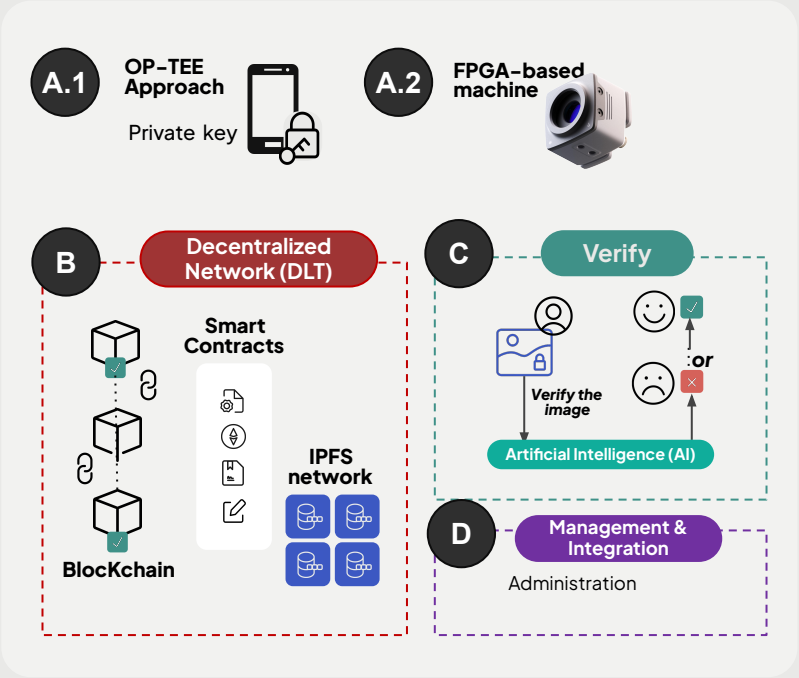
- **AI-Powered Verification**: Multi-layered integrity checks using vector analysis, semantic AI, and watermark validation.
- **System Management & Integration**: Coordination of identities, contracts, APIs, and dashboards to orchestrate a transparent and adaptable ecosystem.

Secublox
IABG

Dual-Stream Execution for High-Impact Innovation in Deepfake prevention and detection

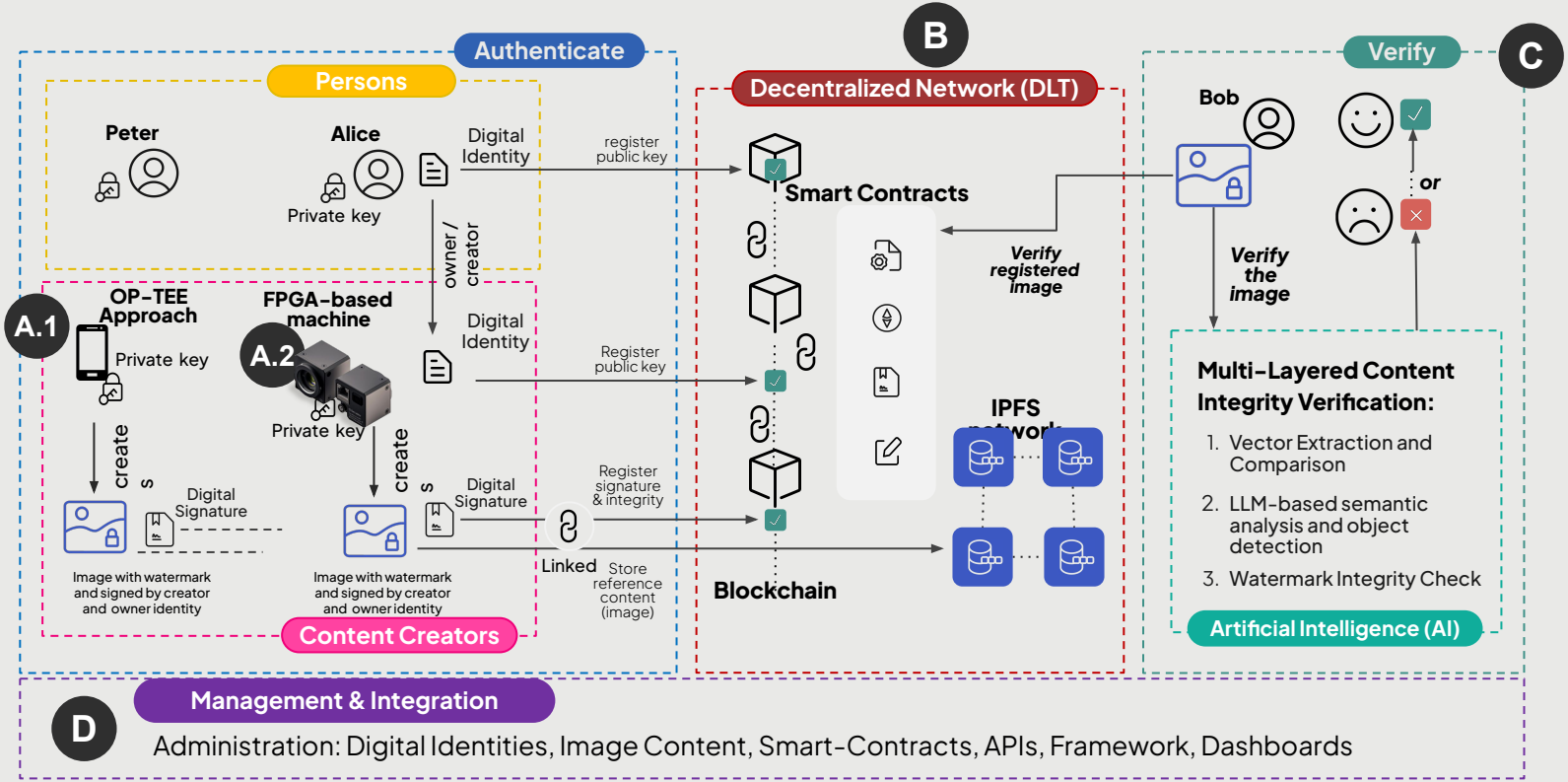
To investigate the described challenges the DeepShield demonstrator for Stage 1 was implemented based on an overall approach, divided into following independent milestones:

- **Milestone A (Authenticate Content)**: Authentication by integrating cryptographic, digital identity, watermarking together with FPGA and TEE logic (A.1: OP-TEE / A.2: FPGA-based machine). In additional digital signed data, transferred to the decentralized network (integrated with Module B and D).
- **Milestone B (DLT Network):** Prototype of a decentralized network to register digital identities (devices, accounts), image signatures/integrity and image reference content. In additional to verify existing images (integrated with module A, C and D).
- **Milestone C (Verify Content):** Prototype of the Artificial Intelligence (AI) services to detect deep fakes, semantic integrity and traceability of images registered in decentralized network (integrated with B and D).
- **Milestone D (Management & Integration):** Prototype of the Integration-Framework (Software-Library, APIs, Integration) together DeepShield Platform backend (Dashboard, database, integrated with A, B and D).



Secublox

DeepShield enabled by a Blockchain framework simplifies deepfake prevention with our multi-layered approach, combining FPGAs & TEEs, cryptographic techniques, and DLT to secure media authenticity and manipulation detection



Documentation of our work

Milestone	Abstract & Concept	User Guide, Installation & Software	Evaluation / Demonstration
A.1–OP-TEE Approach	DeepShield_Technical-Concept.pdf	A1_Watermarking_setup_and_usage_guid.pdf README.md file of the corresponding module provides the same content.	DeepShield_Technical-Concept.pdf A1_OP_TEE_Image_Registration_small.mov
A.2–FPGA based Visual Capture Device	A2_FFPGA_VCD_Abstract.docx	A2_FFPGA_VCD_Guideline_Installation.pdf	A2_FFPGA_VCD_Evaluation.docx A2_FFPGA_VCD_Demonstration_small.mov A2_FPGA_VCD_Evaluation_1.mp4 A2_FPGA_VCD_Evaluation_2.mp4 A2_FPGA_VCD_Evaluation_3.mp4
B-Decentralized Network (DLT)	B_DLT_Abstract.docx	B_DLT_Guideline_DeepShield.docx B_DLT_DeepShield_Library.zip deepshield-blockchain-main deepshieldlibrary-main	A2_FFPGA_VCD_Demonstration_small.mov
C-Multi-Layered Content Integrity Verification	DeepShield_Technical-Concept.pdf	C_Multilayered_semantic_integrity_user_guide.pdf README.md file of the corresponding module provides the same content.	DeepShield_Technical-Concept.pdf A1_Semantic_Integrity.mp4 A1_Watermark_extraction_perturbed_cropping.mp4 A1_Watermark_extraction_perturbed_image.mp4
D-DeepShield-Management/Integration	AD_Platform_MI_Abstract.docx	D_Platform_MI_Guideline_Installation.pdf	D_Platform_MI_Demonstration_small.mov

Secublox

A.1 Open Portable Trusted Execution Environment

Achievements after Stage 1

To guarantee secure execution and data storage, the overall application logic should be separated into a general and secure parts.

✓ To this end, the application's architecture is structured into two domains: a Normal World for general execution and a Secure World for storing protected data and running sensitive code.

The image creation and watermarking approach must be guaranteed to be protected from impersonation attacks, which would undermine the credibility of the watermark verification process.

✓ To achieve this, the image creation and watermarking process is designed to be executed within a Trusted Execution Environment.

Scalability and adaptability to different digital platforms are essential.

✓ To ensure scalability and adaptability across digital platforms, we selected the open-source operating system OP-TEE, which runs in the Secure World and provides a portable, standardized runtime environment with API support for developing Trusted Applications facilitating future possible migration to real devices.

Challenges & Technical Limitations

! Capturing a picture within the Secure World on an OP-TEE-enabled Raspberry Pi proved unfeasible due to the lack of camera driver support in OP-TEE.

! Watermark embedding techniques need to be implemented in Rust as OP-TEE does not have extensive support for Python code.

→ Our solution uses a separate Raspberry Pi device without OP-TEE to capture an image which is then transferred to the Raspberry Pi device with OP-TEE for watermarking.

We used a technology mix in which a Python server calls Rust programs which internally uses functions written in C.

Secublox

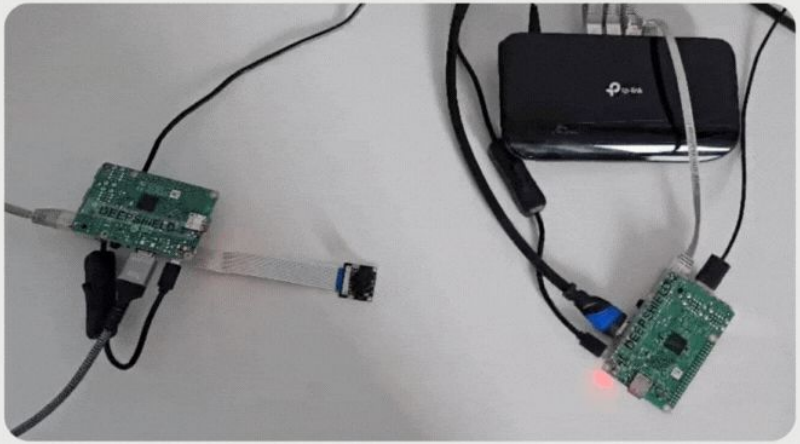
A.1 Open Portable Trusted Execution Environment

Trusted Execution Environments (TEEs)

Achievements after Stage 1 – Demonstration

Secublox

A.1 Open Portable Trusted Execution Environment



Trusted Execution Environments (TEEs)

Achievements after Stage 1 – Demonstration

Secublox

A.1 Open Portable Trusted Execution Environment

Secublox

DLT & Watermarking: Complementary Approaches

- **Phase 1 Goal:**
  Watermarking used to prove content authenticity and provenance.
- **Blockchain Integration:**
  Content metadata (e.g., creator, timestamp) registered on-chain.
- **Limitations Identified:**
  - Watermarking lacked robust creator attribution.
  - Vulnerable to certain manipulation types.

A.1 Watermarking process

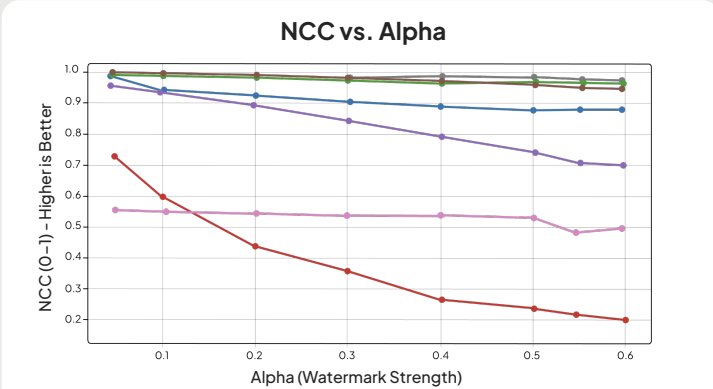
The report evaluates the robustness of a watermarking approach and the effectiveness of a semantic integrity detection module.

Watermarking Robustness:

- The **watermarking method performs well against common distortions** (e.g., brightness, compression, noise) with high NCC, BCR, and CATSO scores across all α values.
- **Performance drops significantly under geometric transformations** (e.g., rotation, cropping), where vector alignment remains but bit recovery fails.
- **Future improvements should focus on adding error correction**, spatial alignment, or invariant embedding techniques to enhance resilience to geometric changes.

Semantic Integrity Detection:

- Evaluated on 4978 CASIA v2.0 samples (splicing and copy-move).
- **"Image embedding" excelled in splicing**; copy-move results were mixed..
- **Key evaluation metrics: F1 score,** precision, and recall.
- **Most reliable methods:** image captioning (0.35 weight), object detection (0.3), image embedding (0.2), and ORB matching (0.15).
- **Method weighting was manually defined;** optimization for better accuracy is still pending.



Tool (or tool set)	Tampering	Accuracy	Precision	Recall	TNR	F1	Total	Pos.	Neg.
Combined	splicing	0.9726	0.9486	0.9994	0.9458	0.9733	3504	1752	1752
Image Captioning	splicing	0.8799	0.8091	0.9943	0.7654	0.8922	3504	1752	1752
Object Detection	splicing	0.9375	0.9233	0.9543	0.9207	0.9385	3504	1752	1752
Image Embedding	splicing	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	3504	1752	1752
ORB Feature Matching	splicing	0.9484	0.9093	0.9960	0.9007	0.9507	3504	1752	1752
Combined	copy-move	0.6634	0.9060	0.3645	0.9622	0.5199	6452	3226	3226
Image Captioning	copy-move	**0.6927**	0.7121	**0.6469**	0.7384	**0.6779**	6452	3226	3226
Object Detection	copy-move	0.6427	0.7621	0.4151	0.8704	0.5375	6452	3226	3226
Image Embedding	copy-move	0.6001	**1.0000**	0.2002	**1.0000**	0.3336	6452	3226	3226
ORB Feature Matching	copy-move	0.5250	0.8675	0.0589	0.9910	0.1103	6452	3226	3226
Combined	**All**	**0.7722**	**0.9310**	**0.5880**	**0.9564**	**0.7208**	**9956**	**4978**	**4978**

A.2 FPGA –based Visual Capture Device

FPGA-based machine vision sensors for ultra-fast processing and customizable encryption

Achievements after Stage 1

Streamlining deepfake detection and verification to strengthen media integrity and operational trust in complex multi-domain, multinational Defence environments.

Scalability and adaptability to different digital platforms are essential.

A FPGA-based visual capture device with real-time encryption and metadata handling for secure image processing in real-time.

Design and implementation of a Routing Node as integration layer, for modular, API-driven integration, that allows to work seamlessly across different digital platforms.

Challenges & Technical Limitations

Fast and iterative implementation approach requires continously updated FPGA code, did not match with static TEE at current development stage.

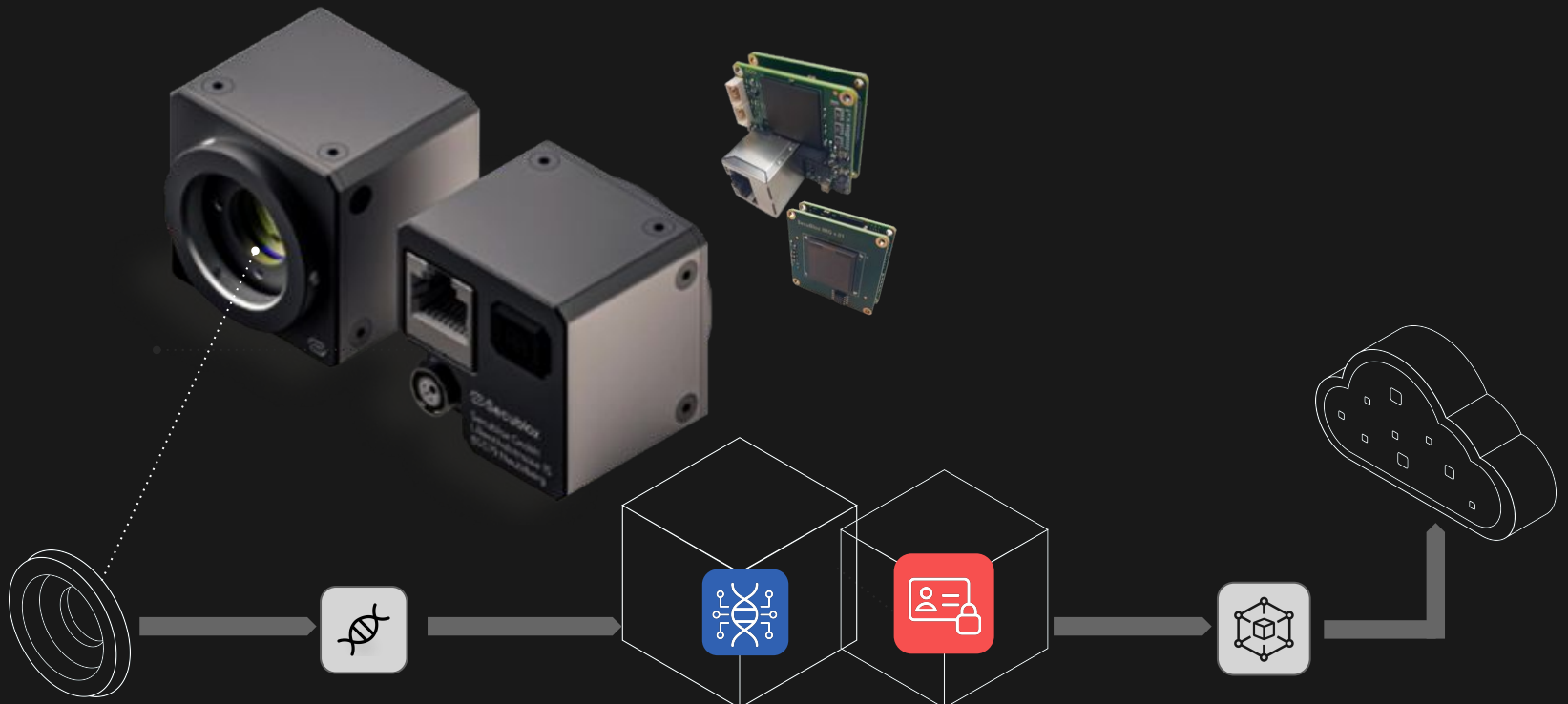
Watermarking consumes a lot of resources in the FPGA and the selected HW design at the beginning could not be changed quickly enough.

TEE architecture by hardware design observed and feasible for Stage 2.

Watermarking feasible with FPGA that offers more resources.

2. The implementation:

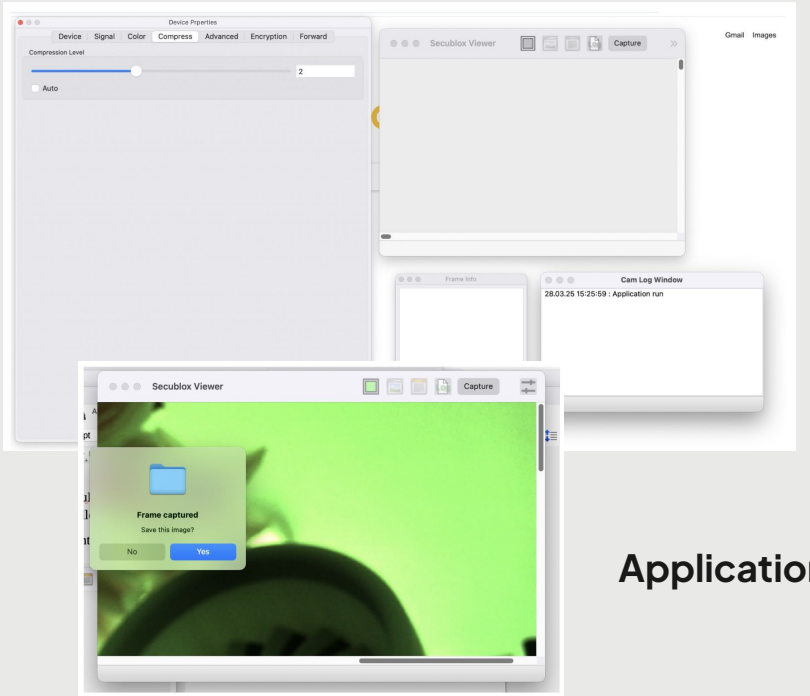
Secublox



Secublox

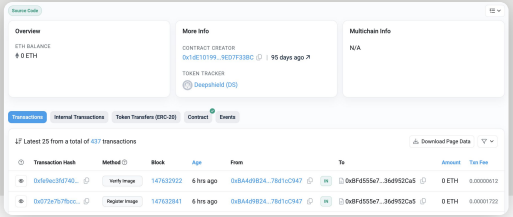
A.2 FPGA –based Visual Capture Device & Application

Device



Application

B Decentralized Network (DLT)



Achievements after Stage 1

Scalability and adaptability to different digital platforms are essential.

Decentralized authentication and verification system to prevent deep fakes and ensure digital content integrity.

Able to continuously adapt to new deepfake techniques. .

- Design, implementation and operation of a decentralized network to register digital identities, image signatures/integrity and image reference content.
- External techniques like Watermark Integrity Check, LLM-based semantic analysis, Object-, DeeFake- and AI-detection integrated.
- End 2 end approach for Image Content: Creation & Secure Watermarking -> Registration -> Verification & Context Integrity enabled by Blockchain

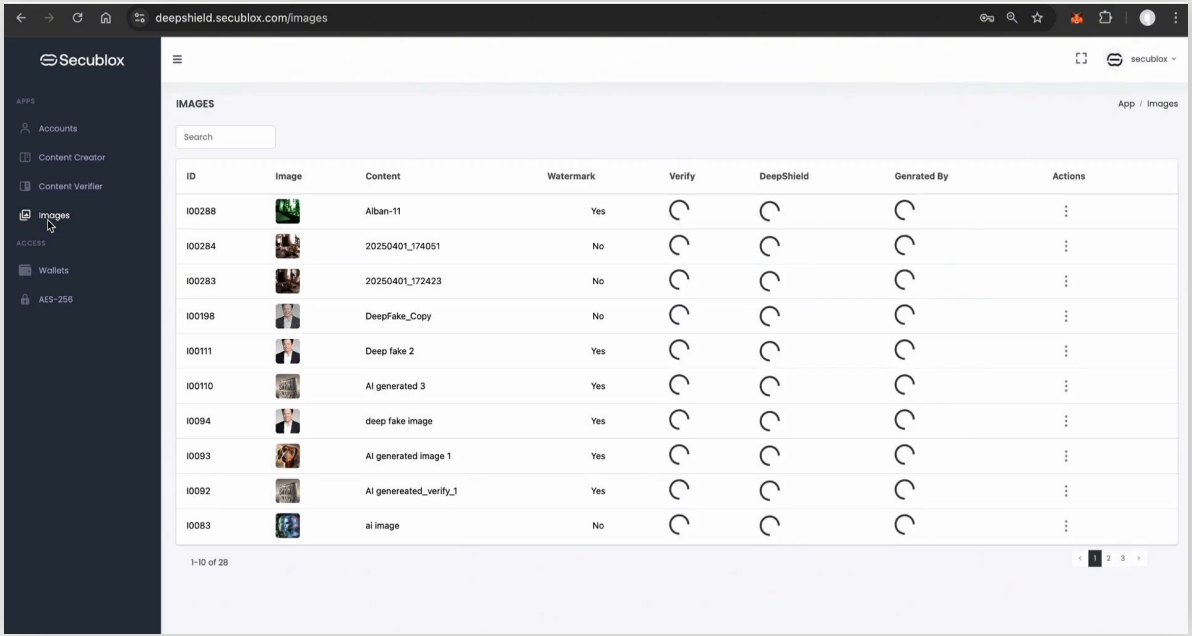
Challenges & Technical Limitations

! Higher scalability (>TPS) and reduced latency required, to finalize the DeepShield product for market readiness.

→ Flexible blockchain architecture designed and implemented to be flexible an avieved an optimized overall approach, for the needed scalability in Stage 2.

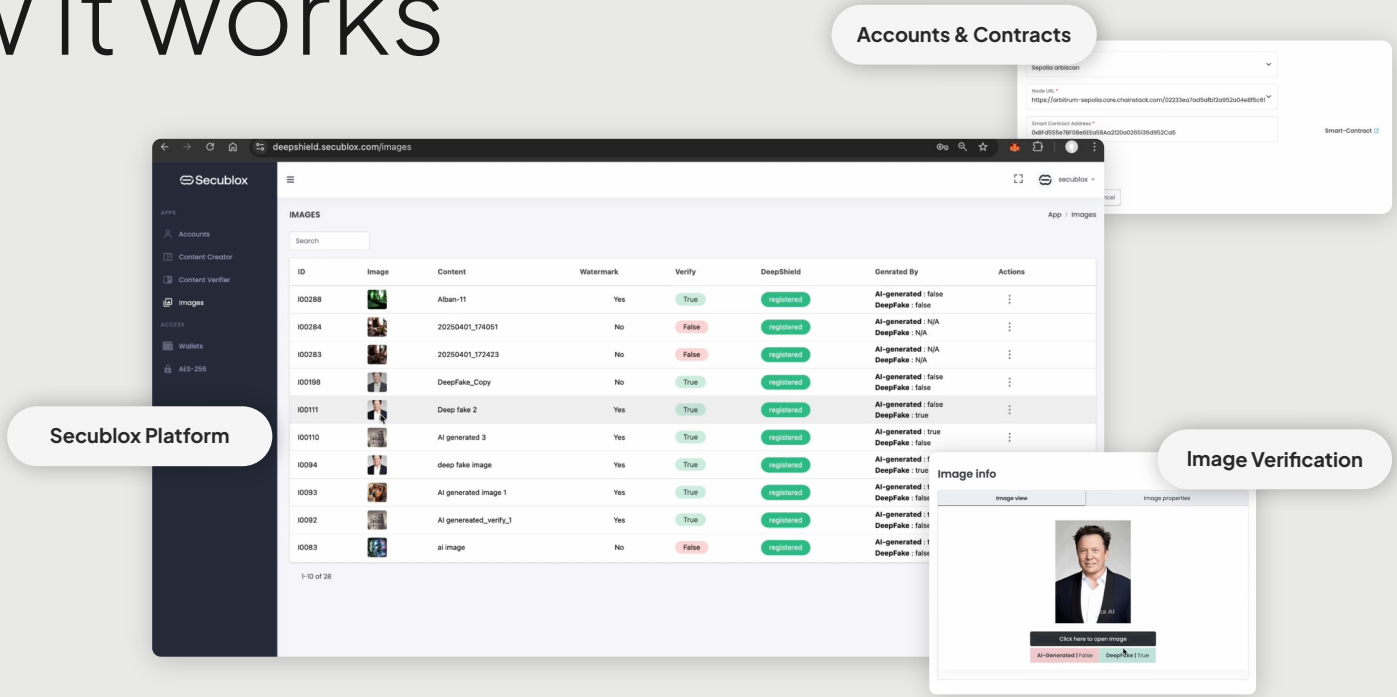
Secublox platform

- Accounts & Contracts
- Content Creator
- Content Verification
- Images
- Wallets



Software

How it works



C Multi-Layered Content Integrity Verification

Achievements after Stage 1

Streamlining deepfake detection and verification to strengthen media integrity and operational trust

Scalability and adaptability to different digital platforms are essential.

Should be able to continuously adapt to new deepfake techniques.

- ✓ Design, implementation and operation of a decentralized approach to register and verify image content in an authenticated way.
- ✓ Techniques like Watermark and Sementic Integrity Check, LLM-based semantic analysis and Object detection.
- ✓ External DeeFake- and AI-detection-services integrated.

Challenges & Technical Limitations

- ! Semantic integrity lacks established benchmarks and datasets.
- ! Integration of diverse tools used for semantic integrity verification.

→ Executed an in-depth literature review, which led us to discover the CASIA 2.0 dataset.

→ Introduced a negative sampling strategy, generating image pairs that do not contain semantic manipulations, thus allowing us to measure false positives and better calibrate precision metrics.

→ Design and implementation of abstract interface set to standardize the interaction

Secublox

C Multi-Layered Content Integrity Verification

Multi-layered Semantic Integrity ✓

To Reach Milestone C, we developed a holistic approach to streamline deepfake detection and verification, ensuring media integrity and operational trust. By integrating various techniques, including watermarking, semantic integrity checks, Large Language Model (LLM)-based semantic analysis, and object detection, we created a robust framework for continuous adaptation to emerging deepfake tactics.

Achievements after Stage 1 – Demonstration

Achievements:

- Decentralized Registration and Verification: Designed and implemented a decentralized approach for authentic registration and verification of image content.
- Multi-Faceted Techniques: Successfully integrated:
  - Watermarking for tamper-evident protection.
  - Semantic Integrity Checks via:
    - Image Captioning and Text Embedding Similarity.
    - Object Detection and Bounding Box Analysis.
    - Image Embedding Comparisons.
    - ORB KeyPoint Matching for geometric integrity.

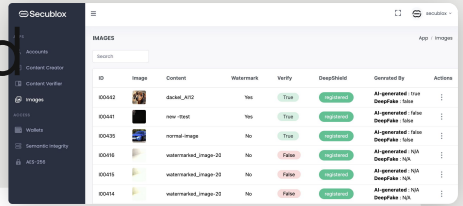
Secublox

C Multi-Layered Content Integrity Verification

Multi-layered Semantic Integrity

Achievements after Stage 1 – Demonstration

D DeepShield-Management-Dashboard



Achievements after Stage 1

At least three different use cases will be demonstrated.

Scalability and adaptability to different digital platforms are essential.

Should be able to continuously adapt to new deepfake techniques.

UC-1: Registration of image content by platform, OP-TEE application or FPGA Image Capture Device
UC-2: Multi-Layered Content Integrity Verification
UC-3: Continuously adapt to new deepfake techniques

External techniques like Watermark Integrity Check, LLM-based semantic analysis, Object-, DeeFake- and AI-detection integrated.

Challenges & Technical Limitations

Fast and iterative implementation approach requires continously updated FPGA code, did not match with TEE.

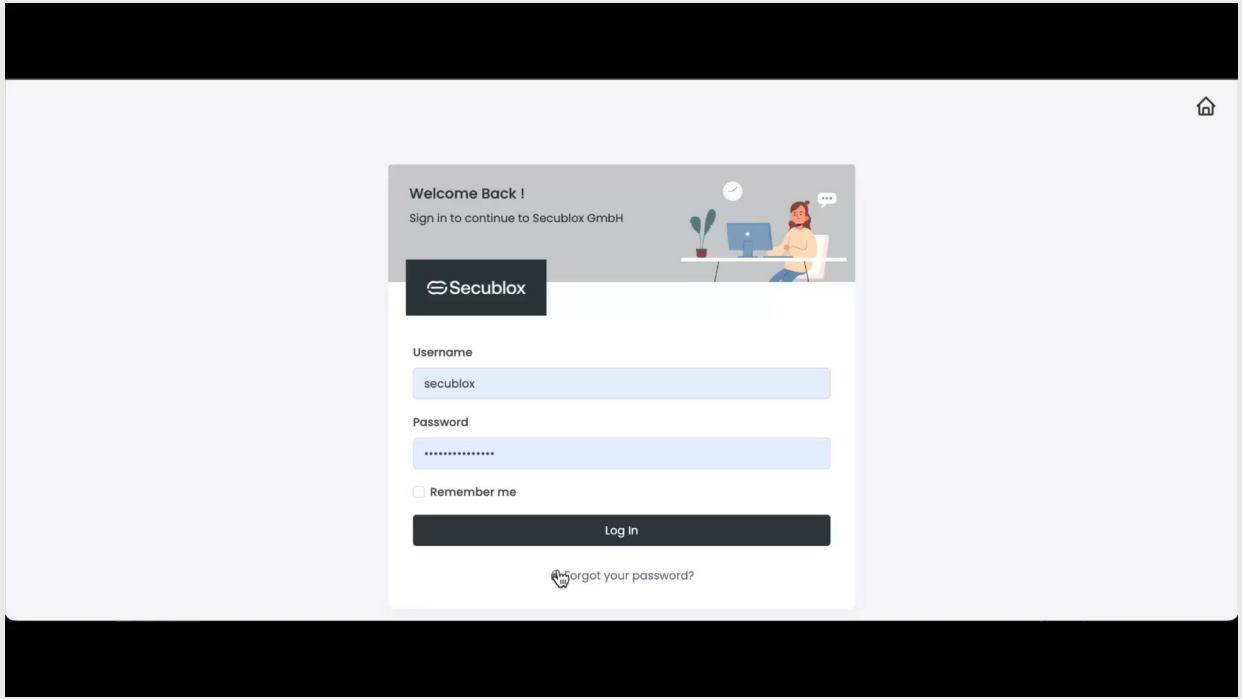
Watermarking consumes a lot of resources in the FPGA and the selected HW design could not be changed quickly enough.

TEE architecture by hardware design observed and feasible for Stage 2.

Watermarking feasible with FPGA that offers more resources.

Secublox

B/D DeepShield-Platform & Blockchain



Achievements after Stage 1 – Demonstration

In summary: Criteria to be fulfilled and what we developed

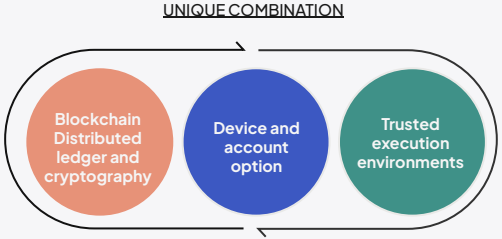
Funke Deep Fake criteria

The prototype must demonstrate how deepfakes images can be reliably detected and authenticated.	✓	The prototype reliably authenticates deepfake images using Trusted Execution Environments (TEEs) to embed unique cryptographic watermarks, which are verified via blockchain-managed public keys, ensuring secure and tamper-proof content origin tracking. Multi-layered integrity checks, including vector comparisons, semantic analysis, and object detection, detect unauthorized manipulations, while smart contracts automate key management. Scalable APIs and SDKs enable integration across platforms, ensuring adaptability to evolving deepfake techniques.
It can be AI-supported and should be able to continuously adapt to new deepfake techniques.	✓	The DeepShield approach is AI-supported and inherently adaptable to emerging deepfake techniques. Its multi-layered architecture combines watermarking for origin detection, blockchain for traceable metadata and ownership, and semantic integrity checks to identify meaning shifts. Each layer hosts diverse tools and can be extended, enabling continuous evolution and integration with external solutions when needed.
At least three different use cases (e.g. social media, news portals, video conferencing systems) will be demonstrated by the end of the process.	✓	The prototype addresses diverse use cases by enabling content verification across social media, news portals, and video conferencing. Through pre-upload detection APIs, it can authenticate images and videos in real-time, ensuring trusted content on platforms before distribution. Additionally, its multi-layered verification and adaptable infrastructure make it applicable across digital media, providing a consistent solution to verify content under various contexts.
Scalability and adaptability to different digital platforms are essential.	✓	Scalability and adaptability through modular, API-driven integration that allows it to work across different digital platforms seamlessly. Its cloud-based infrastructure and decentralized blockchain for key management support high data volumes, enabling efficient scaling across platforms.

The iphone-approach for anti-deep fake: Best of breed of technologies for the best solution

Just like the iPhone transformed the tech world by combining existing technologies like touchscreens, GPS, and cameras into a single, innovative device, DeepShield merges cryptographic watermarking, AI analysis, and blockchain verification into a unified solution.

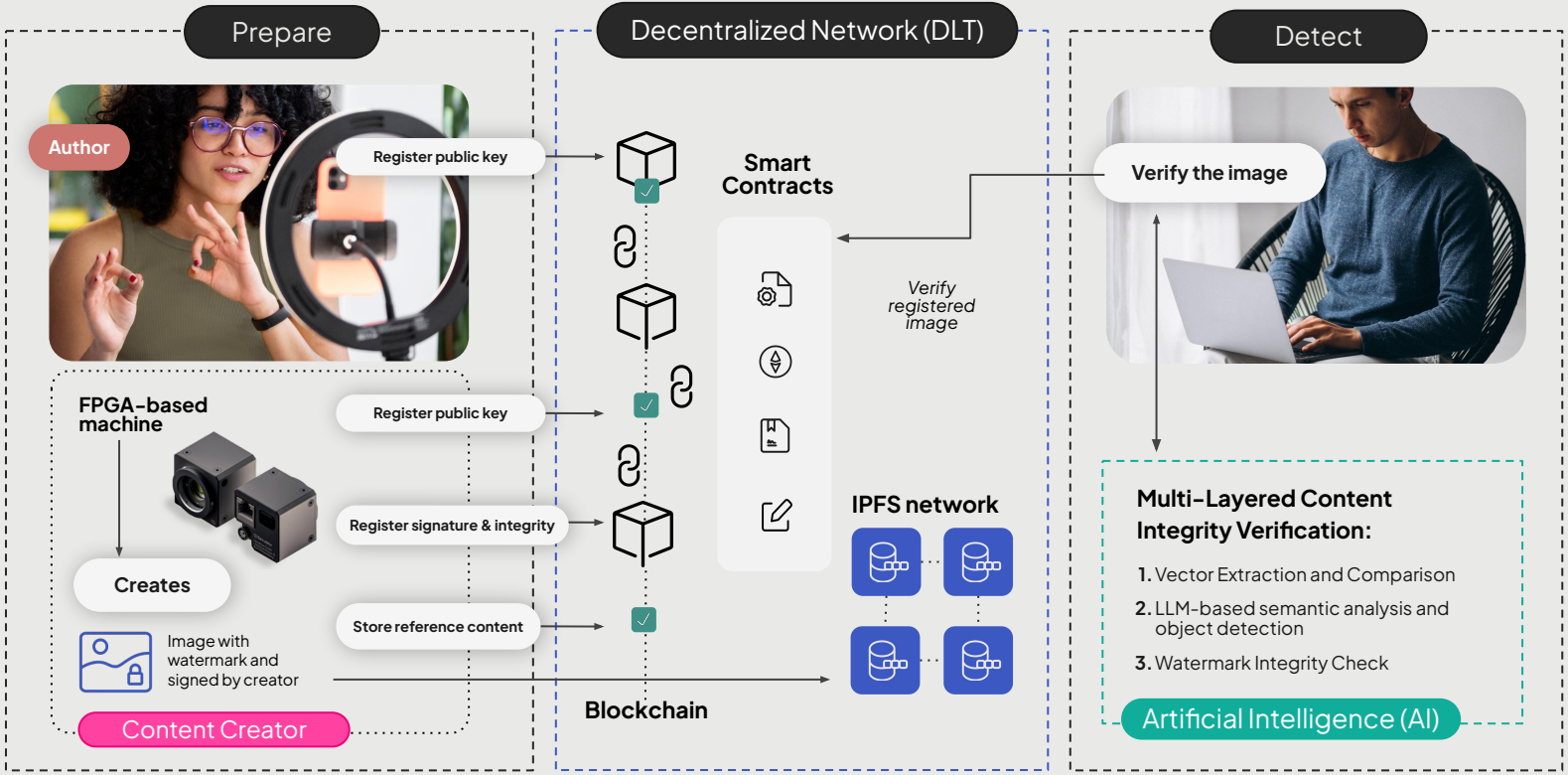
DeepShield:



- Trusted Execution Environments (TEEs)
- Decentralized Verification
- blind spread spectrum watermarking
- Immutable Record
- Distributed Ledger Technology (DLT)
- FPGA-based machine vision sensors for ultra-fast processing and customizable encryption
- Cryptographic techniques

Secublox

Content authenticity and deepfake prevention powered by a blockchain framework with a multi-layered approach to ensure media authenticity and detect manipulation.

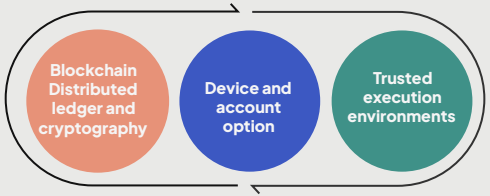


The iphone-approach for anti-deep fake: Best of breed of technologies for the best solution

Just like the iPhone transformed the tech world by combining existing technologies like touchscreens, GPS, and cameras into a single, innovative device, DeepShield merges cryptographic watermarking, AI analysis, and blockchain verification into a unified solution.

This novel integration offers flexibility, allowing for hardware enhancements such as Trusted Execution Environments (TEEs), and sets a new benchmark for digital media security. By blending the best-in-class components, DeepShield disrupts the market, creating a scalable, secure platform that reshapes content authenticity and integrity standards.

UNIQUE COMBINATION



eep security review
Full approach not on the edge
No Smartphone / FaceID support
No zero knowledge proofs
EU AI Act Compliance
Scalability / Latency

Trusted Execution Environments (TEEs)

Decentralized Verification

blind spread spectrum watermarking

Immutable Record

semantic integrity

Distributed Ledger Technology (DLT)

cryptographic techniques

FPGA-based machine vision sensors for ultra-fast processing and customizable encryption

Integration & Interoperability

Multi-layered Semantic Integrity

Combining AI with Blockchain brings many advantages for anti-deep fake: Heidari, A., Navimipour, N. J., Dag, H., Talebi, S., & Unal, M. (2024). A novel blockchain-based deepfake detection method using federated and deep learning models. Cognitive Computation, 16(3), 1073–1091. https://doi.org/10.1007/s12559-024-10255-7 and Yun, J., et al. (2024). DRPChain: A new blockchain-based trusted DRM scheme for image content protection. PLOS ONE, 19(9), e0309743. And Madushanka, T., Kumara, D.S., & Rathnaweera, A.A. (2024). SecureRights: A Blockchain-Powered Trusted DRM Framework for Robust Protection and Asserting Digital Rights.