Signing Right Away

Yejun Jang

May 26, 2024

1 Introduction

The advent of generative models has made hyperrealistic content creation accessible to internet users worldwide. Simple text prompts can now generate synthetic images [1–3], videos [4], and audio tracks [5], with OpenAI's DALL-E [2] and DALL-E 2 [3] available to the public. This raises concerns over disinformation and fraud, exemplified by a May 2023 incident where an AI-generated image of an explosion at the Pentagon was shared on a verified Twitter account, causing a significant dip in the S&P 500 index [6]. Various attempts have been made to address this "fake data" problem, including neural network classifiers, digital signing, and digital watermarking, each with its own strengths and limitations. These approaches are reviewed in Section 2. In Section 3, we introduce the concept of Signing Right Away (SRA), explain how it secures content provenance [7], and address several important security details of an SRA device. The roadmap for development is laid out in Section 4, and we conclude the article in Section 5.

2 Related Works

2.1 Neural Network Classifiers

One solution to combat generated content is using deep learning-based classifiers. A neural network classifier C is trained to discriminate real from fake data, predicting a Boolean value $r = C(x)$ to determine whether the content x is fake (or real). However, this approach struggles with heavily compressed content [8] and generalizing to different types of generators [9]. It also consumes massive computational resources during training. Major generative models, including Diffusion [10], Variational Autoencoders (VAEs) [11], Generative Adversarial Networks (GANs) [12], and Generative Flows (Glow) [13], efficiently approximate data distributions. GANs, for instance, already use classifiers (discriminator networks) in their construction, explaining why naive classification approaches struggle with GAN-based image generators like NeuralTextures [14] [9].

2.2 Digital Signing

Another solution is cryptographically binding digital content to a pair of keys and a digital signature, detectable through a verification algorithm. The keys consist of a public key pk and a private key sk. The signing algorithm S creates a tag $t = S(sk, x)$ for content x with the private key sk, and the tag and content are verified by computing $r = V(pk, x, t)$. Digital signing uses a cryptographic hash function, denoted $H(\cdot)$, to detect even slight alterations in content. This approach does not rely on statistical learning and is proven to enable secure communication, being actively adopted in content provenance standards [7].

2.3 Digital Watermarking

Digital watermarking embeds noise-like signals into images, detectable even after brightness, hue, or contrast changes. This technique complements digital signing by embedding an undetectable signal for tracing copyright infringements and authenticating banknotes [15]. The watermark remains functional after slight modifications, and can be used alongside digital signing methods [16].

3 Signing Right Away (SRA)

We propose a method to secure content provenance using encrypted communication and hardware isolation, named Signing Right Away (SRA). This approach uses an authenticated stream cipher to secure the connection between the sensor and the signal processor, which then digitally signs the sensory data inside a Trusted Execution Environment (TEE) as soon as the content is created.

The C2PA standard certifies the source and history (provenance) of media content. SRA aims to enhance the security of C2PA applications by addressing the lack of hardware security guidelines, preventing fake content injection into the camera-computer interface. Generative models are publicly available, making it crucial to protect the connections between the real and digital worlds with advanced hardware security standards. SRA addresses this problem.

In current systems, the camera module sends an unencrypted, raw bitstream to the motherboard via the Camera Serial Interface 2 (CSI-2) [17]. This design is vulnerable to attacks, where a simple HDMI to CSI-2 adapter [18] can inject fake data. SRA secures the connection between the camera and the Image Signal Processor (ISP) using authenticated encryption, signing the data inside a dedicated TEE as soon as it is created. Important details for implementing SRA include:

3.1 Untamperable Connection from the Sensor to its Encryption Chip

Minimizing the attack surface is crucial, as a single point of attack can nullify the verification system. Ensuring the connection from the sensor to its encryption chip is secure is essential. Potential hardware design choices include hard-wiring

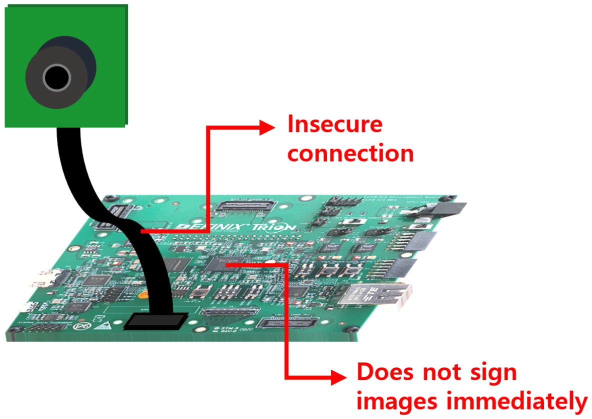

Figure 1: Main issues with the current camera-computer interface: insecure connection and lack of immediate signing by the image signal processor.

the sensor to the board, hiding connection pins inside the PCB, integrating the sensor with the encryption chip, and reducing the chipset manufacturing scale. Digital watermarking [19] [16] can detect fake data injection. Assuming secure connections, we explore and implement appropriate designs.

3.2 Secure Handling of the Private Key

Past incidents, like Elcomsoft hacking Canon's [20] and Nikon's [21] image verification systems, highlight the importance of securely handling private keys. The private keys in SRA should only be accessible inside the sensor's encryption chip and secure chamber. Collaborating with computer security experts can help address this issue.

3.3 Finite Rate of Signing, Incorporating Metadata

Setting a reasonable upper bound to the signing rate can control the spread of falsely signed content. Including metadata such as device ID, geolocation, and timestamp (with the creator's consent) enhances trustworthiness. This allows viewers to validate the information, giving more authenticity to the content.

3.4 Secure and Efficient Image Processing and Signing

Before verification, the image must be securely processed, converted to a popular format (e.g., JPEG), and signed. Using system DRAM with encryption for memory protection, as in [22], ensures real-time processing of high-bitrate video feeds. Memory encryption prevents DMA (Direct Memory Access) attacks that could inject arbitrary images for signing.

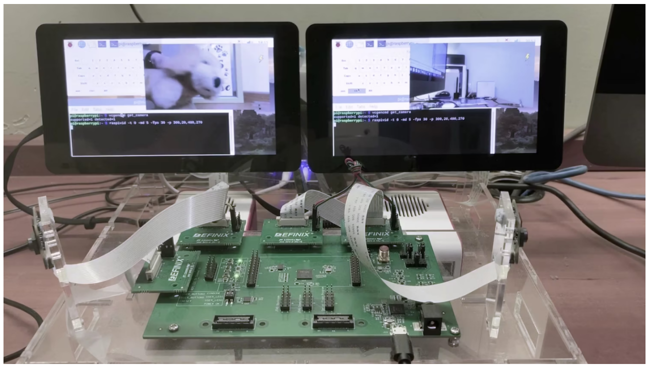

Figure 2: The Trion T20 MIPI Development Kit with a dual camera feed example. The Trion T20 is an FPGA developed by Efinix. It provides hardened MIPI CSI-2 and D-PHY interfaces, which can be used to connect camera modules such as the Raspberry Pi Camera Module.

4 Development Plans

The development procedure consists of five stages:

4.1 Simple Streaming, Image Preprocessing and Compression

We begin by implementing a simple video stream from the camera module to the Raspberry Pi, using the Trion T20 as a bridge. The raw Bayer image from the CSI-2 video stream will be formatted to an RGB tensor using OpenCV and displayed on the monitor. Snapshots will be converted to JPEG using PIL (Python Imaging Library) upon pressing a button. This stage helps in learning about the CSI-2 interface and establishing a clear implementation timeline.

4.2 Encrypted Streaming using Stream Cipher

The second step involves implementing the Salsa20 stream cipher on the Trion T20 FPGA. We decrypt the stream using libraries like OpenSSL or libsodium and preprocess it. Salsa20, developed by Daniel J. Bernstein [23], is one of the most secure stream ciphers, offering decent performance in hardware and software.

4.3 Signing Content using the C2PA Rust Library

Next, we sign the generated snapshot using the C2PA Rust Library. Initially, we implement C2PA signing with self-signed certificates, ensuring the signed image can be verified on another device. We then update the system to sign the image with trusted certificates from a known Certificate Authority (CA), and delve into understanding the library's inner workings to identify hardware-implementable modules.

4.4 Secure Chamber Design

We design the secure chamber using another FPGA, incorporating decryption, preprocessing, and C2PA signing functionalities. We consider critical details like secure handling of private keys, finite signing rates, and secure, efficient image processing and signing. To enhance security, we may design an encrypted preprocessing scheme ensuring decrypted preprocessed data matches preprocessed data from encrypted inputs.

4.5 Designing ASICs and Continual Improvement

We improve the prototypes, designing Application Specific Integrated Circuits (ASICs) for the encryption chip and secure chamber. Ensuring untamperable connections from the sensor to the encryption chip, possibly by integrating the sensor with the encryption module or implementing digital watermarking. We compile designs, fabricate, and test functionality in a PCB, and eventually implement the system in a mobile phone prototype. Journalists, police officers, and investigators may beta test the improved prototype, leading to integration as a new mobile phone feature.

5 Conclusion

In this era of rapidly evolving digital content creation and distribution, the integrity and authenticity of media have become paramount. With the democratization of sophisticated generative models, the potential for misuse in creating and spreading disinformation is significant. In response, our work has introduced Signing Right Away (SRA), a novel approach that melds hardware-based solutions with cryptographic techniques to safeguard the provenance of digital content right from its inception.

By leveraging encrypted communication and hardware isolation, SRA ensures that content is authenticated at the source, thereby minimizing the risk of tampering and unauthorized modifications. This approach not only enhances the security and trustworthiness of digital media but also aligns with the C2PA standard's objectives to establish a robust framework for content provenance and authenticity.

Our proposed system underscores the importance of securing the digital ecosystem against potential threats and highlights the role of technological innovation in preserving the integrity of content. As we move forward, the implementation of SRA in devices will be a critical step towards mitigating the risks associated with fake data, thereby fostering a safer and more trustworthy digital environment.

As we embark on the journey of developing and refining this technology, the collaboration between researchers, industry stakeholders, and regulatory bodies will be crucial. Together, we can navigate the challenges and opportunities that lie ahead, ensuring that the digital realm remains a space for authentic and reliable information. The development of SRA represents a significant milestone in our quest to protect the sanctity of digital content, and we are optimistic about its potential to shape a more secure and trustworthy digital future.

References

[1] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," 2021. Accessed: 2024-03-26.

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents." `https://arxiv.org/abs/2204.06125`, 2022. Accessed: 2024-03-26.

[3] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, W. Manassra, P. Dhariwal, C. Chu, Y. Jiao, and A. Ramesh, "Improving image generation with better captions." `https://cdn.openai.com/papers/dall-e-3.pdf`, 2023. Accessed: 2024-03-26.

[4] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh, "Video generation models as world simulators," 2024.

[5] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "Audiolm: a language modeling approach to audio generation," 2023.

[6] D. O'Sullivan and J. Passantino, "'verified' twitter accounts share fake image of 'explosion' near pentagon, causing confusion," 2023. Updated 11:35 AM EDT, Tue May 23, 2023.

[7] Coalition for Content Provenance and Authenticity (C2PA), "C2PA Specifications 1.3." `https://c2pa.org/specifications/specifications/1.3/index.html`, 2023. Accessed: 2024-03-26.

[8] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," 2019.

[9] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? understanding properties that generalize," 2020.

[10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020.

[11] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.

[13] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," 2018.

[14] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," 2019.

[15] Google DeepMind, "Identifying ai-generated content with synthid." `https://deepmind.google/technologies/synthid/`, 2023. Accessed: 2024-03-27.

[16] O. K. Lapidot, "The first camera to embed content authenticity initiative's "digital watermark" – leica m11-p," *CineD*, October 2023.

[17] MIPI Alliance, "MIPI Camera Serial Interface 2 (CSI-2®)." `https://www.mipi.org/specifications/csi-2`, Mar. 2024. Accessed: 2024-03-26, public version is not available.

[18] Xiaokudedain Store, "Raspberry Pi HDMI-in Module, to CSI-2 C779, Input TC358743 Supports up to 1080p25fps for Raspberry Pi 4B." `https://www.aliexpress.com/item/1005005515591253.html`, March 2024. Accessed: 2024-03-27.

[19] G. Nelson, G. Jullien, and O. Yadid-Pecht, "Cmos image sensor with watermarking capabilities," in *2005 IEEE International Symposium on Circuits and Systems*, pp. 5326–5329 Vol. 5, 2005.

[20] ElcomSoft Co. Ltd., "Canon Original Data Security System Compromised: ElcomSoft Discovers Vulnerability." `https://www.elcomsoft.com/PR/canon_101130_en.pdf`, 11 2010. Accessed: 2024-03-26.

[21] ElcomSoft Co. Ltd., "ElcomSoft Discovers Vulnerability in Nikon's Image Authentication System." `https://www.elcomsoft.com/PR/nikon_110428_en.pdf`, April 2011. Accessed: 2024-03-26.

[22] A. Inc., "Secure enclave." `https://support.apple.com/guide/security/secure-enclave-sec59b0b31ff/web`, 2024. Accessed: 2024-03-28.

[23] D. J. Bernstein, "The salsa20 family of stream ciphers," tech. rep., Department of Mathematics, Statistics, and Computer Science, The University of Illinois at Chicago, Chicago, IL, 12 2007. Available at `https://cr.yp.to/snuffle/salsafamily-20071225.pdf`.