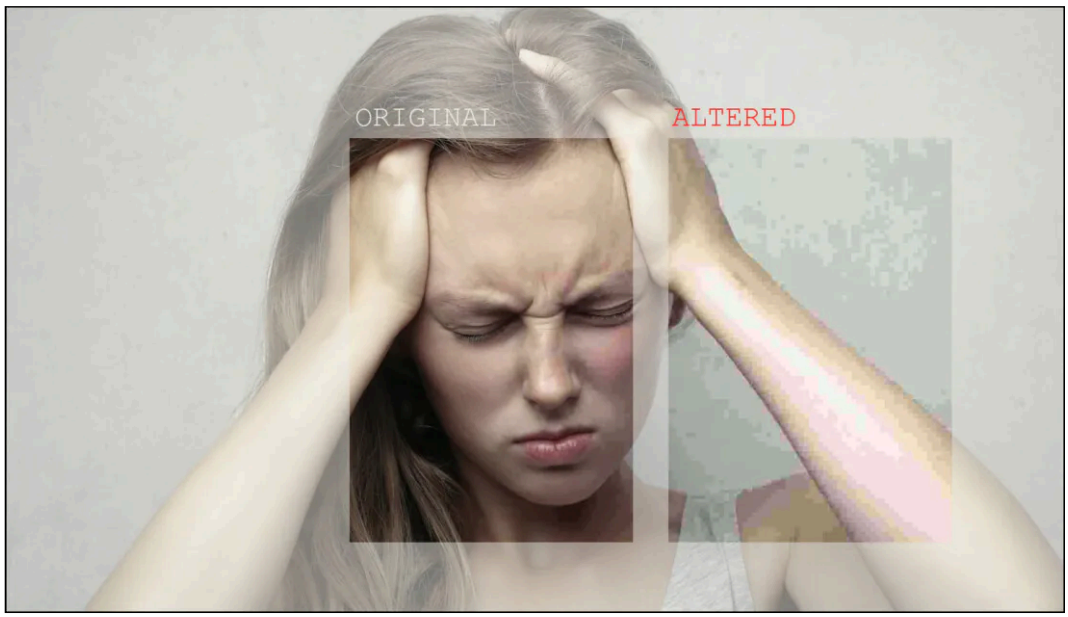
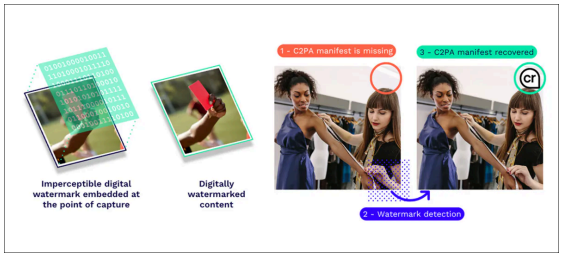
Self-Authenticating Images Through Simple JPEG Compression

Originally published at https://www.unite.ai/self-authenticating-images-through-simple-jpeg-compression/



Concerns about the risks posed by tampered images have been showing up regularly in the research over the past couple of years, particularly in light of a new surge of AI-based image-editing frameworks capable of amending existing images, rather than creating them outright.

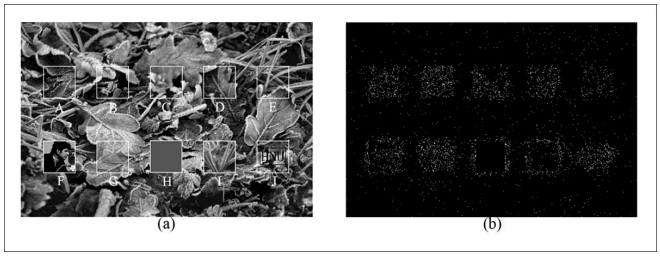
Most of the proposed detection systems addressing this kind of content fall into one of two camps: the first is *watermarking* – a fallback approach built into the image veracity framework now being promoted by the Coalition for Content Provenance and Authenticity (C2PA).



The C2PA watermarking procedure is a fallback, should the image content become separated from its original and ongoing provenance 'manifest'. Source: https://www.imatag.com/blog/enhancing-content-integrity-c2pa-invisible-watermarking

These 'secret signals' must subsequently be robust to the automatic re-encoding/optimization procedures that often occur as an image transits through social networks and across portals and platforms – but they are often not resilient to the kind of lossy re-encoding applied through JPEG compression (and despite competition from pretenders such as webp, the JPEG format is still used for an estimated 74.5% of all website images).

The second approach is to make images tamper-evident, as initially proposed in the 2013 paper *Image Integrity Authentication Scheme Based On Fixed Point Theory*. Instead of relying on watermarks or digital signatures, this method used a mathematical transformation called *Gaussian Convolution and Deconvolution* (GCD) to push images toward a stable state that would break if altered.



From the paper 'Image Integrity Authentication Scheme Based On Fixed Point Theory': tampering localization results using a fixed point image with a Peak Signal-to-Noise (PSNR) of 59.7802 dB. White rectangles indicate the regions subjected to attacks. Panel A (left) displays the applied modifications, including localized noise, filtering, and copy-based attacks. Panel B (right) shows the corresponding detection output, highlighting the tampered areas identified by the authentication process. Source: https://arxiv.org/pdf/1308.0679

The concept is perhaps most easily understood in the context of repairing a delicate lace cloth: no matter how fine the craft employed in patching the filigree, the repaired section will inevitably be discernible.

This kind of transformation, when applied repeatedly to a grayscale image, gradually pushes it toward a state where applying the transformation again *produces no further change*.

This stable version of the image is called a *fixed point*. Fixed points are rare and highly sensitive to changes – any small modification to a fixed point image will almost certainly break its *fixed* status, making it easy to detect tampering.

As usual with such approaches, the artefacts from JPEG compression can threaten the integrity of the scheme:

On the left, we see a watermark applied to the face of the iconic 'Lenna' (Lena) image, which is clear under normal compression. On the right, with 90% JPEG compression, we can see that the distinction between the perceived watermark and the growth of JPEG noise is lowering. After multiple resaves, or at the highest compression settings, the majority of watermarking schemes face issues with JPEG compression artefacts. Source: https://arxiv.org/pdf/2106.14150

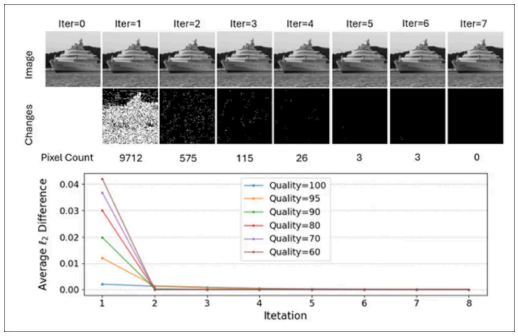
What if, instead, JPEG compression artefacts could actually be used as the central means of obtaining a fixed point? In such a case, there would be no need for extra bolt-on systems, since the same mechanism that usually causes trouble for watermarking and tamper detection would instead form the basis of tamper detection framework itself.

JPEG Compression as a Security Baseline

Such a system is put forward in a new paper from two researchers at the University of Buffalo at the State University of New York. Titled *Tamper-Evident Image Using JPEG Fixed Points*, the new offering builds on the 2013 work, and related works, by officially formulating its central principles, for the first time, as well as by ingeniously leveraging JPEG compression itself as a method to potentially produce a 'self-authenticating' image.

The authors expand:

> *'The study reveals that an image becomes unchanged after undergoing several rounds of the same JPEG compression and decompression process.*
>
> *'In other words, if a single cycle of JPEG compression and decompression is considered a transformation of the image, referred to as a JPEG transform, then this transform exhibits the property of having fixed points, i.e., images that remain unaltered when the JPEG transform is applied.'*



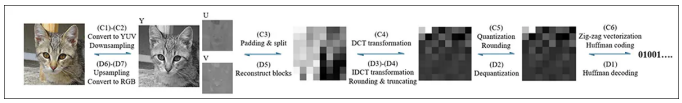
From the new paper, an illustration of JPEG fixed point convergence. In the top row we see an example image undergoing repeated JPEG compression, with each iteration showing the number and location of changing pixels; in the bottom row, the pixel-wise L2 distance between consecutive iterations is plotted across different compression quality settings. Ironically, no better resolution of this image is available. Source: https://arxiv.org/pdf/2504.17594

Rather than introducing external transformations or watermarks, the new paper defines the JPEG process itself as a dynamic system. In this model, each compression and decompression cycle moves the image toward a fixed point. The authors prove that, after a finite number of iterations, any image either reaches or approximates a state where further compression will produce no change.

The researchers state*:

> *'Any alterations to the image will cause deviations from the JPEG fixed points, which can be detected as changes in the JPEG blocks after a single round of JPEG compression and decompression…*
>
> *'The proposed tamper-evident images based on JPEG fixed points have two advantages. Firstly, tamper-evident images eliminate the need for external storage of verifiable features, as required by image fingerprinting [schemes], or the embedding of hidden traces, as in image watermarking methods. The image itself serves as its proof of authenticity, making the scheme inherently self-evident.*
>
> *'Secondly, since JPEG is a widely-used format and often the final step in the image processing pipeline, the proposed method is resilient to JPEG operations. This contrasts with the original [approach] that may lose integrity traces due to JPEG.'*

The paper's key insight is that JPEG convergence is not just a byproduct of its design but a mathematically inevitable outcome of its operations. The discrete cosine transform, quantization, rounding, and truncation together form a transformation that (under the right conditions) leads to a predictable set of fixed points.

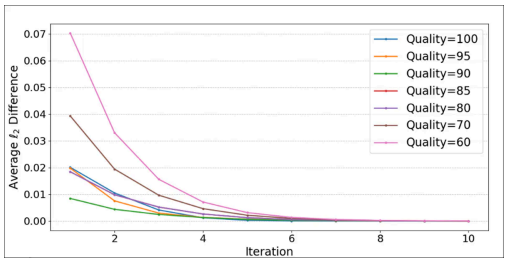


Schema for the JPEG compression/decompression process formulated for the new work.

Unlike watermarking, this method requires *no embedded signal*. The only reference is the image's own consistency under further compression. If recompression produces no change, the image is presumed authentic. If it does, tampering is indicated by the deviation.

Tests

The authors validated this behavior using one million randomly generated eight-by-eight patches of eight-bit grayscale image data. By applying repeated JPEG compression and decompression to these synthetic patches, they observed that convergence to a fixed point occurs within a finite number of steps. This process was monitored by measuring the pixel-wise L2 distance between consecutive iterations, with the differences diminishing until the patches stabilized.



L2 difference between consecutive iterations for one million 8×8 patches, measured under varying JPEG compression qualities. Each process begins with a single JPEG-compressed patch and tracks the reduction in difference across repeated compressions.

To evaluate tampering detection, the authors constructed tamper-evident JPEG images and applied four types of attacks: *salt and pepper* noise; *copy-move* operations; *splicing from external sources*; and *double JPEG compression* using a different quantization table.

Example of fixed point RGB images with detection and localization of tampering, including the four disruption methods used by the authors. In the bottom row, we can see that each perturbation style betrays itself, relative to the generated fixed-point image.

After tampering, the images were re-compressed using the original quantization matrix. Deviations from the fixed point were detected by identifying image blocks that exhibited non-zero differences after recompression, enabling both detection and localization of tampered regions.

Since the method is based entirely on standard JPEG operations, fixed point images work just fine with regular JPEG viewers and editors; but the authors note that if the image is recompressed at a different quality level, it can lose its fixed point status, which could break the authentication, and needs to be handled carefully in real-world use.

While this isn't just a tool for analyzing JPEG output, it also doesn't add much complexity. In principle, it could be slotted into existing workflows with minimal cost or disruption.

The paper acknowledges that a sophisticated adversary might attempt to craft adversarial changes that preserve fixed point status; but the researchers contend that such efforts would likely introduce visible artifacts, undermining the attack.

While the authors do not claim that fixed point JPEGs could replace broader provenance systems such as C2PA, they suggest that fixed point methods could complement external metadata frameworks by offering an additional layer of tamper evidence that persists even when metadata is stripped or lost.

Conclusion

The JPEG fixed point approach offers a simple and self-contained alternative to conventional authentication systems, requiring no embedded metadata, watermarks, or external reference files, and instead deriving authenticity directly from the predictable behavior of the compression process.

In this way, the method reclaims JPEG compression – a frequent source of data degradation – as a mechanism for integrity verification. In this regard, the new paper is one of the most innovative and inventive approaches to the problem that I have come across over the past several years.

The new work points to a shift away from layered add-ons for security, and toward approaches that draw on the built-in characteristics of the media itself. As tampering methods grow more sophisticated, techniques that test the image's own internal structure may start to matter more.

Further, many alternative systems proposed to address this problem introduce significant friction by requiring changes to long-established image-processing workflows – some of which have been operating reliably for years, or even decades, and which would demand a far stronger justification for retooling.

* *My conversion of the authors' inline citations to hyperlinks.*

First published Friday, April 25, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More