

Identification, Not Detection: Blind Owner Identification and Tamper Localization for Images with an Analytic False-Owner Bound

Adam Weisman

Vereye

October 2026

Abstract

Image watermarks are evaluated as detectors: is *some* mark present, at a tolerated false-alarm rate? Provenance needs three stronger answers from a copy alone: *who* owns the image, *where* it no longer matches its registration, and a computed bound on ever naming a wrong owner. Vereye answers all three blindly and without machine learning. A 64-bit owner key is sealed in one deterministic pass. A copy that was cropped, rotated, rescaled, re-encoded, re-coloured or pasted into another picture is read with a geometric solve computed in closed form from a key-free lattice, followed by a bounded number of rectifications and soft decodes (at most 1,771 checks of the 24-bit code). The regenerated mark is then compared cell by cell, localizing edits as bit-level facts. On unmarked images, the probability of naming a registered owner is bounded below $6 \cdot 10^{-13}$ per verification for 10^{11} registered images, under a stated uniformity assumption on the 24-bit check; no key at all was decoded from 46,467 attacked copies of unmarked images under the two protocols. On our implementation of the SynthID-Image protocol (all 29 transformations at its 512 px) Vereye identifies the exact owner in 99.70 % / 95.80 % of cases (random / worst) on 300 photos never used, with the current reader; on the first 100 held-out photos, under a rule fixed before that run and with an earlier reader, 99.0 % / 86.4 %, behind SynthID-O’s reported values at both strengths; at 1024 px, not the protocol’s size, 99.90 % / 98.93 % on 100 held-out photos; on the 50 photos the open methods ran on, Vereye identifies 99.2 / 86.2 % while WAM and TrustMark, at the protocol’s 0.1 % calibration, detect 92.7 / 81.7 % and 82.6 / 72.5 %; both score above the values the SynthID-Image paper reports for them, so our attacks are probably milder than its. On the WAM protocol it identifies 98.74 % (COCO, reader of 31 October 2026) and 99.05 % (DIV2K at 2K, reader of 27 October 2026) of copies over its 35 attack settings with no false owner; WAM and TrustMark, run with their published thresholds on the same attacked unmarked copies, flag 0.9–2.7 % of them (WAM reports 0.04 % on unattacked images; Section 5.3 discusses the difference). Vereye fails most on worst-strength combined attacks (11, 0, 88, 8 % on the four held-out combinations) and identifies 100 % of greyscale copies.

1 Introduction

Regulation now requires synthetic media to be marked in a machine-readable way (EU AI Act, Art. 50) [17]. Platforms, newsrooms, insurers and courts need to know who published an image and whether it was altered after it passed through screenshots, social-media recompression and edits. Provenance metadata such as C2PA [18] carries this information, but most of these operations remove it. A watermark survives them only if it can be read *blindly*, from the copy alone.

Learned watermarks (HiDDeN, StegaStamp, RivaGAN, TrustMark, WAM, Video Seal) and Google’s SynthID have made large gains in robustness [1–7]. They are designed and evaluated as *detectors*: a mark is declared present when enough of its bits agree, at a chosen false-positive rate, typically 0.1 % [3]. At internet scale, a 0.1 % false-positive rate means a million wrong flags per billion images. A detector that recovers only part of a message cannot name an owner. Run with their published thresholds on identical attacked images, WAM and TrustMark flag 0.9–2.7 % of *unmarked* copies (WAM reports 0.04 % on unattacked images; Section 5.3), and recover their full message in only 67–71 % of marked ones (Section 5).

This paper addresses these three questions. For any copy, Vereye returns the exact owner or nothing, a map of what changed, and a verdict whose wrong-owner probability is bounded analytically. It needs no neural network, training data or GPU, and does not store the original. Our contributions are:

1. **Identification under geometry.** A deterministic watermark whose geometry is *computed* from a key-free synchronization structure with a fixed set of transform-domain operations; the remaining ambiguities (the lattice’s symmetries, mirror and the tile offset) are resolved by a bounded number of decodes.
2. **An analytic false-owner bound** that holds at registry scale (Eq. (4)) and is consistent with zero keys decoded from 46,467 attacked copies of unmarked images.
3. **Tamper localization without the original.** Once the owner is identified, every cell of the copy is tested against the regenerated mark and reported as read, not read or unobservable (Figure 1).
4. **A private registry** whose records are signed with a post-quantum signature. It stores no image and nothing from which one can be rebuilt.
5. **A comparison on identical pixels** under two published protocols. Every method decodes identical attacked pixels, and each competitor’s own published numbers are shown beside our measurements.

2 Related work

Learned watermarking. HiDDeN [1] trained an encoder–decoder pair through differentiable noise layers. StegaStamp [5] added physical robustness, and RivaGAN [6] attention-based embedding. TrustMark [2] made such marks resolution-independent. WAM [4] predicts a per-pixel mask and localized messages, and Video Seal [7] extends the approach to video. SynthID-Image [3] watermarks generated images at internet scale and defines the evaluation protocol we reproduce. Marks rooted in the generator, such as Stable Signature [26] and Tree-Ring [27], and post-hoc marks such as InvisMark [28] are evaluated in the same way. All of these are detectors: the decision is a threshold on bit agreement, calibrated to a false-positive rate. **Classical watermarking.** Spread-spectrum embedding [8] and quantization index modulation [9] established the communication view of watermarking [11]. Template and Fourier–Mellin methods [10,12] and autocorrelation self-reference [24] made marks resilient to affine maps; periodic templates are in turn exposed to template-estimation [32] and copy attacks [25]. **Content authentication.** Fragile and semi-fragile marks localize edits but must be read against a reference or break under recompression [13,14]. Learned schemes such as EditGuard [22] and OmniGuard [23] embed a localization watermark and predict a tamper mask with a network; they report masks and bit accuracy, not an identified owner with a bounded false-match rate. **Lookup without a watermark.** Perceptual-hash registries such as PhotoDNA and NeuralHash [33,34] find an image by a robust content hash; they need no embedding, but match by a similarity threshold rather than by an exact key. Vereye combines a robust identifier with a reference regenerated from the identified key, so localization survives re-encoding and geometry while no image is stored.

3 Problem and verdicts

A verification receives a single file and returns one of three verdicts. **Exact match:** every pixel of the file matches a registered seal. A cryptographic fingerprint of the pixels decides this, and the registry holds the fingerprint, never the image. **Owner identified:** the watermark names a registered owner, and the report states where the copy differs from what was sealed. **We could not read a seal:** no registered owner can be named, and the system does not guess. This does not mean the image was never sealed. Table 1 contrasts these capabilities with existing approaches.

Scope and threat model. The mark is designed to survive the processing a published image undergoes in good faith or by negligence: re-encoding, resizing, cropping, rotation and perspective, screenshots, filters, colour and tone changes, overlays and local edits. Removal by an adversary who knows the scheme, regeneration of the image by a generative model, photographs of screens and print are out of scope; under them the reader is designed to report that it could not read a seal (this has not been measured), and naming a wrong owner by chance is bounded by Eq. (4). Periodic marks are also exposed to template-estimation

Table 1. Capabilities. “Partial”: the method recovers a short message with a bit-error tolerance; owner identification then depends on an external lookup with no stated false-match bound. Vereye’s entries are qualified by its measured limits: it fails on worst-strength combined attacks (Section 8), and its cell-level localization has a mean IoU of 0.48 with the edited area (Table 7).

Capability	Vereye	SynthID	WAM	TrustMark	C2PA
Names the exact owner from a copy	✓	detects only	partial	partial	until stripped
Survives crop, rotation, rescale, re-encode	✓	✓	✓	partly	—
Locates edits without the original	cell level	—	mark extent	—	—
Analytic wrong-owner bound	$< 6 \cdot 10^{-13}$	0.1 % FPR	—	—	signature
Marks any image, not only generated ones	✓	✓ (post hoc)	✓	✓	✓
No neural network, training or GPU	✓	—	—	—	✓
Post-quantum signed records	✓	n/a	n/a	n/a	classical
No image stored	✓	n/a	n/a	n/a	n/a

Table 2. Threat model: adversary classes, whether this paper covers them, and what holds.

Adversary or event	Scope	What holds
Benign processing: re-encoding, resizing, crop, rotation, perspective, screenshots, filters, tone and colour	in scope	identification measured in Section 5; measured limits in Section 8
Local edits of a sealed image	in scope	owner identified and edits localized, measured in Section 6
Chance: an unmarked or foreign image	in scope	naming a registered owner bounded by Eq. (4)
A seal copied onto another picture	partly	the registered owner is named; not bounded by Eq. (4); the report’s difference from the registered photo is not measured
Informed removal, template estimation	out of scope	designed to end in <i>could not read a seal</i> ; not measured
Regeneration by a generative model, screen photographs, print	out of scope	designed to end in <i>could not read a seal</i> ; not measured
Forging or altering registry records	in scope	records signed with ML-DSA-65; tree heads signed with ML-DSA-65 and Ed25519 and pinned; no witnesses (Section 4.7)
Enumerating the registry online	in scope	rate-limited lookups through a separate key service (Section 4.7)
Compromise of the registry host, or of the store together with the key service	partly	recently seen keys and the registration of known images are exposed; offline key search becomes possible (Section 4.7)

and removal attacks [32], which are not evaluated here. A seal can, however, be copied onto another picture [25]: the reader then names the registered owner, and the report shows how far the picture is from the registered photo. Eq. (4) bounds chance coincidences of keys only; misattribution through a copied seal is not bounded by it, and how often the report flags a copied seal as differing from the registered photo has not been measured. Records are signed and kept in a Merkle log whose signed heads are pinned (witnesses are future work), so a copied seal cannot add or change a record. Table 2 summarizes the threat model.

4 Method

We state the physics and mathematics the design rests on; implementation parameters are withheld.

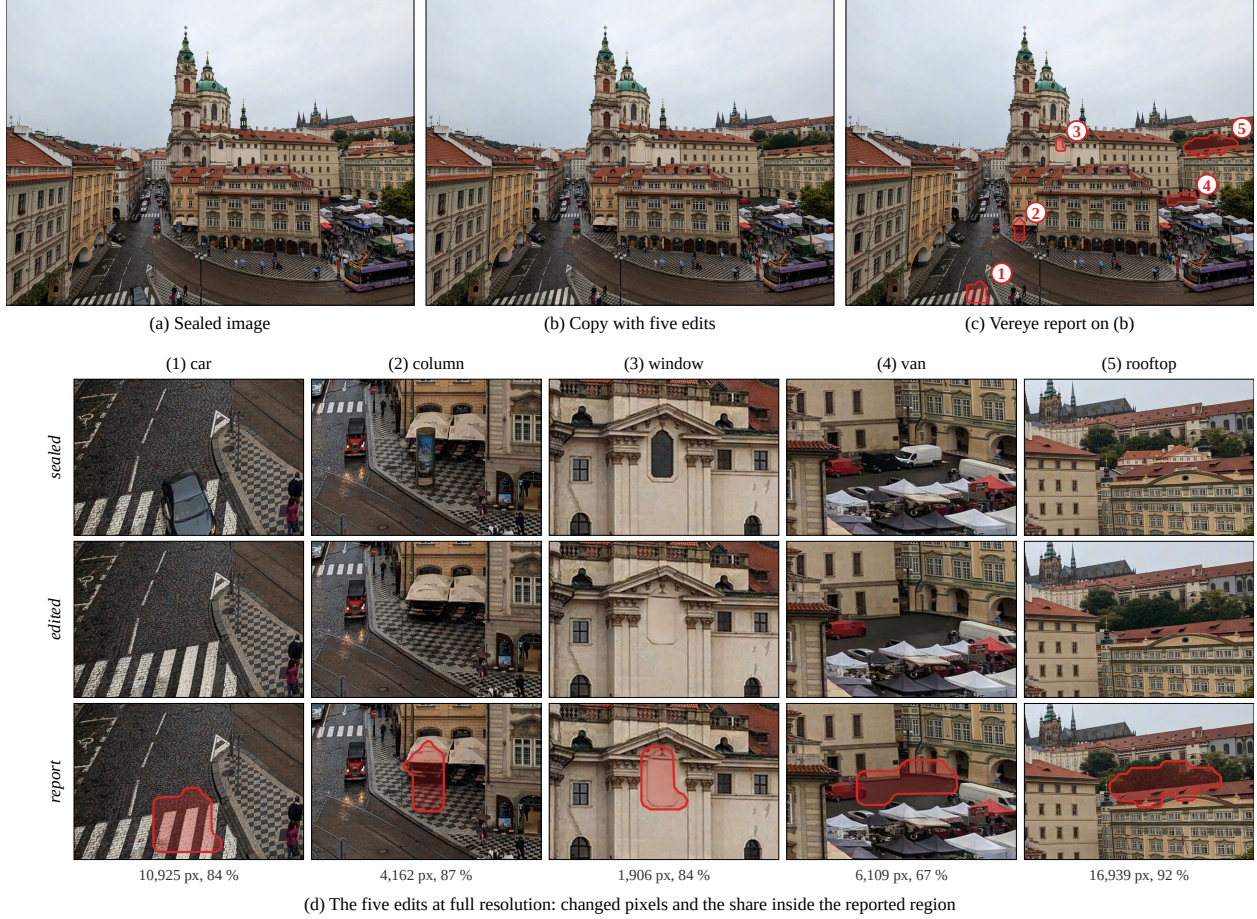


Figure 1. One blind verification of a hand-edited copy. (a) A sealed photograph. (b) A copy in which five objects were removed by hand in an image editor: a car on the crossing, an advertising column, a church window, a van and a rooftop. Together they change 40,041 pixels (1.3 % of the image) and are hard to see. (c) Vereye’s report on (b), computed from the copy alone: the owner key is read, the mark the owner wrote is regenerated, and the regions where it no longer reads back are tinted. All five edits are reported, and nothing is reported more than 48 px from an edit on this photograph. (d) Each edit at full resolution, numbered as in (c), with its changed pixels and the share of them inside the reported region.

4.1 Channel

The sealed image is $x + w$, where w is a keyed, deterministic signal shaped by a perceptual model. A copy is

$$y = \mathcal{A}(x + w) + n, \quad (1)$$

where $\mathcal{A}$ is an unknown composition of a geometric map H (crop, rotation, scale, perspective, mirror) and photometric operations (tone, colour, filtering, lossy coding), and n collects quantization and noise. The host x is unknown to the reader and is by far the largest term. The design problem is to make w invisible to the eye yet separable from x after $\mathcal{A}$.

4.2 Visibility budget

The mark has luminance, chrominance and synchronization components. For each component c , let $p_c(x)$ be the mean magnitude of its change in CIELAB, weighted by the local Jacobian of CIELAB with respect to the channel it changes. The strengths s_c are set per image so that the pooled visibility meets one budget,

$$\left(\sum_c (s_c p_c(x))^\beta \right)^{1/\beta} \leq \tau(x), \quad (2)$$

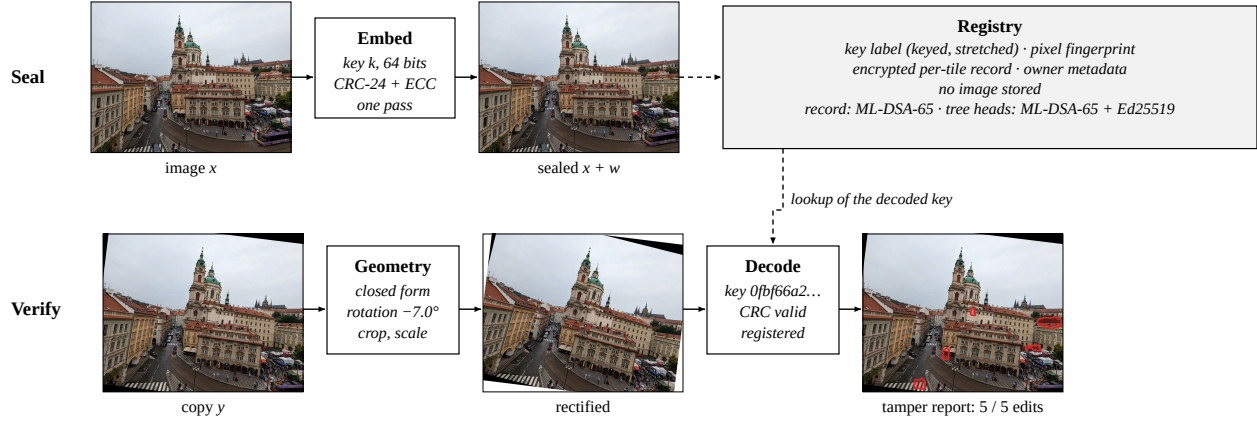


Figure 2. Sealing and verification on real data. *Seal*: one deterministic pass writes the mark for a random 64-bit key; the registry receives a signed record that holds no image. *Verify*: the copy of Figure 1(b), additionally rotated, cropped and re-encoded, is read blindly. The geometry is solved in closed form up to the lattice’s symmetries, which a bounded number of decodes resolves, and the key is accepted only because its integrity check passes and it is registered. The tamper report then compares the copy with the regenerated mark in the copy’s own coordinates.

with a Minkowski exponent β for the summation of luminance and colour visibility, in the tradition of perceptual masking models [15, 16]. Equation (2) is stated schematically: the threshold $\tau(x)$, the exponent β and the Jacobian weights are among the withheld parameters. Smooth regions carry chrominance only. The synchronization component is written along the direction of the colour-difference plane where the eye is least sensitive. The strengths are set once per image from this budget, so the same visibility holds on smooth skies and on textured streets (Figure 9).

4.3 Geometry from a key-free lattice

The mark is periodic over a grid of square tiles. A tile-periodic signal has a discrete spectrum on a lattice Λ^* . Under an affine map $u \mapsto Au + t$, which moves a point u of the sealed image to $Au + t$ in the copy, the lattice and the phases of its frequencies k transform in closed form:

$$\Lambda^* \mapsto A^{-\top} \Lambda^*, \quad \varphi_k \mapsto \varphi_k - 2\pi \langle A^{-\top} k, t \rangle. \quad (3)$$

The linear part is therefore solved from where the spectral energy sits, and the translation from its phases. The lattice is found with a fixed set of transform-domain operations rather than a sweep over angles and scales, in the tradition of template and Fourier–Mellin registration [10, 12] and autocorrelation self-reference [24]. A lattice determines the pose only up to its symmetry group (8 orientations of a square lattice, mirror, and the integer tile offset); these hypotheses, and a few photometric variants of the soft values, are decided by the CRC within the budget C of Eq. (4). Perspective is solved from local estimates of the same quantities.

4.4 Identification

The 64-bit key k is extended by a 24-bit cyclic redundancy check (CRC) and an error-correcting code, and the codeword is repeated in every tile, so any visible region that covers one full tile, or partial tiles that together cover every code bit, carries all of it. After rectification the reader computes, for each code bit c_i , a log-likelihood ratio $\ell_i = \log p(r \mid c_i = 0) - \log p(r \mid c_i = 1)$ from the received samples r . It pools these over all visible tiles, including partial ones, and list-decodes [29, 30]. A candidate $\hat{k}$ is accepted only if

$$\text{CRC}(\hat{k}) = 0 \quad \wedge \quad \hat{k} \in \mathcal{R},$$

where $\mathcal{R}$ is the set of registered keys. No threshold on bit agreement and no score enters this acceptance. Synchronization gates, calibrated on unmarked images, decide only whether a decode is attempted: they can cause a miss and they bound C , but they never name an owner.

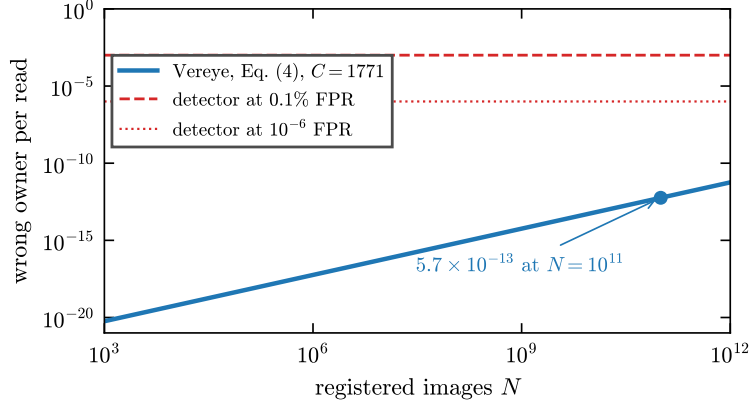


Figure 3. The wrong-owner probability of Eq. (4) against registry size (an upper bound for an unmarked image), next to the nominal per-read false-alarm rates at which detectors are usually operated. Those rates bound a different event, declaring a mark present; the comparable Vereye quantity is the pre-registry rate of a CRC-valid unregistered key, at most $C 2^{-24} = 1.1 \cdot 10^{-4}$ per verification, which the figure does not show.

4.5 The false-owner bound

Let a verification check at most C candidate codewords against the CRC; the reader enforces $C \leq 1,771$ (a hard cap: a decode that would exceed it is not run). On an unmarked image, each candidate passes the 24-bit check with probability 2^{-24} , under the assumption that the check field of a candidate decoded from noise is uniformly distributed, and must then also coincide with one of N registered 64-bit keys. By the union bound,

$$P_{\text{wrong owner}} \leq C \cdot 2^{-24} \cdot N \cdot 2^{-64} = C N 2^{-88}. \quad (4)$$

The cap was added to the reader after the benchmark runs of Section 5; its value is the largest number of checks measured in development. The stored results of those runs do not record the number of checks per verification, so for them the value of C rests on that development measurement rather than on the cap. With $N = 10^{11}$ registered images, $P \leq 5.7 \cdot 10^{-13}$ per verification: at one billion verifications per year, one wrong owner in about 1,750 years (Figure 3). Without the CRC assumption, the registry factor alone still bounds the probability by $C N 2^{-64} = 9.6 \cdot 10^{-6}$, because keys are drawn uniformly at random and independently of the image. Before the registry, a verification decodes a CRC-valid but unregistered key with probability at most $C 2^{-24} = 1.1 \cdot 10^{-4}$ under the same assumption; such a key is never reported as an owner, and the verdict states only that no registered mark was found. This is the event comparable with a detector’s false-positive rate. Empirically, no CRC-valid key at all, registered or not, was decoded from 46,467 attacked copies of unmarked images (exact 95 % upper bound $6.4 \cdot 10^{-5}$ per read, treating the copies as independent although they derive from 800 images; this is consistent with, but far weaker than, Eq. (4)). On a marked image a wrong candidate lies close to the true codeword; it passes the check only if its difference from the true codeword is a multiple of the generator polynomial [19, 31], so this case is governed by the minimum distance of the CRC at length 88 rather than by 2^{-24} , and Eq. (4) does not cover it. No wrong key was decoded from 45,469 marked copies (the 45,269 benchmark copies of Section 5 and the 200 edited copies of Section 6).

4.6 Tamper evidence as bit facts

Once the owner is known, the reader regenerates the mark that was written, maps it through the solved geometry, and tests each small cell of the copy. The cell’s mark *reads back*, *does not read back*, or is *unobservable* (saturated, flat or outside the copy). A compact per-tile record in the registry, which holds no image content, lets the comparison run where the record is kept. Changes are reported only where cells fail *coherently*. The reference flip rate is estimated on the copy itself: the observed bits are split into strata by the strength of their projection, and each stratum’s rate is the median over tiles of its rate within the tile, with a floor. An edit covers a minority of the tiles and cannot raise this median. The bits of a cell share its

samples, so counts are divided by a design effect $D \geq 1$ measured on the copy (a robust dispersion of block counts against the binomial model). Let a block R of 3×3 neighbouring cells have n_R observable bits, of which d_R read opposite to the regenerated mark, and let $\bar{p}_R$ be the mean reference rate of those bits. The block is reported as changed when

$$\Pr[\text{Bin}(n_R/D, \bar{p}_R) \geq d_R/D] \leq \alpha/B, \quad (5)$$

where B is the number of blocks tested and $\alpha = 10^{-4}$ (Bonferroni). For independent bits with unequal rates, the binomial at their mean rate bounds the upper tail beyond its mean [35]. Under the binomial model the block test produces a false region on at most 10^{-4} of untouched copies; the model is approximate and the additional rules below are calibrated empirically, so the measured rate is what we report: 0 of 80 untouched copies (95 % upper bound 3.7 %). The additional rules are these: a pair of strongly flipped neighbouring cells seeds a region outside any changed block; large flat-filled areas and erased blocks are reported; and colour and brightness are compared with a coarse colour summary of the registered image. Their thresholds are set above every untouched copy measured during development, and no bound covers them. The report states only these facts and never claims an edit it cannot observe.

4.7 Registry

For each seal, the registry stores a keyed, stretched label of the owner key (never the key itself), a pixel fingerprint, a compact per-tile record for the tamper comparison, and owner metadata. It never stores the image, a thumbnail, or anything from which one can be rebuilt. Each record is signed with ML-DSA-65 (FIPS 204, post-quantum) [20]; records are leaves of an RFC 6962 Merkle log [36], stored encrypted at rest, whose signed tree heads carry both ML-DSA-65 and Ed25519 signatures. Records are therefore designed to remain unforgeable even against a future quantum adversary. Lookups are designed to pass through a separate key service with per-caller rate limits. It computes the label as a keyed hash of an Argon2id stretch of the key (16 MiB, one pass, with a single global salt, because a blind lookup by key cannot use a per-record salt), so that enumeration is rate-limited (one hit per about 43 days at 50 lookups per second and 10^{11} records). An attacker who holds both the store and the key service’s secret can still search offline at about 0.6 core-years per recovered key; recent keys are held in the host’s memory cache. An exact match is decided by the fingerprint alone, but the record is reached only through the key read blindly from the same pixels: the registry is never searched by fingerprint. The fingerprint is stored in clear inside the encrypted log, so whoever compromises the host can test whether a known image is registered. The key is drawn at random and carries no information about the owner; the link between a key and its owner exists only inside the signed record.

5 Evaluation

5.1 Protocols

We reproduce two published protocols. **SynthID-Image** [3]: all 29 of the transformations the protocol names besides identity, in its six groups, each applied twice: at a *random* strength drawn uniformly from the transformation’s range (for example a JPEG quality or a rotation angle anywhere in its interval) and at its *worst* strength, the most severe end of that range (for example the lowest JPEG quality, the strongest blur or the smallest resize). [3] publishes no ranges; ours are stated with the harness. The score is the true-positive rate at 0.1 % FPR, averaged over transformations without weighting. We evaluate at the protocol’s 512×512 on 100 COCO val2017 photos that were never sealed, read or looked at while Vereye was developed, under a decision rule written down before the run; the protocol itself uses 10,000 ImageNet validation images. Two sets at 1024×1024 , which is not the protocol’s resolution, are reported as well: 100 DIV2K training photos, also held out, and the 100 DIV2K validation photos on which the writer and the reader were tuned, a development set and not evidence. **WAM** [4]: native-resolution COCO val (200 images) and the 100 DIV2K validation images (the development set above), under 35 attacks applied with WAM’s own code; WAM’s evaluation also reports inpainting and splicing, which we did not run. Every attack is seeded per (image, attack, strength), so all methods decode identical attacked pixels, and the unmarked copies receive the same

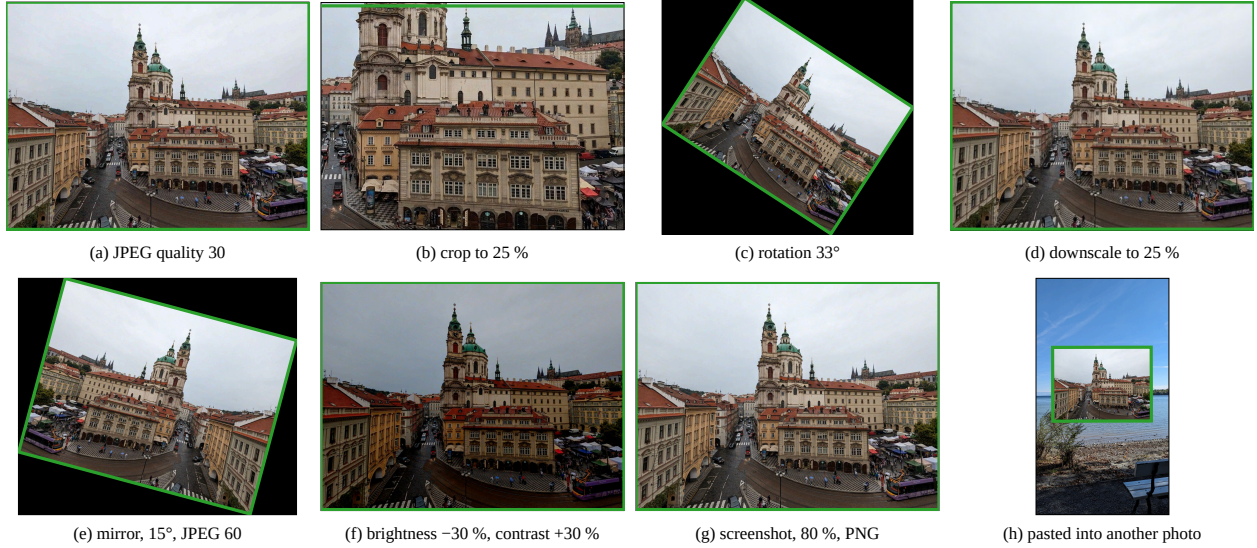


Figure 4. One sealed photograph under eight attacks, each read blind. The green outline is where the reader locates the sealed photograph in the copy, computed from the mark alone. This includes a 25 % crop, a mirrored and rotated copy, a screenshot and the photograph pasted into another picture. In every case the exact owner is identified and no other key is named.

attacks. Competitors run with their official decoders; on the SynthID protocol their watermarks are merged into the full-size image bicubically, as [3] does for every method.

Choices the protocols left open are stated with the harness, which is public with every image, verdict and script. Vereye is scored by exact identification: the 64-bit key read from the attacked copy alone must pass its 24-bit check and be registered. The competitors are scored as [3] scores them: WAM by its own detection score, TrustMark, which has none, by the fraction of its 100 raw bits decoded correctly; each threshold is the 99.9th percentile of that score on the worst-strength unmarked copies, every copy decoded against its own photo’s message, and a detection is a score above it. [3] reports detection (its Table 1) and payload recovery (its Table 2); SynthID-O’s values from both are shown. For the held-out sets the code was frozen before each run (the reader and writer byte-identical to that run’s build, every file hashed: 0a4e72601 for the COCO set, 2c6862cd0 for the DIV2K-training set), and the decision rule was fixed before the first run, its text hashed in our own logs (not timestamped by a third party). Figure 4 shows one sealed photograph under eight attacks.

5.2 SynthID-Image protocol

Table 3 gives the held-out results at the protocol’s resolution. Vereye named the exact registered owner of 99.03 % of the random-strength copies (conservative 95 % bounds 98.52–99.41: the wider of a photo-resampled interval and an exact interval on the clustered effective sample size) and 86.38 % of the worst-strength copies (85.08–87.61); it named no owner on the 6,000 unmarked twins, from which no key passed the check at all. By the rule registered before the run - ahead only if the lower bound exceeds both of SynthID-O’s reported values and neither rival scores higher here than [3] reports for it; behind if the upper bound is below both; otherwise not distinguishable at this sample size - Vereye is behind SynthID-O at random strength and behind it at worst strength. SynthID-O’s values are its authors’, on 10,000 ImageNet images with their own strengths: 99.98 / 99.72 % in their detection table and 99.96–99.97 / 98.84–98.9 % in their payload table. WAM and TrustMark, measured on 50 of the same photos and attacks, detect 92.69 / 81.66 % (WAM, its detector) and 82.62 / 72.55 % (TrustMark, raw bits); recovering their full message, the like-for-like task, they reach 80.41 / 68.14 % and 75.59 / 62.97 %; on exactly those 50 photos Vereye identifies the exact owner in 99.17 / 86.21 %. The rule’s calibration check fails: trustmark bitacc random: 81.10 % here (95 % 78.99–83.09) vs 78.61 % reported; trustmark rawacc random: 82.90 % here (95 % 80.86–84.80) vs 78.61 % reported; trustmark rawacc worst: 72.62 % here (95 % 70.25–74.90) vs 68.73 % reported; wam mask random: 92.69 %

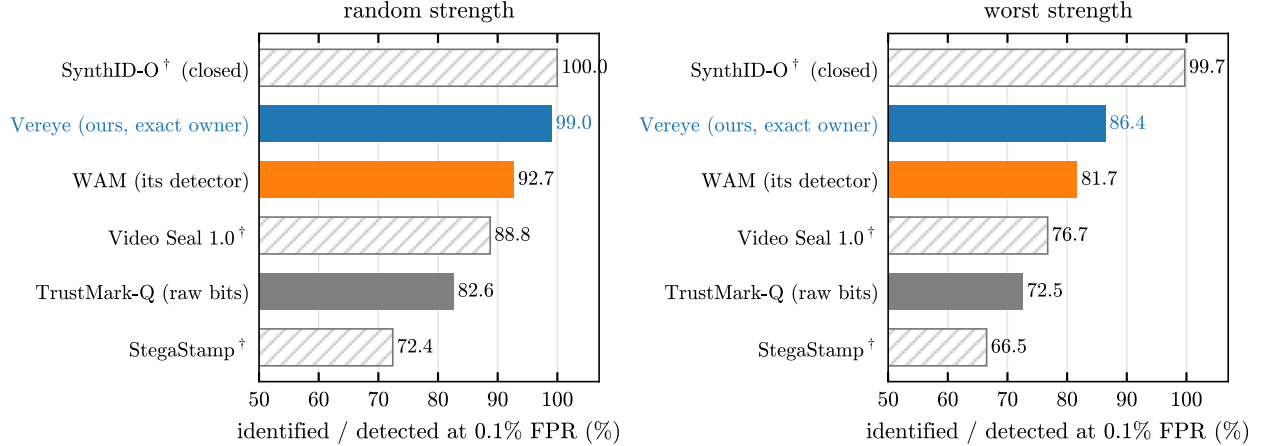


Figure 5. Table 3 as a chart, sorted (the axis starts at 50 %). Hatched bars ([†]) are reported in [3] on its own images and strengths, a different set; the others were measured here on the held-out photos and attacked pixels (the rivals on the first 50 of them). Vereye is scored only on exact identification (the 64-bit key, check-valid and registered); the others on detection at 0.1 % FPR.

here (95 % 91.23-93.98) vs 90.62 % reported. Both rivals score above the values [3] reports for them here, so our implementation of the attacks is probably milder than Google’s for them, which flatters every rate measured here, Vereye’s included. Against SynthID-O’s per-group detection values (Table 1 of [3]) Vereye is lower in 5 of the 12 group and strength cells (random combination: 93.00 % vs 99.96 %; worst color: 99.86 % vs 100.00 %; worst combination: 26.75 % vs 98.06 %; worst noise: 99.75 % vs 99.96 %; worst spatial: 87.50 % vs 99.97 %). Vereye’s losses concentrate in the worst-strength combined attacks (11, 0, 88, 8 % on the four, Figure 6). At 1024×1024 , not the protocol’s resolution, the held-out DIV2K training photos give 99.90 / 98.93 % and the development set 99.93 / 99.48 %; larger images hold more copies of the mark. Over all sets no key was read from any of the 36,000 attacked unmarked copies.

A second held-out run, made after the reader gained the small-image rules (designed after the first held-out run was measured, partly on its photos, which therefore no longer count as unseen for this reader) and not pre-registered, used 300 further COCO val2017 photos at the protocol’s 512×512 , sealed into the production registry and attacked by the same harness: Vereye named the exact registered owner of 99.70 % of the random-strength copies (99.50–99.84) and 95.80 % of the worst-strength copies (95.13–96.43), and read no key from any of its 18,000 unmarked twins. It is reported as a measurement, not a test of the rule above; its worst-strength rate remains below SynthID-O’s reported values. Every verdict, seed and script of these runs is in a benchmark package available on request; the SHA-256 of its file list (every file’s SHA-256, sorted by path) is fd5015c1e9b2ed7a 88aa37df13e94d2b 9bbd32937d4e35da a7b5f3d36ef0e0b5.

5.3 WAM protocol

Our attack set includes the protocol’s extreme settings (perspective 0.8, crop 0.5, JPEG 40, median filter 17), so WAM and TrustMark score slightly lower here than in their paper; Vereye faces the same rows. Figure 7 resolves Table 4 by attack, and Figure 8 places every method on the plane of accuracy against false alarms. Vereye’s zero counts bound its false-alarm rate at $4.3 \cdot 10^{-4}$ on COCO and $8.6 \cdot 10^{-4}$ on DIV2K per copy (exact 95 % upper bounds). WAM’s false alarms are higher than its authors report: on unattacked copies it flagged 5 of 200 COCO images and 1 of 100 DIV2K images, against 0.04 % reported in [4] at the same threshold of 0.07. We have not identified the cause; image pre-processing, resolution handling and the checkpoint are possible sources. The rates in Table 4 are as measured with WAM’s official code and published threshold.

5.4 Invisibility

Table 5 measures every method on the same 200 benchmark images, each marked at its published default and stored as it ships. Vereye’s mark is clearly below WAM’s on every metric. TrustMark-Q is fainter still and identifies less: 76.5 / 60.7 % on the SynthID protocol and 26 % under geometric attacks (Tables 3 and 4).

Table 3. SynthID-Image protocol, every measurement: the first held-out set (100 COCO photos never used in development, 512×512 , 29 transformations, two strengths: 5,800 attacked and 6,000 unmarked copies), the second held-out set (300 further COCO photos, 512×512 , current reader, not pre-registered), and two sets at 1024×1024 , which is not the protocol’s size (held-out and development). Identified (Vereye: the exact registered owner) or detected (the others) at 0.1 % FPR, unweighted mean over transformations. *Interval / exact*: for Vereye the 95 % interval over photos (random / worst); for the others the full message recovered. Reported rows are from [3], on 10,000 ImageNet images with its own strengths, the lower end of each reported interval; [3] also reports InvisMark, TrustMark-P and VideoSeal-0.0, which are not shown.

Method	Random	Worst	Interval / exact	False positives	PSNR	Source
SynthID-O (closed; its authors, ImageNet)	99.98	99.72	payload table 99.96 / 98.84	nominal 0.1 %	—	reported [3]
Vereye (exact owner), first held-out set (100 photos)	99.03	86.38	98.52–99.41 / 85.08–87.61	0 / 6,000	41.6 dB	measured, held out
Vereye, second held-out set (300 photos, current reader)	99.70	95.80	99.50–99.84 / 95.13–96.43	0 / 18,000	—	measured, held out, not pre- registered
Vereye, 100 held-out DIV2K photos at 1024×1024	99.90	98.93	99.61–100.00 / 98.21–99.45	0 / 6,000	41.5 dB	measured, held out; not the proto- col’s size
Vereye, 100 devel- opment DIV2K photos at 1024×1024	99.93	99.48	99.75–100.00 / 98.95–99.79	0 / 6,000	41.4 dB	measured; develop- ment set, not evidence
WAM (its detector)	92.69	81.66	80.41 / 68.14	0.069 %	37.8 dB	measured, 50 photos
Video Seal 1.0	88.75	76.74	—	nominal 0.1 %	—	reported [3]
TrustMark- Q (raw bits)	82.62	72.55	75.59 / 62.97	0.0 %	41.6 dB	measured, 50 photos
StegaStamp	72.41	66.51	—	nominal 0.1 %	—	reported [3]

Because each method runs at a single operating point, these comparisons do not separate robustness from strength; robustness at matched visibility was not measured. Vereye spends its budget in texture, where the eye masks it (Figure 9), and per-pixel metrics count that energy in full. Figure 10 shows the mark on two further photographs.

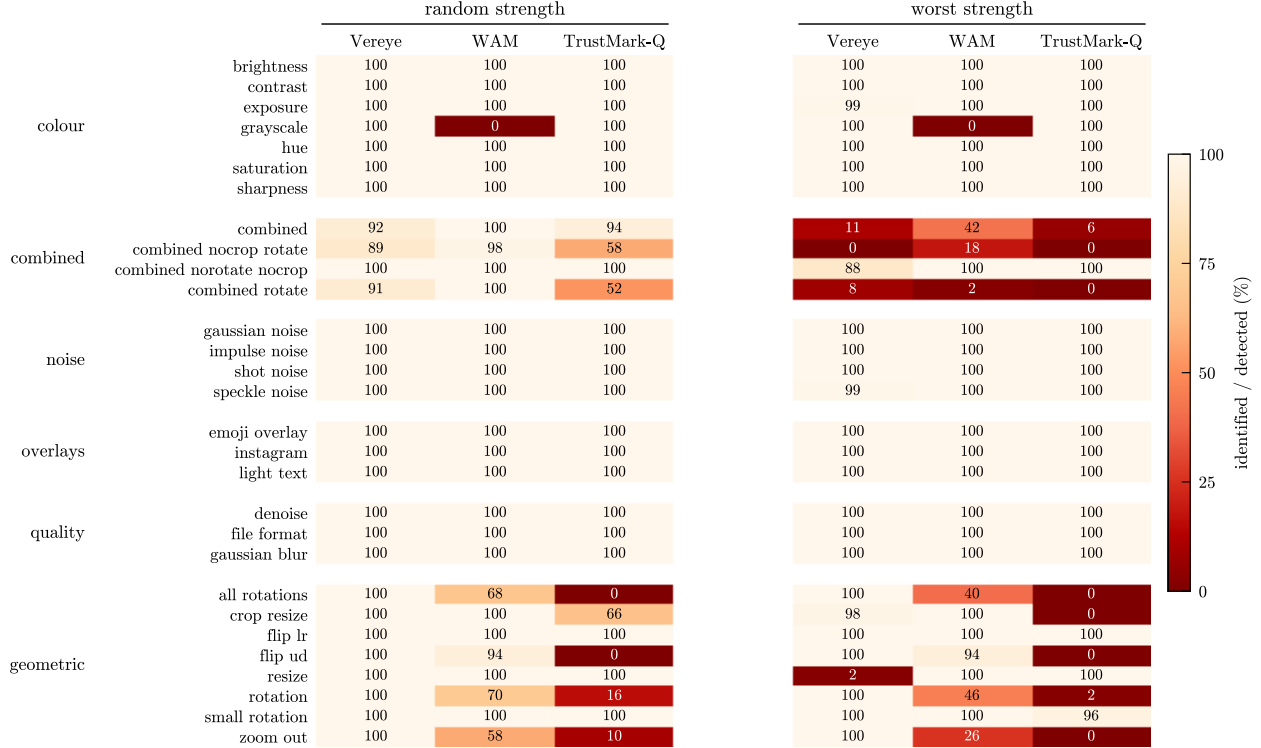


Figure 6. SynthID-Image protocol per transformation on the held-out set, in the paper’s six groups, random and worst strength, for the three methods measured on identical pixels (WAM: its detector; TrustMark-Q: raw bits); white is 100 %, warmer colours mark losses.

Table 4. WAM protocol: 200 COCO and 100 DIV2K images, 35 attacks each. Rates (%) by attack group, each method by its own detection rule (Vereye: the exact registered key; WAM: mask score above 0.07; TrustMark: its presence flag). False alarms are counted on the identically attacked unmarked copies. *Exact*: the full message or key recovered. Reported rows are from [4].

Method	None	Geometric	Valuemetric	False alarms	Exact
Vereye — COCO (reader of 31 October 2026)	100.0	98.3	99.0	0 / 7,000	98.74
WAM — COCO (measured)	100.0	94.8	96.3	162 (2.3 %)	70.27
WAM — COCO (reported [4])	100.0	99.3	100.0	—	—
TrustMark — COCO (measured)	100.0	26.1	91.5	178 (2.5 %)	66.67
TrustMark — COCO (reported [4])	99.8	36.3	90.1	—	—
Vereye — DIV2K (2K, reader of 27 October 2026; 31 of 3,500 copies not run)	100.0	98.9	99.1	0 / 3,467	99.05
WAM — DIV2K (measured)	100.0	87.8	95.3	33 (0.9 %)	71.29
WAM — DIV2K (reported [4])	100.0	96.1	100.0	—	—
TrustMark — DIV2K (measured)	100.0	26.3	91.9	93 (2.7 %)	66.97

6 Tamper localization

Table 6 measures tamper detection on a corpus: 40 camera photographs, each sealed, then edited in five ways at seeded positions and re-saved as JPEG, with WAM marked and edited identically. Vereye finds 197 of the 200 edits, names the correct owner on 199 of the 200 edited copies and no wrong owner on any, and reports no changed area on any of the 80 untouched copies (95 % upper bound 3.7 %). An edit counts as found when a reported region overlaps it, a lenient criterion, so Table 7 also gives the overlap quality. WAM

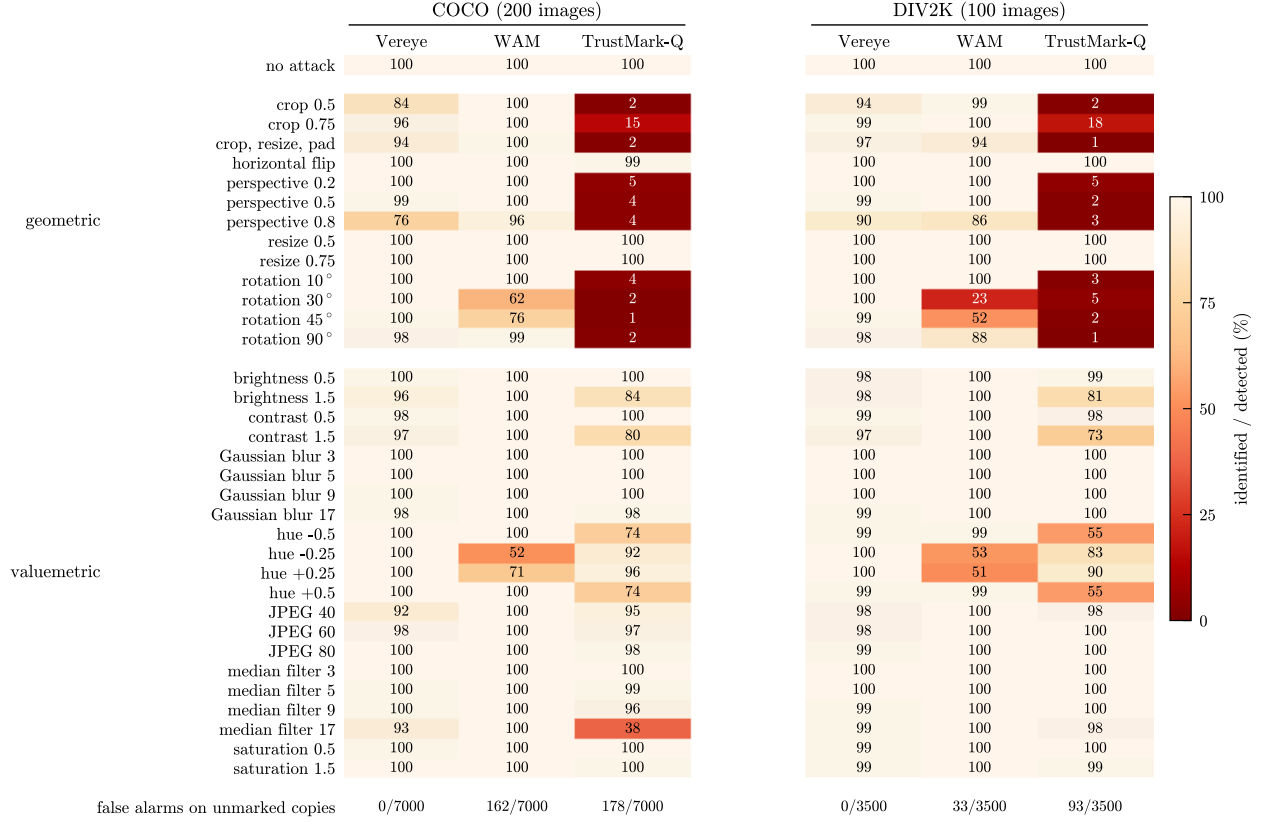


Figure 7. WAM protocol per attack: the share of attacked copies identified (Vereye) or detected (WAM, TrustMark-Q, each by its own rule), on the same colour scale as Figure 6. The last row counts false alarms on the identically attacked unmarked copies.

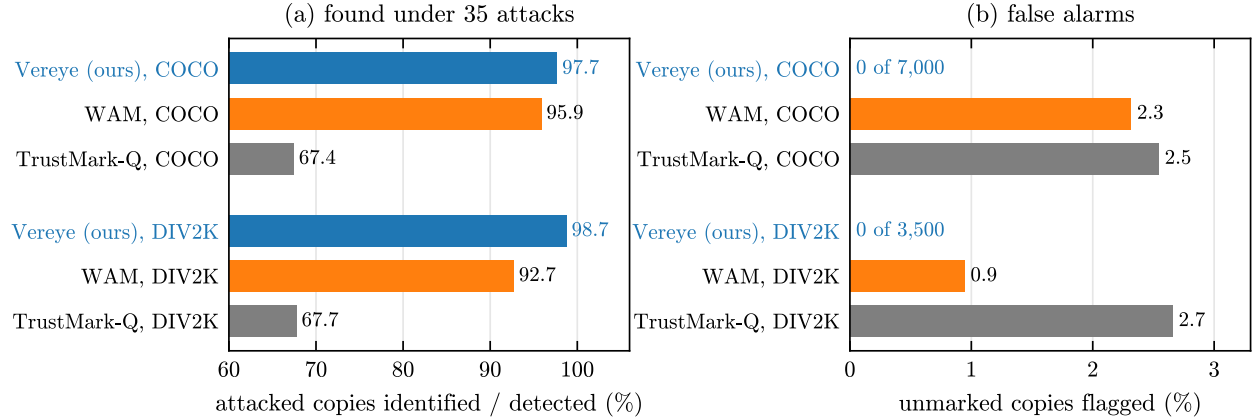


Figure 8. WAM protocol, all 35 attacks pooled. (a) Attacked copies identified (Vereye, exact key) or detected (WAM, TrustMark-Q, each by its own rule). (b) False alarms on the identically attacked unmarked copies: WAM and TrustMark-Q flag 0.9–2.7 %, Vereye none.

misses almost every recolouring, because its mask detects the presence of its watermark, which a colour change leaves intact; on the edits it finds, its masks are tighter (mean IoU 0.74 against 0.48 for Vereye, whose reports follow its cells). Vereye also flags more area outside the edits (2.2 % of the image on average, 14.1 % at most), and its precision on the small 3 % patches is 0.50. Brightness-only edits keep the mark readable and are not reported (Section 8). The set is small: 40 photographs, untouched copies limited to JPEG re-saves and a $0.9\times$ resize, and WAM as the only baseline; EditGuard, OmniGuard [22, 23] and public

Table 5. Invisibility on the SynthID-protocol images (200 COCO images, 512×512), mean over images. LPIPS: AlexNet features [21]; ΔE_{00} : CIEDE2000 colour difference. Lower is better for LPIPS and ΔE_{00} .

Method	PSNR (dB)	SSIM	LPIPS	ΔE_{00}	Stored as
Vereye	41.65	0.990	0.016	1.14	JPEG
WAM	38.41	0.978	0.037	3.11	PNG
TrustMark-Q	41.87	0.991	0.002	0.52	PNG

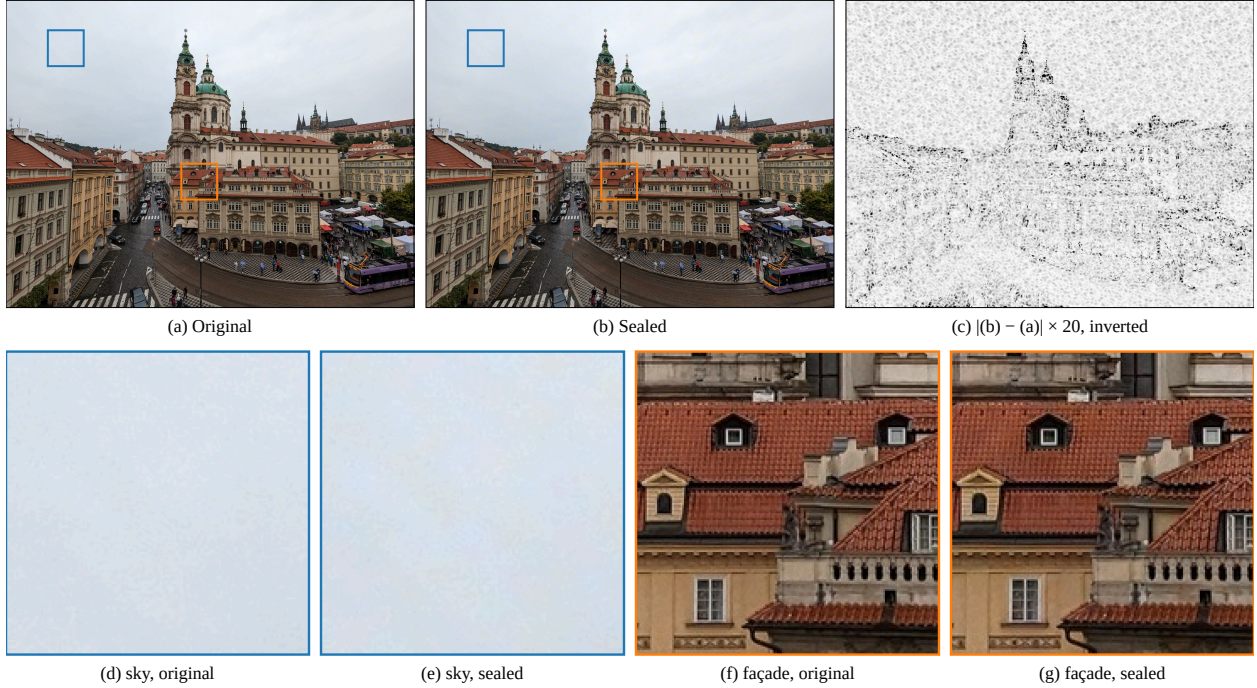


Figure 9. Invisibility. (a) Original, (b) sealed and (c) their absolute difference (mean over the R, G and B channels), amplified twenty times and inverted (white: unchanged). (d)–(g) The boxed 180×180 px regions at full resolution, enlarged $4\times$ without interpolation: (d, e) a smooth sky, original and sealed; (f, g) a detailed façade, original and sealed. The mark sits in texture, where the eye cannot separate it from the photograph; smooth regions carry only colour, at a low level. On this photograph: PSNR 40.8 dB, SSIM 0.994, LPIPS 0.034, mean CIEDE2000 colour difference 1.2. Table 5 compares all methods on the benchmark images.

forensic sets were not evaluated. Figure 11 repeats the verification of Figure 1 after further processing.

7 Where the gains come from

Every failure was compared against a decode at the true geometry and against the host-cancelled signal. This separates losses in the decoder from losses in the channel. In these development runs, the largest gain came from the decoder, not the mark (Table 8).

8 Cost and limitations

Cost. Sealing is one deterministic pass. No GPU and no model weights are needed, so the reader also runs in a browser. The cost of a verification depends on the copy: the number of decodes is bounded by the check budget C but varies from copy to copy.

Limitations. Items (i)–(v) are measured. (i) Worst-strength combined attacks (crop, rotation, darkening and heavy JPEG together): on the held-out set 11, 0, 88, 8% of copies are identified on the four combinations (combined, combined nocrop rotate, combined norotate nocrop, combined rotate). Decoding each missed

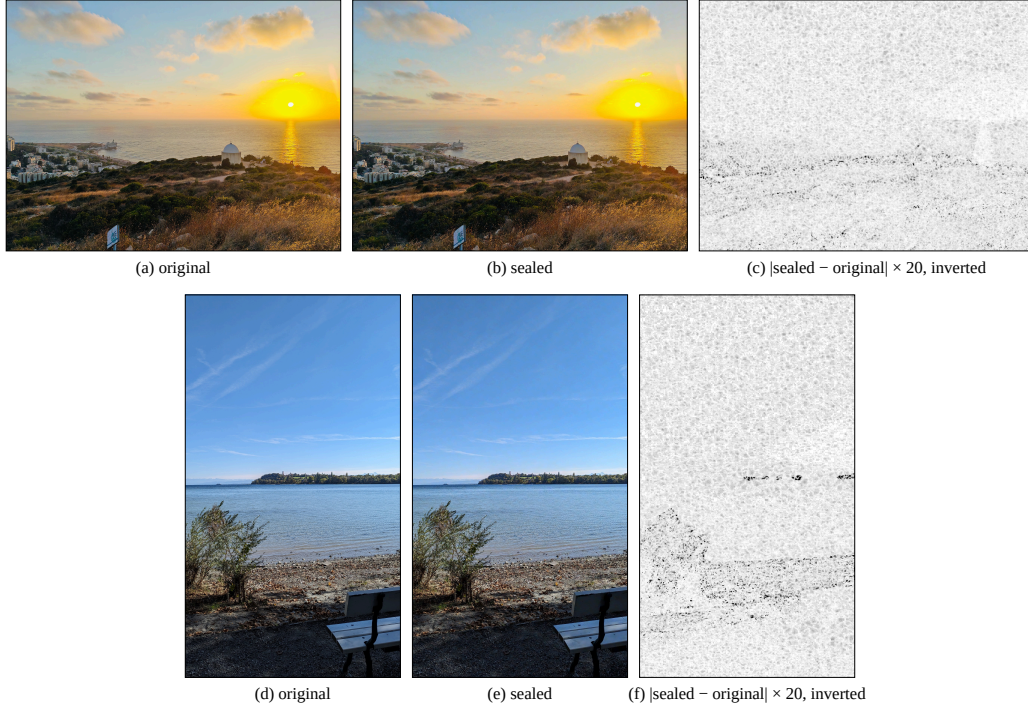


Figure 10. The mark on two further photographs, sealed by the product at the same visibility budget: original, sealed, and their absolute difference (mean over the R, G and B channels) amplified twenty times and inverted. Sealed copies: PSNR 43.6 dB and SSIM 0.993 for (b), PSNR 42.6 dB and SSIM 0.993 for (e). The difference follows texture and edges; flat sky and walls carry almost nothing visible.

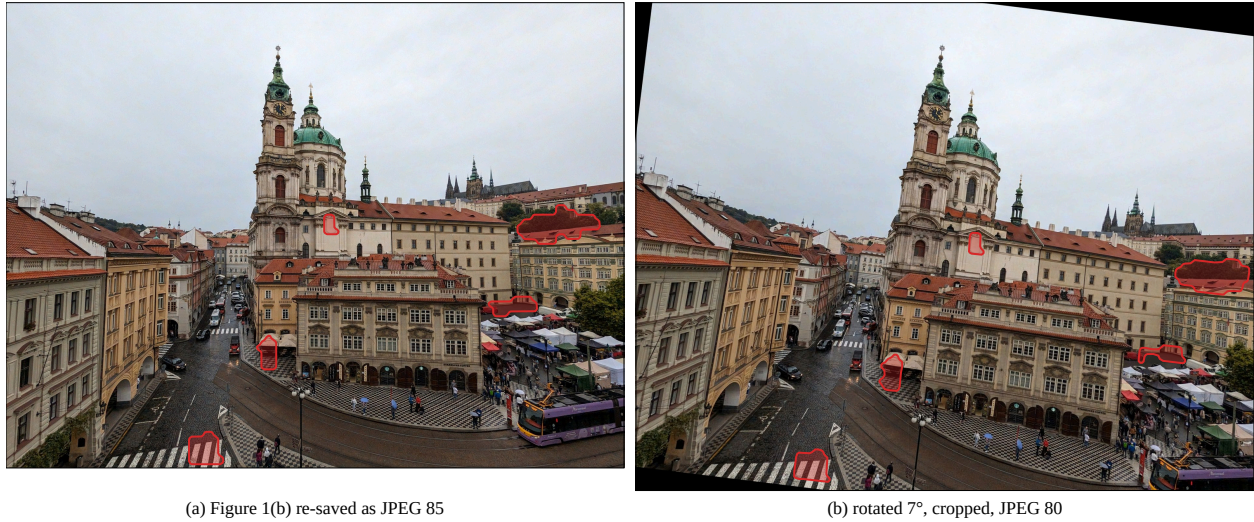


Figure 11. The edited copy of Figure 1(b) after further processing, verified again from the file alone. (a) Re-saved as JPEG at quality 85. (b) Rotated by 7°, cropped and re-saved at JPEG quality 80: the geometry is solved first, then the comparison runs in the copy's own coordinates. In both cases all five edits are reported and nothing is reported more than 48 px away from an edit on this photograph.

copy at its true geometry locates the loss: a 512×512 seal has a quarter of the pixels of a 1024×1024 one at the same per-pixel visibility, darkening and contrast halve the mark and JPEG at quality 50 removes most of its colour; the colour pilot that gives the geometry then sits at the noise level of the pose search (z 6–10 at the true pose), while the payload itself still decodes at the true pose in about three quarters of these misses.

Table 6. Tamper detection from the copy alone: 40 photographs (1024 px), each edit re-saved as JPEG quality 90; measured 21 October 2026. An edit is *found* when the reported regions overlap it. *Owner identified*: the verdict names the sealing key (no wrong key was named on any copy). Untouched copies: re-saved at quality 90, and resized to 0.9× and re-saved at quality 75.

Edit (area of the image)	Vereye: edit found	Vereye: owner identified	WAM: edit found
Patch pasted from another photo (20 %)	39 / 40	39 / 40	40 / 40
Object removed, region inpainted by blur (10 %)	38 / 40	40 / 40	40 / 40
Text banner (10 %)	40 / 40	40 / 40	40 / 40
Region recoloured, hue +60° (15 %)	40 / 40	40 / 40	3 / 40
Small patch pasted (3 %)	40 / 40	40 / 40	40 / 40
All edits	197 / 200	199 / 200	163 / 200
Untouched copies with a reported change (> 1 % of the image)	0 / 80	78 / 80 identified	2 / 80

Table 7. Localization quality on the tamper set of Table 6, computed from the stored per-copy results: intersection over union (IoU), F1, precision and recall of the reported area against the edited area (means), and the share of the image flagged outside the edit (mean / maximum).

Method	IoU	F1	Precision	Recall	Flagged outside the edit
Vereye	0.48	0.61	0.79	0.62	2.2 % / 14.1 %
WAM	0.74	0.77	0.96	0.76	0.22 % / 2.5 %

Redistributing the same visibility budget (pilot bands, a luminance pilot inside the budget) did not help in our measurements; a stronger mark at small sizes did, at a visibility cost. The quarter-size resize, which the held-out run also missed, was a reader defect (the tile grid of a downscaled small seal was never tried); it was fixed after the run, and the held-out numbers are those of the frozen build. (ii) Greyscale conversion removes the chrominance half of the mark; 100 % of greyscale copies of the held-out set are identified. (iii) Heavy re-encoding of already degraded copies destroys the payload itself. (iv) Brightness-only edits keep the mark readable and are not localized; hue rotations and colour inversions of the whole image are localized only in part (local recolouring is found in 40 of 40 cases, Table 6). (v) Non-projective distortions (lens or row-wise warps) are outside the geometric model; on one of three photographs a 6 % row-wise warp of an untouched copy was reported as a changed area. SynthID-O remains ahead on its own protocol; it is closed and deployed only for Google’s own content. (vi) Not evaluated: a held-out image set, robustness at matched visibility, a classical template or self-reference baseline, and the copy and template-removal attacks.

Table 8. Ablation on the SynthID-Image protocol, at the same visibility budget. Rates (%): true-positive rate at 0.1 % FPR for random and worst strength, and pooled over all cases. Each row is a single development run of the 2026-10-20 harness (26 transformations, 200 COCO images at 512 px) on the benchmark images themselves (no held-out set); rows on 60 images are not directly comparable with rows on 200, dashes mark rates that were not recorded, and the reference and intermediate runs are not part of the reproduction kit.

Configuration	Random	Worst	All cases	Images
Reference (19 October 2026)	87.85	78.73	83.29	200
+ decoder stages that no longer discard readable evidence	—	—	89.09	200
+ per-image perceptual budget, revised synchronization	—	—	90.42	60
+ synchronization in the least visible chrominance direction	97.76	85.32	91.54	60
Final	97.54	85.00	91.27	200

9 Conclusion

This paper treats provenance as identification with localization and an analytic error bound, rather than as detection. Vereye names the exact owner of an image from a copy alone and shows what changed without the original. On unmarked images it bounds the chance of naming a wrong owner below one in a trillion per verification, for 10^{11} registered images and at most 1,771 checks, under the stated assumption on the 24-bit check, without machine learning. On our implementation of the SynthID-Image protocol, on held-out photos at its own resolution, it is behind SynthID-O’s reported values at random strength and behind them at worst strength, while being held to a stricter rule; it fails most on worst-strength combined attacks.

Competing interests

The author is the founder of Vereye, which develops the system described in this paper, and the inventor named on its pending patent applications.

References

- [1] J. Zhu, R. Kaplan, J. Johnson, L. Fei-Fei. HiDDeN: Hiding data with deep networks. *ECCV*, 2018.
- [2] T. Bui, S. Agarwal, J. Collomosse. TrustMark: Universal watermarking for arbitrary resolution images. arXiv:2311.18297, 2023.
- [3] S. Goyal et al. SynthID-Image: Image watermarking at internet scale. arXiv:2510.09263, 2025.
- [4] T. Sander, P. Fernandez, A. Durmus, T. Furon, M. Douze. Watermark Anything with localized messages. arXiv:2411.07231, 2024.
- [5] M. Tancik, B. Mildenhall, R. Ng. StegaStamp: Invisible hyperlinks in physical photographs. *CVPR*, 2020.
- [6] K. A. Zhang, L. Xu, A. Cuesta-Infante, K. Veeramachaneni. Robust invisible video watermarking with attention. arXiv:1909.01285, 2019.
- [7] P. Fernandez et al. Video Seal: Open and efficient video watermarking. arXiv:2412.09492, 2024.
- [8] I. J. Cox, J. Kilian, F. T. Leighton, T. Shamoan. Secure spread spectrum watermarking for multimedia. *IEEE Trans. Image Processing* 6(12), 1997.
- [9] B. Chen, G. W. Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Trans. Information Theory* 47(4), 2001.
- [10] S. Pereira, T. Pun. Robust template matching for affine resistant image watermarks. *IEEE Trans. Image Processing* 9(6), 2000.
- [11] I. J. Cox, M. L. Miller, J. A. Bloom, J. Fridrich, T. Kalker. *Digital Watermarking and Steganography*, 2nd ed. Morgan Kaufmann, 2008.
- [12] J. J. K. Ó Ruanaidh, T. Pun. Rotation, scale and translation invariant spread spectrum digital image watermarking. *Signal Processing* 66(3), 1998.
- [13] C.-Y. Lin, S.-F. Chang. A robust image authentication method distinguishing JPEG compression from malicious manipulation. *IEEE Trans. Circuits and Systems for Video Technology* 11(2), 2001.
- [14] J. Fridrich. Security of fragile authentication watermarks with localization. *Proc. SPIE* 4675, 2002.
- [15] A. B. Watson. DCT quantization matrices visually optimized for individual images. *Proc. SPIE* 1913, 1993.
- [16] X. Zhang, B. A. Wandell. A spatial extension of CIELAB for digital color image reproduction. *SID Symposium Digest* 27, 1996.
- [17] Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 50.

- [18] Coalition for Content Provenance and Authenticity. C2PA Technical Specification, version 2.1, 2024.
- [19] W. W. Peterson, D. T. Brown. Cyclic codes for error detection. *Proc. IRE* 49(1), 1961.
- [20] National Institute of Standards and Technology. FIPS 204: Module-lattice-based digital signature standard, 2024.
- [21] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CVPR*, 2018.
- [22] X. Zhang, R. Li, J. Yu, Y. Xu, W. Li, J. Zhang. EditGuard: Versatile image watermarking for tamper localization and copyright protection. *CVPR*, 2024.
- [23] X. Zhang, Z. Tang, Z. Xu, R. Li, Y. Xu, B. Chen, F. Gao, J. Zhang. OmniGuard: Hybrid manipulation localization via augmented versatile deep image watermarking. *CVPR*, 2025.
- [24] M. Kutter. Watermarking resistance to translation, rotation, and scaling. *Proc. SPIE* 3528, 1998.
- [25] M. Kutter, S. Voloshynovskiy, A. Herrigel. The watermark copy attack. *Proc. SPIE* 3971, 2000.
- [26] P. Fernandez, G. Couairon, H. Jégou, M. Douze, T. Furon. The Stable Signature: Rooting watermarks in latent diffusion models. *ICCV*, 2023.
- [27] Y. Wen, J. Kirchenbauer, J. Geiping, T. Goldstein. Tree-Ring watermarks: Fingerprints for diffusion images that are invisible and robust. *NeurIPS*, 2023.
- [28] R. Xu et al. InvisMark: Invisible and robust watermarking for AI-generated image provenance. arXiv:2411.07795, 2024.
- [29] N. Seshadri, C.-E. W. Sundberg. List Viterbi decoding algorithms with applications. *IEEE Trans. Communications* 42(2/3/4), 1994.
- [30] I. Tal, A. Vardy. List decoding of polar codes. *IEEE Trans. Information Theory* 61(5), 2015.
- [31] P. Koopman, T. Chakravarty. Cyclic redundancy code (CRC) polynomial selection for embedded networks. *Proc. IEEE/IFIP DSN*, 2004.
- [32] A. Herrigel, S. Voloshynovskiy, Y. Rytsar. The watermark template attack. *Proc. SPIE* 4314, 2001.
- [33] H. Farid. An overview of perceptual hashing. *Journal of Online Trust and Safety* 1(1), 2021.
- [34] Apple Inc. CSAM Detection: Technical Summary (NeuralHash), 2021.
- [35] W. Hoeffding. On the distribution of the number of successes in independent trials. *Annals of Mathematical Statistics* 27(3), 1956.
- [36] B. Laurie, A. Langley, E. Kasper. Certificate Transparency. RFC 6962, 2013.