Stable Bits, Not Strong Ones: Selecting PRNU Bits for On-Device Cryptographic Camera Provenance

Jamie Newton
Apertrue
jamie@apertrue.photo

***Abstract*—Photo-Response Non-Uniformity (PRNU) is the standard camera fingerprint, but forensics consumes it as a correlation. A cryptographic provenance system consumes a binary code over a pixel subset committed at enrolment and succeeds or fails on its bit-error rate. The forensic subset convention, magnitude selection, is the wrong default here: a training-free mask selected by sign-stability across enrolment frames lowers key-tier bit-error rate from 0.026 to 0.006 on the reference iPhone, shares $\sim 7\,\%$ of its pixels with magnitude, lifting the worst scene frame from 4.7σ to 14.7σ. A neural extractor built to beat it specialises instead: 97.6σ on the enrolling phone falls to 9.9σ on held-out cameras; the untrained classical extractor holds 21.7σ. Scene-enrolled selection, however, admits impostors (7/300 trials): controlled flat-field enrolment is a security requirement. Exact-key recovery from ordinary photos fails as a reliable mode (0.45–0.49), so the system is two-tier: a binding exact-key tier from enrolment flats, and a per-photo threshold tier whose pass line sits 6.8 effective-σ below the impostor mean, with zero false accepts in 308 published-pairing trials (300 at the corpus operating point, 8 at the deployed configuration) and a measured same-model boundary. The extractor is normative: bit-exact across C, Python and Swift.**

I. Introduction

A. The problem: PRNU for a cryptographic consumer, not a correlator

Photo-Response Non-Uniformity (PRNU), the fixed pattern of per-pixel gain variation left by sensor manufacturing, is the standard physical fingerprint for source-camera attribution. Two decades of forensics (Lukáš, Fridrich and Goljan [1] onward) score it the same way: extract a noise residual, correlate it against a reference fingerprint, and threshold the correlation (NCC, PCE, or an AUC/EER on a dataset). Every learned extractor — from SPN-CNN [2] through PhotoNecf [3] and its successors, with denoiser backbones adopted from image restoration [4] and ViT-based models [5] — improves that same correlation score.

A cryptographic provenance system does not consume a correlation. It consumes a *binary code*. To turn a fingerprint into a key (a fuzzy extractor with code-offset and repetition decoding) or to prove membership under a Hamming threshold in zero-knowledge (the companion paper [6]), what matters is the **sign-bit agreement** — the fraction of selected residual bits whose sign matches the enrolled reference — over a *fixed subset of pixels chosen once at enrolment*. Bit-error rate on a committed subset, not correlation over the whole frame, is the quantity that decides whether the system works.

This changes what "a reliable pixel" means, and the change is the point of this paper. The forensics convention selects the pixels with the largest fingerprint **magnitude** ($|K|$), the strongest signal. That is the natural default, and for correlation it is reasonable. But a cryptographic consumer does not care how strong a pixel's signal is; it cares whether that pixel's *sign* comes out the same way every time. A high-magnitude pixel that sits on a scene edge, or that the camera's processing nudges across zero from frame to frame, is worse than useless: it is a confident, repeatable error.

B. Thesis: select for stability, not strength

We select bits by **sign-stability**: for each pixel, the consistency of its residual sign across the enrolment frames, $s_i = |\mathrm{mean}_f \,\mathrm{sign}(\mathrm{residual}_{f,i})|$. This is a fixed, per-device, enrolment-time mask: classical, training-free, and computed in seconds from residuals the device already has. It is also, we find, substantially better than magnitude for the cryptographic objective:

- **It beats magnitude selection.** On our reference iPhone, key-tier bit-error rate falls from 0.026 to 0.006 at the 896-bit key selection ($0.11 \rightarrow 0.006$ over the wider 8 064-bit reliable set), enough to turn exact-key recovery from marginal into reliable. The two criteria agree on only $\sim 7\,\%$ of the pixels they select: magnitude is not a noisy approximation to stability, it selects a *different and worse* population. The improvement is largest exactly where it is needed — the worst single scene frame moves from 4.7σ to 14.7σ above chance, well clear of a bar the standard criterion misses.
- **It beats a neural extractor we built to beat it.** We trained a crypto-aligned network (a RAW-domain NAFNet [7] with a differentiable bit-error loss and a learned per-pixel reliability head, supervised on the exact statistic the circuit verifies). On the phone it was trained on it reached 97.6σ. But that was specialisation: on three held-out cameras the network-selected variant fell to 9.9σ and the raw network residual to 4.0σ, against 21.7σ for the classical wavelet extractor with no training at all. The lesson is not that ML fails, but *where* it helps: it rescues a starved signal in open-world forensics (a single reference

photo); an enrolment system with two dozen controlled RAW frames reaches the ceiling with a classical statistic, and a fixed cryptographic commitment cannot use per-query adaptive selection anyway.

C. What the finding unlocks, and what this paper claims

Stability selection is what makes the rest feasible on a phone. It gives a two-tier design: an exact-key tier from controlled enrolment flats (BER 0.006, recovered by a short repetition code), and — because exact-key recovery from ordinary scene photos is falsified *as a reliable mode* (BER 0.45–0.49 at design time) — a per-photo **threshold** tier, where one image's roughly one million pixel-votes yield a $\tau \cdot N$ ($\tau = 0.475$) decision whose pass line sits **6.8 effective-σ below the impostor mean** (empirically, zero false accepts in 308 cross-camera trials with the closest impostor 5.5σ short of the line; Section VI-C).

Concretely, this paper contributes: (1) the stability-selection finding, with the magnitude comparison and the $\sim$7 %-overlap result; (2) the ML negative and the reframing of when learned extraction helps; (3) the two-tier architecture and the falsification that motivates it; (4) a capture-quality characterisation for a cryptographic consumer (why ProRAW is unusable and the RAW sidecar is load-bearing); (5) a content-masking negative result the forensics literature does not report; and (6) cross-camera validation on fifteen real phones (zero false accepts at the corpus operating point) together with a measured boundary of the criterion itself: scene-enrolled stability selection admits impostors, which makes controlled enrolment a security requirement (Section VI-D). The extractor itself is normative and bit-exact across a C reference, a Python validator and the shipped Swift — the extractor *is* the specification, which is what lets a cryptographic circuit commit to its output.

Everything here is measured on real hardware (an iPhone 13 Pro Max primary, and a fifteen-camera corpus for cross-device claims). The zero-knowledge and privacy machinery that consumes these bits is the subject of the companion paper [6]; this paper is about earning bits worth committing to.

II. Background

A. PRNU as a sensor fingerprint

An image sensor's pixels do not respond identically to light. Manufacturing variation leaves each pixel with a fixed multiplicative gain offset (Photo-Response Non-Uniformity), so a captured image I can be modelled as

$$I = I^0 \cdot (1 + K) + \Theta, \tag{1}$$

where I^0 is the noise-free image, K is the per-pixel PRNU factor (the fingerprint), and Θ collects other noise. K is deterministic, unique per sensor, and stable over the device's life; it is present in every image and cannot be removed without destroying the image. Since Lukáš, Fridrich and Goljan (2006) [1] it has been the dominant tool for source-camera attribution, and Chen et al. (2008) [8] established the maximum-likelihood estimator and the correlation detector still used today.

Because K is multiplicative with intensity, it is estimated from a denoising residual: $W = I - \mathrm{denoise}(I)$ suppresses I^0 and leaves an estimate of $I \cdot K + \Theta$. Averaging residuals over many flat-field frames concentrates K; a single scene frame gives a noisier estimate dominated by whatever content the denoiser failed to remove.

B. Extraction: the wavelet residual

We use a single-level Haar wavelet residual with BayesShrink soft thresholding [9], a classical content-suppressing PRNU estimator. It is deliberately simple and deterministic: the extractor is normative (Section VI-F), so its exact behaviour is part of the specification a cryptographic circuit commits to, not an implementation detail free to drift. It is not bit-exact to a reference `skimage` implementation (they agree on $\sim$83 % of signs, as expected from different boundary and rescaling choices); what matters is that our C reference, Python validator and shipped Swift agree with *each other* bit-for-bit, so the enrolled code is reproducible by construction.

C. The consumer: a binary code, not a correlation

Classical attribution ends at a correlation and a threshold. A cryptographic provenance system does not. It needs one of two things, and both consume a *binary code* rather than a real-valued correlation:

- **A key (exact-recovery tier).** A fuzzy extractor turns a noisy bit string into a stable key: a code-offset sketch plus repetition decoding corrects up to a bounded fraction of bit-flips, so the same key is recovered from any fresh reading whose bit-error rate is under the code's radius. The relevant quantity is BER on a fixed selected subset, and it must be low (a few per cent) for a short code to succeed.
- **A threshold proof (membership tier).** Rather than recover a key, prove that a fresh reading is within Hamming distance $\tau \cdot N$ of an enrolled reference (in the clear, or in zero-knowledge as in the companion paper [6]). Here the relevant quantities are the genuine BER (must sit well below τ) and the impostor BER (must sit at chance, $\sim$0.5), with the separation setting the false-accept rate.

In both cases the enrolled object is a **fixed subset of pixels chosen once at enrolment** plus their reference signs. The system commits to that subset; it cannot, at verification time, look at the fresh image and re-choose which bits to trust. This "fixed at enrolment" constraint is what separates the selection criteria of Section III from the adaptive and learned approaches of Section IV, and it is why the natural forensics choice (rank pixels by fingerprint magnitude) turns out to be the wrong default for this consumer.

D. The magnitude convention

Full-frame weighted correlation is the default detector; when the forensics literature commits to a *subset*, it selects by magnitude. The fingerprint digest retains the k largest

fingerprint values and their positions [10], and binarised representations quantise those values to signs [11]. The field has long distrusted strong *residual* components at test time as scene-contaminated — enhanced SPN attenuates them [12], phase SPN discards magnitude outright [13] — but enrolment-time subset selection has stayed magnitude-ranked, and every learned extractor in the recent literature inherits that framing implicitly, optimising an extractor so that the strong-signal correlation is higher. Section III shows that for a *bit* consumer this criterion selects the wrong pixels, and that a different enrolment-time statistic, sign-stability, is both better and, for a fixed commitment, more principled.

III. Stability selection: the finding

Numbers in this section are measured on the reference iPhone 13 Pro Max and the fifteen-camera corpus (Section VI has the full battery and the extraction pipeline).

A. Definition

Enrolment captures a set of F frames of a bright, even surface (flats) at base ISO ($F = 24$ in the campaigns reported here). For each we compute the wavelet residual W_f (single-level Haar BayesShrink; Section II), the standard content-suppressed PRNU estimate. The per-pixel **stability** is the consistency of the residual's *sign* across those frames:

$$s_i = \left| \frac{1}{F} \sum_f \operatorname{sign}(W_{f,i}) \right| \in [0, 1]. \tag{2}$$

A pixel whose residual sign is the same in every enrolment frame has $s_i = 1$; one that flips like a coin has $s_i \approx 0$. Selection takes the top-N pixels by descending s_i; the deployed enrolment breaks ties by ascending index, while the frozen evaluation artefacts behind Section III-C and Figure 1 keep their original sort order, which we preserve so every printed number remains reproducible (on the perfect-stability plateau the tie rule is material). The enrolled code is the sign of the mean residual at those positions. The selected index set and the code are fixed at enrolment and never recomputed; as Section III-D notes, that is exactly what a cryptographic commitment requires. One deployment guard is required on multi-lens hardware: a position may join the selection only with *photosite evidence* — residual energy above a self-calibrating floor — to exclude ISP edge-fill padding, whose copied pixels are perfectly sign-stable yet device-independent and would otherwise pack the fingerprint with worthless bits (the companion paper measures this failure and the fix [6]).

Contrast the forensics default, **magnitude** selection, which takes the top-N pixels by $|K_i|$, the estimated fingerprint strength. Both are enrolment-time masks over the same residuals; they differ only in the scalar they rank by.

B. Magnitude selects the wrong pixels

The two criteria are not two estimates of the same thing. On the reference device, the top-32 704 pixels chosen by stability and the top-32 704 chosen by magnitude **overlap on only $\sim 7\,\%$ of positions** (against a $0.27\,\%$ chance baseline: correlated, but far from the same set). Magnitude is not a noisy approximation to stability that a larger N would reconcile; it selects a different and, for a bit consumer, worse population.

TABLE I
Separation (z-score above impostor chance under bit-independence) at the 32 704-bit operating point, reference iPhone 13 Pro Max, 24-frame flats enrolment. Flats column: mean of 25 held-out flats; scenes: mean of 33 hand-held photos. Recomputed from the frame record, 2026-07.

Selection	Flats	Scenes	Worst scene frame
Magnitude ($\lvert K\rvert$)	46σ	15σ	4.7σ
Stability	109σ	28σ	14.7σ

The reason is measurable. On the flat-field enrolment behind Table I there are no scene edges to blame; instead, the magnitude estimator's intensity normalisation concentrates its top ranks in the dimmest photosites — median intensity 83 DN at the top magnitude positions, against 559 at the stability positions and 370 frame-wide — where the estimate is inflated while the per-frame sign signal is weakest. On content-bearing enrolment a second face appears: "strong" often means "on a scene edge" or "in a region the sensor's processing pushes around" (Section VI-D measures where that leads). Either way, such a pixel contributes a large term to a correlation and a confident, *repeatable* error to a sign code. Stability measures the property the code actually needs — does this bit come out the same way every time — directly, and a single strong-but-inconsistent pixel is exactly what it rejects. Multi-frame enrolment is what makes this visible: with one frame, magnitude is all you can measure; with two dozen, the pixels that actually hold their sign reveal themselves.

C. Measured: stability wins

We report separation as a z-score (σ above the impostor-chance mean) at the 32 704-bit operating point, for controlled flats and for ordinary hand-held scene photos, and (the quantity a per-photo system lives or dies by) for the *worst single scene frame*. Reference iPhone 13 Pro Max: Table I.

The worst-frame row is the headline. A per-photo provenance system is only as good as its weakest genuine photo, and 4.7σ under magnitude selection sits below what a low false-accept rate needs; stability lifts the same worst frame to 14.7σ — with no extra bits, no training, and no change to the extractor, only a change to which bits are committed.

At the exact-key (identity) tier the effect is a bit-error rate collapse: over the top-8 064 reliable set, key-tier BER falls from **0.110 (magnitude) to 0.006 (stability)** — a $17\times$ reduction (recomputed from the residual cache, 2026-07; BERs here score the enrolled code against a held-out verification aggregate — per-capture rates are higher, Section III-D). On the shipped 896-bit key selection the gap is $0.026 \to 0.006$ (23 vs 5 flipped bits of 896), well inside the $r = 7$ repetition code's radius in the mean-BER sense; on the held-out normative-extractor run, decoding saw zero flipped key bits (0/896; Section VI-B). Figure 1 traces both criteria across the full

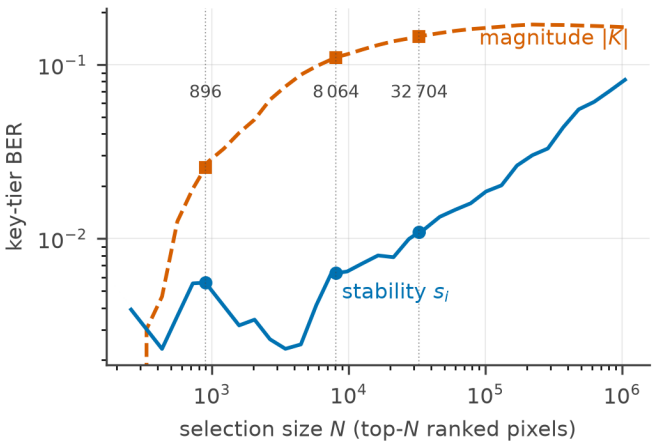

Fig. 1. Key-tier bit-error rate against selection size N for the two ranking criteria, recomputed from the residual cache (enrolled bits vs a held-out verification aggregate). Markers: the operating points reported in the text. The small-N wobble in the stability curve is tie order inside the perfect-stability plateau.

range of selection sizes: the gap is not a property of one operating point but of the whole design space.

D. Why it is the right criterion for this consumer

Two properties make stability selection not just better but *correct* for a cryptographic fingerprint, where magnitude is merely conventional:

- **It is a fixed enrolment-time mask.** The selected indices and the code are computed once and frozen. A zero-knowledge commitment [6] or a fuzzy-extractor sketch binds to exactly this fixed object; it cannot use a per-query, adaptive selection that looks at the fresh photo to decide which bits to trust. Stability is protocol-native in a way an adaptive or learned per-image reliability estimate is not (Section IV).
- **It measures reliability directly, not a proxy.** The consumer's failure mode is a flipped selected bit; stability is the empirical anti-flip probability over enrolment. There is no modelling gap between what is optimised and what is verified — unlike correlation-optimised selection, which improves a score the circuit never sees.

Sign-stability is also the coarse, integer-only form of a per-pixel signal-to-noise statistic, and the finer form is better still: ranking by $|\mathrm{mean}_f W|/\mathrm{std}_f(W)$ over the shipped enrolment's frames gives per-frame key-tier BER 0.029, against stability's 0.047 and magnitude's 0.109 (896 bits, mean over 25 held-out flats; the two reliability rankings share 55 % of selected positions, magnitude under 10 %). The dichotomy that matters is reliability against strength; within the reliability family, the t-refinement is a drop-in improvement for a future enrolment engine at the cost of one accumulator, while the deployed system ships the sign-stability form, whose integer counters suit fixed-point enrolment.

E. Cross-camera: the separation is device-specific, and rejection holds

A finding that only helped the enrolling phone would be suspect, and the criterion's impostor side needs its own evidence. The saved fifteen-camera evaluation (Section VI-C; 300 trials) establishes the true-negative side at the magnitude/wavelet operating point: impostor separation sits at chance (mean 0.0, range -4.8 to $+3.5$) with **zero false accepts**. A per-device *stability* run on the same corpus (Section VI-D) shows why the enrolment regime matters: under the corpus's scene-photo enrolment, stability selection admits impostors that magnitude does not, because sign-stability rewards device-independent structure wherever content-bearing enrolment lets it in. The stability gains reported above are therefore stated as the controlled-flats, on-device result, and Section VI-D makes the flats-and-guard enrolment a security requirement rather than a capture preference. This also distinguishes the criterion from the learned extractor of Section IV, whose device-specific gains did *not* survive transfer.

IV. A Crypto-Aligned Neural Extractor, and Why It Did Not Win

We built the learned extractor the literature's trajectory points at, aimed it at the exact cryptographic statistic, and it lost to a classical one-line criterion. The result is not "ML fails" but an account of *when* learned extraction helps.

A. The gap we set out to fill

Post-PhotoNecf source-attribution work swapped DnCNN for stronger restoration backbones (invertible networks [4] among them) and added blind-noise modules, ViT-based models [5] and Hadamard-product scoring [14], but every one of them optimises a correlation-style identification metric (NCC, PCE, AUC). None optimises what a cryptographic consumer eats: sign-bit agreement over a selected subset, and the per-pixel *reliability* that chooses the subset. That gap is real, and on its face it is exactly where a learned extractor should help.

B. The design: aligned to the circuit's statistic

We trained a RAW-domain extractor (a NAFNet-style backbone [7], Bayer-native with pixel-shuffle CFA handling) against a loss built from the quantities the circuit verifies, not a correlation:

- L_{bit}: a differentiable bit-error surrogate, tanh-relaxed sign agreement between the network residual and the enrolled reference, per pixel.
- L_{rel}: a second head predicts each pixel's bit-flip probability p_i, supervised by the *actual* flips observed across augmented re-captures. This replaces top-$|K|$ selection with a learned reliability map, and p_i also yields soft-decision decoding weights $w_i = \log\big((1-p_i)/p_i\big)$.
- Plus the field's own triplet and frequency losses, for comparability.

We called the object *crypto-aligned fingerprint extraction*: the network is trained against the exact statistic the zero-knowledge circuit consumes. To our knowledge no prior extractor aligns to an in-circuit consumer.

C. What happened: specialisation, not a better extractor

On the phone it was trained on, the learned reliability head produced large gains. Using wavelet bits at the network-selected positions (a composite that keeps the classical extractor but takes the net's *selection*), separation on flats reached 97.6σ at 32k bits, roughly twice the classical wavelet-plus-magnitude baseline, and the scene-weighted variant lifted the worst single scene frame past the 7σ with no extra bits, above the saved-run 6σ acceptance line (Section VI-C). Taken alone, this reads like a win.

It was not. The gain was **specialisation to the enrolling device**. Trained on twelve cameras and evaluated on three held out entirely, the network's separation collapsed: 4.0σ (network residual) and 9.9σ (network-selected reliability) against 21.7σ **for the classical wavelet extractor with no training at all**. The learned extractor is worse than doing nothing on a camera it has not seen. Its on-phone brilliance did not transfer because what it learned was largely that device's specifics, not a universal notion of which pixels are reliable.

D. Reconciliation: when learned extraction helps

The negative is not a contradiction of the ML literature — it delimits it. The open-world forensics setting those extractors target is *data-poor at query time*: a single reference photo, a single questioned photo, and a network's job is to rescue a starved signal it can generalise from a large training corpus. Our setting is the opposite: an *enrolment* system that captures a few dozen controlled RAW frames of the device in front of it (24 in the campaigns reported here). With that much matched data, a classical statistic already reaches the ceiling, and the one that matters, sign-stability across those frames (Section III), is measured directly rather than predicted.

There is also a structural reason a learned per-image reliability map cannot help our consumer even where it is accurate: the cryptographic commitment is fixed at enrolment (Section II-C). A network that decides, at verification time, which bits of the fresh photo to trust is computing an adaptive selection the committed circuit can never use. The reliability that matters must be baked into the enrolled mask, which is exactly what stability selection is, and what a per-query network is not.

We therefore de-prioritised the learned extractor. The training code and the crypto-aligned objective are retained for a future, proper-scale attempt (a larger backbone, a seventy-camera corpus, longer training); the negative here is against a small, undertrained network, not a claim that no learned extractor could ever generalise. But for the system this paper describes, classical wavelet extraction with stability selection is the extractor, and the ML campaign's lasting contribution is the reframing above.

V. Two-tier architecture and single-frame viability

A. The falsification: no exact key from an ordinary photo

The appealing design is one tier: recover the same exact key from any photo the camera takes. We measured whether that is possible, and it is not. Against a controlled flat-field enrolment, the design-time measurement (at the configuration that fixed the operating point) put single-frame scene bit-error rate at **0.45–0.49**, and 0.39 even averaged over a burst of five, against a decode requirement below roughly 0.40 even with a strong ($r = 63$) repetition code. Under the final stability masks the per-scene range is broader (0.325–0.465 over the 33-scene battery), so an occasional near-flat scene dips toward decodability; what the measurements falsify is exact-key recovery as a *reliable* mode — a system cannot depend on the photo content cooperating. (Same-scene bursts do not rescue it: the content is correlated across the burst, so averaging does not wash it away the way independent-scene aggregation optimistically suggests.)

The intuition is that a scene photo's residual is dominated by content the denoiser could not remove, not by the faint multiplicative fingerprint; the fingerprint is there, but at a per-bit signal-to-noise ratio far below what exact decoding needs. This is a property of the physics, not of the extractor: no amount of masking or soft-decision decoding closes it (Section VI-E).

B. The pivot: a million pixel-votes per photo

Exact recovery per bit is the wrong question. The right one is aggregate: across the whole selected set, is the fresh photo's residual closer to the enrolled device than chance would allow? A single ordinary scene frame, compared over the top $\sim$1M enrolment bits, sits about 12σ **below the chance Hamming distance** — an overwhelming aggregate signal, even though no individual bit is reliable. A million voices are pixels, not pictures: each bit is a coin barely biased toward the truth, but a million barely-biased coins decide the question with enormous margin.

This gives the **threshold tier**: prove $\mathrm{popcount}(\mathrm{fresh} \oplus \mathrm{enrolled}) < \tau \cdot N$ (structurally a binarised correlation detector, but over a subset committed at enrolment and provable in zero knowledge, which is the distinction Section II-C requires). At the frozen operating point ($N = 32\,704$, $\tau = 0.475$ — frozen analytically before any corpus measurement; the companion records the provenance, its §3.3), an impostor's Hamming distance under bit-independence is binomial around $N/2$ with a standard deviation of $\sim$90 bits; measured across the corpus, per-photo impostor BER has standard deviation 0.0037, an effective $N \approx 18\,300$ rather than $32\,704$ (the companion's correlation-aware analysis, its §3.3). We therefore state the margin in measured units: the pass line sits **6.8 effective-σ below the impostor mean** (9.05 under the idealised binomial), and we quote no model tail probability. The empirical statement: across 308 real cross-camera attempts (the 15-camera corpus plus the on-device eight-capture second-phone

demonstration, Section VI-C) there were zero false accepts, bounding the false-accept rate below $1.0\,\%$ at $95\,\%$ confidence, and the closest any impostor came was 3.5σ — still 5.5σ short of the line. On the genuine side, all 33 measured scene frames pass the deployed line (false-reject $\leq 8.7\,\%$ at $95\,\%$ confidence for this battery), with worst-case BER 0.460 against the 0.475 threshold (a 490-bit margin, measured at the shipped enrolment's bind selection). One photo suffices; no scale to a literal million bits is needed, because tightening τ on the stable 32k subset already delivers the required margin.

Threat model. These false-accept numbers quantify rejection of *passive* impostors: other real cameras' genuine photos. An active adversary is a different object. PRNU is estimable from a device's published photos (the fingerprint-copy attack [15]), the committed selection mask is device-correlated (Section VI-F), and a crafted RAW file carrying an estimated fingerprint would pass a bare Hamming check. Freshness and capture binding are the companion system's machinery [6]; within this paper's scope, a threshold pass attests that a reading is fingerprint-consistent with the enrolled device, not that it is a fresh capture.

C. The two tiers

- **Exact-key (identity) tier.** Controlled enrolment flats, base ISO, $N = 896$ bits, $r = 7$ repetition code, measured BER 0.006, inside the code's correction radius (zero flipped key bits on the held-out normative run, Section VI-B). Produces a stable per-device key/commitment. Used at enrolment and for identity-mode proofs.
- **Threshold (per-photo) tier** (the companion paper's *bind tier* [6]). Any ordinary capture, verified against its same-shutter RAW sidecar (Section V-D), $N = 32\,704$, $\tau = 0.475$. Produces a per-photo "fingerprint-consistent with the enrolled device" decision (see the threat model above). Used for everyday provenance; a processed photo without the sidecar gets best-effort forensic verification only.

The exact-key tier's 896 bits decode through the $r = 7$ repetition code to a 128-bit value. Its security property is binding, not secrecy: the code-offset helper data exposes six of every seven raw bits, and PRNU itself is estimable from a device's published photos (the threat model of Section V-B), so the recovered value is a stable, device-bound commitment preimage rather than a secret key — which is how the companion system consumes it [6].

The tiers share one extractor and one selection criterion (Section III); they differ only in N, in the presence of a decoding code, and in what capture conditions they demand. Stability selection is what makes the exact-key tier reliable at all and the threshold tier's worst-frame margin comfortable (Section III-C).

D. What a photo is verified against: the RAW sidecar

A per-photo guarantee needs a fresh reading, and the question is whether the *processed* image a user actually shares (a demosaiced, tone-mapped HEIC) carries enough fingerprint to verify. We tested eight processed HEIC scenes against a RAW flat-field enrolment at the stability-selected bits. In this eight-scene sample the outcome splits: about half survive strongly ($+15\sigma$ to $+28.5\sigma$ over the RAW baseline of 19.6σ) and about half are scrubbed down to $+1\sigma$ to $+2.6\sigma$, plausibly because on-device fusion touches different scenes by different amounts. The mean is $+9.6\sigma$, but the variance is the point: on this sample, a processed HEIC is a coin flip.

The consequence for the system is a design requirement, not a defeat: a *per-photo cryptographic guarantee requires the RAW sidecar*: verification is performed against the RAW reading captured at the same shutter, and the HEIC is bound to it in the companion system [6] rather than verified directly. Best-effort verification of an arbitrary processed photo (no sidecar) remains meaningful, since a strong positive is real roughly half the time, but its absence proves nothing, so it is a forensic mode, not the guarantee.

TABLE II
MEASURED KEY-TIER BER ON THE TOP-8 064 MAGNITUDE-SELECTED BITS, BY CAPTURE FORMAT.

Capture	Key-tier BER
ProRAW (demosaiced, 4-channel; G2 channel all-zero)	0.42–0.48
Classic Bayer RAW, ISO base, bright flats (Halide DNG)	0.11

VI. EVALUATION

All measurements use one primary device (an iPhone 13 Pro Max, classic Bayer RAW via a third-party AVFoundation capture app) and, for cross-camera claims, a fifteen-iPhone corpus (PhotoNecf-released RAW [3]).

A. Capture quality: the RAW format is the first lever

The dominant factor in per-bit quality is the capture format, before any selection. Measured key-tier BER on the top-8 064 magnitude-selected bits is given in Table II.

ProRAW is unusable: it is already demosaiced and tone-mapped, so the raw sensor pattern is largely gone. Classic Bayer RAW at base ISO on bright flats is the workable regime, and on iPhone it is only reachable through third-party AVFoundation capture, not the stock camera. On the same top-8 064 reliable set, switching magnitude $\rightarrow$ stability then takes BER $0.11 \rightarrow 0.006$ ($0.026 \rightarrow 0.006$ at the 896-bit key selection; Section III-C).

B. Separation: stability vs magnitude (reference device)

Table I (Section III-C) records the z-score above impostor-chance at $N = 32\,704$, reproduced from Section III-C for the evaluation record.

The normative C extractor (Section VI-F), evaluated on held-out flats under the shipped enrolment's own stability selection (a different mask and code generation from Table I's campaign battery), gives 81σ (flats) and 25.6σ (scenes, worst 12.6σ), with key-tier BER 0/896, better than the independent

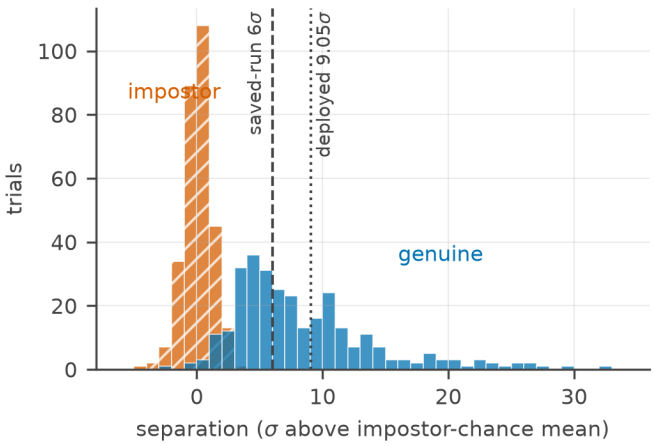

Fig. 2. Fifteen-camera corpus at the 32 704-bit operating point: separation for all 300 genuine and 300 impostor trials. No impostor reaches the saved run's 6σ acceptance line (dashed); the deployed $\tau \cdot N$ pass line (dotted) sits at 9.05σ. The weak genuine tail reflects the corpus's scene-photo enrolment and small crops (see text).

pure-Python reimplementation's 5/896 on the same data (the ctypes validator, which calls the same C, is bit-identical by construction).

C. Cross-camera impostor test (fifteen devices)

Fifteen real iPhones, 30-photo enrolment and 20 test photos each, 512-pixel crops, z-score at 32k (Figure 2). Over 300 impostor trials, the impostor separation mean is **0.0** (range -4.8 to 3.5), i.e. chance, with **zero false accepts** at the saved run's 6σ acceptance line, and *a fortiori* at the deployed $\tau \cdot N$ pass line, which sits higher still (9.05σ; end of this subsection). Genuine separation is 8.3σ mean at the magnitude/wavelet operating point the saved run reports, with a weaker tail (range -2.4 to 32.5) attributable to the corpus enrolling from *scene* photos rather than flats and to small crops, both fixable capture levers. Under these corpus conditions, 110 of 300 genuine trials clear the deployed 9.05σ line — a false-reject statement about scene-photo enrolment at small crops, not about the deployed flats-enrolled system, whose genuine margins are the Section III-C battery. The saved 15-camera run does not include a per-device stability selection, so no cross-camera stability figure is claimed here; the stability gains are the on-device result (Section III-C). The headline is the true-negative side: no impostor is accepted at the measured (magnitude/wavelet) operating point. Put against the frozen operating point, the margin is wide: breaching $\tau \cdot N$ with $\tau = 0.475$ requires an impostor 9.05σ below its own mean, and the luckiest of 300 corpus trials reached 3.5σ — 5.5σ short. The on-device second-phone demonstration (2026-07-10) adds eight more attempts at 49.6–50.4 % Hamming, all chance. The published pairing is not exhaustive, and all-pairs testing on the same corpus (210 ordered pairs, 4 200 per-photo trials; the companion's §5.8) locates the boundary: six threshold crossings, every one from a single same-model pair — the corpus's two iPhone 13 Pro back-telephoto modules, distinct devices that share roughly 40 % of the genuine deviation from chance — with the remaining 208 ordered pairs at zero accepts across 4 160 trials. Same-model non-unique-artefact suppression is the standard forensic remedy and is not part of this pipeline; it is recorded as a required measurement, not assumed.

D. Stability selection under scene enrolment: a measured negative

Whether stability selection is safe on the *impostor* side is a question the deployment guard of Section III-A makes concrete: sign-stability rewards any structure that repeats, including structure that is not the device's. We measured it by repeating the fifteen-camera protocol of Section VI-C with per-device stability selection (30-frame enrolment maps, ties by ascending index, 32 704 bits) in place of magnitude, scoring the same 300 genuine and 300 impostor trials; the magnitude side reproduces Section VI-C's aggregates exactly. The stability side does not stay clean. Seven of the 300 impostor trials cross the saved run's 6σ line (worst 67σ), and the impostor distribution's standard deviation widens to 7.2 in nominal-σ units (range -9.2 to $+67.0$). The failures decompose into two mechanisms: a few individual photos match five or six foreign devices' codes at once while the selection masks overlap only at the 3.1 % chance level — a device-independent sign field, absorbed from content-bearing enrolment, that any mask samples — and one same-model pair (the corpus's two iPhone 13 Pro telephoto modules; see Section VI-C) sits consistently above chance. Genuine separation inflates for the same reason (individual genuine trials reach 99σ), so we quote no cross-camera stability gain from this corpus in either direction.

The consequence is a requirement the deployed system already meets: enrolment must come from controlled flats, where there is no content to absorb, with the photosite-evidence guard excluding what remains. Scene-enrolled stability selection is measurably unsafe; controlled enrolment is what makes the criterion's gains (Section III-C) available without opening the impostor side. The flats-enrolled record (eight second-phone attempts at chance, Section VI-C) is consistent with this, and a flats-enrolled impostor battery at corpus scale is the remaining verification pass (Section VIII).

E. The content-masking negative

A natural hypothesis is that masking content-heavy regions (block filtering by luminance or local variance, or soft-decision weighting) should improve per-bit quality. It does not, in the RAW/wavelet regime: the wavelet residual is already content-suppressed, so the per-bit limit is the faint fingerprint, not residual dirt, and no mask improves sign agreement. The masks are, however, empirically robust to a forgery concern: under every mask we tried, impostor BER stays pinned at 0.500, so an adversary cannot use mask choice to manufacture a match. The result is a negative the correlation-oriented literature does not report: masking is neither necessary nor helpful at the bit level.

F. The extractor is normative and bit-exact

The extractor is a specification, not an implementation, because a cryptographic circuit commits to its exact output. `denoise.c` (single-level Haar BayesShrink) is the single source of truth: a Python validator drives it through ctypes, and the shipped Swift calls the same C, so all three agree bit-for-bit by construction. A golden-vector harness fixes a reference DNG to plane/residual/index/bit checksums and the resulting commitment; on-device tests reproduce them exactly (planes and residuals SHA-identical, key tier 0/896 bit-for-bit). This is what lets the selected code be committed and later re-derived without ambiguity.

The battery above is one campaign against a single enrolment; the enrolment the shipped system later froze is an independent frame set, on which the same selection measures key-tier BER 0.0033 at the 896-bit selection and 0.0263 at $N = 32\,704$ — the stability finding reproduces across two independent enrolment generations. (The companion paper's raw bind-tier 0.026 [6] is that shipped generation at $N = 32\,704$, not this paper's magnitude 0.026 at the 896-bit selection.) The selection also transfers between the generations: the first generation's mask and code, verified against the second generation's reading, flip 0/896, 6/8 064 and 26/32 704 bits, and the reverse direction behaves the same, so the criterion is not overfitting enrolment-session noise. The mask positions themselves overlap only 13–27 % across the two enrolments; the stable object is the code at whichever positions a generation selects, not the position set.

G. Summary of measured claims

- Format matters first: ProRAW BER 0.42–0.48 vs classic Bayer 0.11.
- Stability beats magnitude: key-tier BER $0.11 \to 0.006$ ($17\times$, top-8 064 reliable set; $0.026 \to 0.006$, i.e. 23 vs 5 flipped bits, at the shipped 896); worst scene frame $4.7\sigma \to 14.7\sigma$ ($n = 33$ scenes); $\sim$7 % selected-pixel overlap.
- ML specialises, does not generalise: 97σ on-phone $\to 4\sigma$ cross-camera vs classical 21.7σ (Section IV).
- Two tiers, both viable: exact-key BER 0.006 ($r = 7$); threshold $\tau = 0.475$ with the pass line 6.8 effective-σ below the impostor mean (9.05 under the idealised binomial), empirical FAR below 1.0 % at 95 % confidence from 0/308; worst scene frame 12σ over top-1M bits.
- Rejection holds at the measured operating points: zero false accepts in 308 real cross-camera attempts (300-trial corpus at the magnitude operating point + on-device eight-capture flats-enrolled second-phone demo); closest corpus impostor 3.5σ, 5.5σ short of the deployed $\tau \cdot N$ pass line; all-pairs testing locates the same-model boundary (six crossings, one telephoto pair; Section VI-C).
- Scene-enrolled stability selection is unsafe: 7/300 impostor trials cross 6σ (worst 67σ) on the same corpus; controlled-flats enrolment with the photosite guard is a security requirement (Section VI-D).
- Masking negative: no per-bit gain, mask-forgery-safe (impostor pinned 0.5).
- Extractor bit-exact across C, Python and Swift.

VII. Related work

A. Classical PRNU and the magnitude convention

Source-camera attribution by PRNU begins with Lukáš, Fridrich and Goljan (2006) [1] and the maximum-likelihood estimator and correlation detector of Chen et al. (2008) [8]. Where this lineage commits to a pixel *subset*, it ranks by estimated fingerprint **magnitude**: the fingerprint digest of Goljan et al. [10] keeps the k largest values and their positions, and sign-binarised matching [11] quantises the same magnitude-ranked object — a precedent that also shows sign codes retain identification power. That choice is well-founded for a correlation detector and is the implicit default in essentially all downstream work; Section III shows it is the wrong default for a *bit* consumer, which is our contribution against this baseline: not a better correlator, but a better *selector* for a different consumer.

B. Learned fingerprint extraction

The learned-extraction line begins with SPN-CNN [2]; PhotoNecf (Web Photo Source Identification via Neural Enhanced Camera Fingerprint, WWW 2023) [3] is the closest work to ours and the source of our cross-camera corpus; the post-PhotoNecf line — restoration backbones adopted as extractors (InvDN [4]), ViT-based models (PDC-ViT [5]), blind-noise modules, Hadamard-product scoring (PRNU-Bench [14]) — improves neural fingerprint enhancement. Every one of these optimises a correlation/identification metric (NCC, PCE, AUC/EER). None targets the sign-bit agreement or the per-pixel reliability that a cryptographic consumer needs, and none reports a cross-camera *negative* for a crypto-aligned objective. We built such an extractor (Section IV), aligned it to the in-circuit statistic, and found it specialises rather than generalises, a result this literature does not have, and one that reframes when learned extraction is worth its cost.

We also inherit PhotoNecf's good engineering ideas where they help (Bayer-native processing, pixel-shuffle CFA handling) and reproduce its losses as our network's baseline, so the ablation is like-for-like.

C. The cryptographic consumer

The consumer side is fuzzy extractors (Dodis, Ostrovsky, Reyzin & Smith [16]) — code-offset sketches and repetition/BCH decoding that turn a noisy biometric into a stable key — and, for the membership tier, Hamming-threshold matching proved in zero-knowledge (the subject of the companion paper [6]). Prior ZK-PRNU work (the PhotoNecf sketch, and the membership scheme that registers a hash of the raw fingerprint, a registration the fingerprint-copy attack makes forgeable [15]) is discussed in the companion's §3.6. The closest prior use of PRNU as cryptographic key material is the PUF scheme of Valsesia et al. [17], which binarises random projections of the fingerprint and applies a polar-code fuzzy

extractor; selection there happens in projection space, not per pixel, and no stability statistic is involved. Reliability-based bit selection is itself standard PUF enrolment practice: index-based syndrome coding stores pointers to the most reliable response bits [18], and image-sensor PUFs winnow unstable bits from dark fixed-pattern noise [19]. What is new here is not reliability masking as a mechanism but its object and its evidence: per-pixel sign-stability for PRNU, a signal whose forensic subset convention is magnitude [10], measured head-to-head against that convention and against a trained extractor. The prior ZK-PRNU work, meanwhile, takes the extracted bits as given and says nothing about how to select individual bits worth committing to. That selection is the gap this paper fills.

D. Sensor physics beyond PRNU

Dark-signal / thermal fingerprinting (dark-current non-uniformity, hot-pixel maps) is an adjacent physical channel. The companion paper [6] measures it on the same iPhone platform and finds it closed by on-die suppression (optically-black clamp; synthesised temperature-model dark frames), a negative result (its §4.7, §5.4) load-bearing for its threat model ("no second fingerprint to fall back on") and positioned against the Entropy-2022 smartphone dark-image identification work [20] (which exploits the post-clamp residual FPN, not a thermal signal); the underlying thermal channel (dark current as a recoverable function of sensor temperature [21]) is genuine but suppressed on that hardware. The companion also measures the same channel on two clamp-free consumer DSLRs (its §5.5), where the thermal signal *does* survive but a per-pixel Meyer–Neldel proportionality makes it a fingerprint rather than a possession-hard challenge-response — so PRNU remains the anchor on the physics as well as on the phone.

E. Positioning summary

Against classical PRNU we change the *selection criterion* for a new consumer; against learned extraction we supply the missing crypto-aligned objective and its cross-camera negative; against the cryptographic literature we supply the extraction and selection it assumes. To our knowledge, sign-stability selection for PRNU — and the finding that it beats both magnitude and a trained extractor for a bit consumer — is new for camera fingerprints (its PUF analogue is acknowledged above). A structured search of the classical and learned-extraction literature confirmed that the classical PRNU lineage fixes reliability by magnitude (Chen et al. 2008 [8]) and that the learned-extraction line optimises correlation metrics (PhotoNecf and successors [3], [5], [14]); we found no prior work selecting PRNU bits by sign-stability for a cryptographic consumer, so we state the claim as "to our knowledge" with that basis.

VIII. Conclusion

For a cryptographic consumer of PRNU, the right selection criterion is sign-stability, not fingerprint magnitude. The two criteria choose largely different pixel populations ($\sim 7\,\%$ overlap against a $0.27\,\%$ chance baseline), and the switch (no extra bits, no training, no change to the extractor) turns exact-key recovery from marginal into reliable and lifts the worst genuine scene frame past the margin a per-photo system needs (Section III-C). A neural extractor aligned to the same objective did not close the gap: its on-phone gains were specialisation and did not survive transfer, which delimits where learned extraction helps (rescuing a starved signal in data-poor forensics) from where a classical statistic reaches the ceiling (a data-rich enrolment; Section IV). Because exact-key recovery from ordinary photos is falsified, the system is two-tier: an exact-key tier from controlled enrolment flats and a per-photo threshold tier verified against the same-shutter RAW sidecar, whose rejection side holds at the measured operating points — zero false accepts across 308 cross-camera attempts, including an on-device second-phone demonstration (Section VI-C). The criterion's own impostor side is measured too, and the result is a boundary rather than a blanket: under scene-photo enrolment, stability selection admits impostors (Section VI-D), which makes the controlled-flats, guarded enrolment the deployed system uses a security requirement. The extractor is normative and bit-exact across C, Python and Swift, so the committed code is reproducible by construction, and the selected code transfers across independent enrolment generations (Section VI-F).

One verification pass remains open on the measured record: a flats-enrolled impostor battery beyond the single second device (Section VI-D closes the corpus question against scene enrolment; the controlled-enrolment analogue at scale is future work). The bits this paper earns are consumed — committed and proved against in zero-knowledge — by the companion paper [6].

Artifact availability

The normative extractor (`denoise.c`), its Python validator, the golden vectors, and the aggregate per-trial scalars behind Figures 1 and 2 accompany the paper. The raw captures and all per-pixel artefacts are fingerprint-bearing (a positional map *is* the fingerprint) and are not released.

Generative AI Usage

The author used Claude for tooling implementation and for drafting the manuscript. All results were independently verified by the author against the controls reported herein. The author takes full responsibility for the content.

References

[1] J. Lukáš, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 205–214, Jun. 2006.

[2] M. Kirchner and C. Johnson, "SPN-CNN: Boosting sensor-based source camera attribution with deep learning," in *Proc. IEEE International Workshop on Information Forensics and Security (WIFS)*, 2019, arXiv:2002.02927. Origin of the learned-extraction line.

[3] F. Qian, S. He, H. Huang, H. Ma, X. Zhang, and L. Yang, "Web photo source identification based on neural enhanced camera fingerprint," in *Proceedings of the ACM Web Conference (WWW) 2023*, 2023, DOI 10.1145/3543507.3583225; arXiv:2302.09228. Source of our cross-camera corpus and the ZK-membership sketch; all-correlation-metric baseline.

[4] Y. Liu, Z. Qin, S. Anwar, P. Ji, D. Kim, S. B. Caldwell, and T. Gedeon, "Invertible denoising network: A light solution for real noise removal," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 365–13 374, the invertible denoiser adopted by learned fingerprint-extraction work.

[5] O. Elharrouss, Y. Akbari, N. Almaadeed, S. Al-Maadeed, F. Khelifi, and A. Bouridane, "PDC-ViT: source camera identification using pixel difference convolution and vision transformer," *Neural Computing and Applications*, vol. 37, no. 9, pp. 6933–6949, 2025, DOI 10.1007/s00521-025-11004-z.

[6] J. Newton, "Prove the camera, not the cloud: Anonymous, hardware-anchored photo provenance," Companion manuscript, 2026.

[7] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *Proc. European Conference on Computer Vision (ECCV), Part VII*, 2022, pp. 17–33, DOI 10.1007/978-3-031-20071-7_2. The restoration backbone our crypto-aligned extractor adapts.

[8] M. Chen, J. Fridrich, M. Goljan, and J. Lukáš, "Determining image origin and integrity using sensor noise," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 74–90, Mar. 2008, the ML PRNU estimator and correlation detector.

[9] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Transactions on Image Processing*, vol. 9, no. 9, pp. 1532–1546, Sep. 2000, DOI 10.1109/83.862633. The soft-thresholding rule inside the normative wavelet extractor.

[10] M. Goljan, J. Fridrich, and T. Filler, "Managing a large database of camera fingerprints," in *Proc. SPIE 7541, Media Forensics and Security II*, 2010, p. 754108, DOI 10.1117/12.838378. The fingerprint digest: the magnitude-ranked subset convention.

[11] S. Bayram, H. T. Sencar, and N. Memon, "Efficient sensor fingerprint matching through fingerprint binarization," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 4, pp. 1404–1413, 2012, DOI 10.1109/TIFS.2012.2192272.

[12] C.-T. Li, "Source camera identification using enhanced sensor pattern noise," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 2, pp. 280–287, 2010, DOI 10.1109/TIFS.2010.2046268.

[13] X. Kang, Y. Li, Z. Qu, and J. Huang, "Enhancing source camera identification performance with a camera reference phase sensor pattern noise," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 393–402, 2012, DOI 10.1109/TIFS.2011.2168214.

[14] F.-A. Croitoru, V. Hondru, and R. T. Ionescu, "PRNU-Bench: A novel benchmark and model for PRNU-based camera identification," arXiv:2509.17581, 2025, Hadamard-product scoring; correlation-metric evaluation.

[15] M. Goljan, J. Fridrich, and M. Chen, "Defending against fingerprint-copy attack in sensor-based camera identification," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 1, pp. 227–236, 2011, DOI 10.1109/TIFS.2010.2099220.

[16] Y. Dodis, R. Ostrovsky, L. Reyzin, and A. Smith, "Fuzzy extractors: How to generate strong keys from biometrics and other noisy data," *SIAM Journal on Computing*, vol. 38, no. 1, pp. 97–139, 2008, preliminary version Eurocrypt 2004. Code-offset sketch + repetition decoding = the exact-key tier's consumer.

[17] D. Valsesia, G. Coluccia, T. Bianchi, and E. Magli, "User authentication via PRNU-based physical unclonable functions," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 8, pp. 1941–1956, 2017, DOI 10.1109/TIFS.2017.2697402. PRNU as key material via binarised random projections and a polar-code fuzzy extractor; no per-pixel selection statistic.

[18] M.-D. Yu and S. Devadas, "Secure and robust error correction for physical unclonable functions," *IEEE Design & Test of Computers*, vol. 27, no. 1, pp. 48–65, 2010, index-based syndrome coding: reliability-selected PUF bits.

[19] Y. Cao, L. Zhang, S. S. Zalivaka, C.-H. Chang, and S. Chen, "CMOS image sensor based physical unclonable function for coherent sensor-level authentication," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 62, no. 11, pp. 2629–2640, 2015.

[20] A. Berdich and B. Groza, "Smartphone camera identification from low-mid frequency DCT coefficients of dark images," *Entropy*, vol. 24, no. 8, p. 1158, 2022, DOI 10.3390/e24081158.

[21] R. Matthews, M. Sorell, and N. Falkner, "Determining image sensor temperature using dark current," arXiv:1901.02113, 2019, journal version: Australian Journal of Forensic Sciences 54(5), 2022, DOI 10.1080/00450618.2021.1892186.