ScreenSense: Detecting Screen Re-photos at Verification Time

Mohid Tanveer
University of California San Diego
mtanveer@ucsd.edu

***Abstract*—The content provenance standard C2PA proposes that cameras cryptographically sign every capture, and recent work such as VerITAS shows how to preserve this provenance through a constrained set of image edits using zk-SNARKs [1]. However, both the standard and prior systems largely assume that the signed image is a direct physical capture of the underlying scene. An attacker can instead display arbitrary content, for example an AI-generated or out-of-context photo, on a screen and re-photograph that screen with a C2PA-signing camera, producing a cryptographically valid but misleading image. In this project we investigate *ScreenSense*, a hybrid detector that runs at verification time and attempts to distinguish screen re-photos from authentic photos using only the decoded JPEG and its EXIF metadata. Our pipeline combines: (1) a random-forest EXIF prior over camera metadata, (2) a wavelet + CNN branch targeting moiré artifacts, and (3) a sub-pixel structure branch that analyzes cross-channel frequency patterns, with a small neural fusion head on top. On our collected dataset, the EXIF prior alone achieves AUC ≈ 1.0 and FPR@95 TPR ≈ 0, while the wavelet CNN reaches AUC 0.99 but with higher false positive rates and the sub-pixel branch under performs. We empirically find that, under our current data collection protocol, capture metadata is the dominant cue for re-photos, and we discuss limitations and directions for making signal-level cues more robust.**

1 Introduction

Verifying where and when a digital image was taken has become increasingly important in the face of large-scale misinformation. The C2PA standard [2] proposes that cameras digitally sign each capture along with selected EXIF metadata and possibly device information, enabling downstream consumers to verify that an image in a news article originated from a trusted device and has not been arbitrarily manipulated. Recent work such as VerITAS [1] shows how to preserve this provenance even after common edits (cropping, resizing, blurring) by replacing the camera's signature with a zk-SNARK that proves the edited image was derived from a properly signed original using only whitelisted transformations.

While these systems significantly strengthen provenance, they intentionally focus on transformations inside the camera-to-editor-to-publisher pipeline. A key open gap is that the camera may itself be pointed at untrustworthy content. An adversary can display an AI-generated or out-of-context image on an LCD/OLED display, photograph the screen with a C2PA-compliant camera, and obtain a capture that verifies as genuine with respect to the camera and the signing key, even though the underlying scene is not what it purports to be. This *screen re-photo* attack threatens "glass-to-glass" security, in which the user wants assurance that the signed image corresponds to a live physical scene and not to content replayed from another device.

In this work we explore whether lightweight, verification-time signal analysis can help close this gap. Motivated by forensic work on moiré pattern detection in digital photos [3], and by the observation in VerITAS that focal length and other EXIF features can correlate with suspicious captures, we design and implement a hybrid detector that takes as input a JPEG and its EXIF metadata and outputs a probability that the image is a screen re-photo.

Our contributions are:

- We formalize a threat model in which the verifier only considers C2PA-verified images, so attackers cannot freely tamper with metadata: any modification that breaks the signature will cause the content credentials to appear invalid or untrusted. Within this model, we focus on distinguishing genuine scene captures from re-photos of screens.
- We implement a three-signal hybrid detector: (1) an EXIF prior trained as a random forest over engineered metadata features, (2) a wavelet + CNN branch designed to pick up moiré and other high-frequency artifacts, and (3) a sub-pixel analysis branch that looks for color substructure consistent with LCD/OLED stripe layouts,

fused via a small Torch-based linear head.

- We evaluate this system on a dataset of photos containing both authentic scenes and re-photos of LCD and OLED displays, reporting overall performance as well as leave-one-display-type-out (LOD) and leave-one-camera-out (LOC) metrics. We find that EXIF metadata alone is extremely strong under our collection protocol, while moiré provides a useful but secondary cue and our current sub-pixel design is not yet reliable.

Problem statement. At a high level, we model verification-time re-photo detection as a constrained binary classification problem. The verifier observes an image x (decoded from JPEG) and a metadata vector m (EXIF fields covered by a valid C2PA manifest), and must predict a label $y \in \{0,1\}$ indicating whether x is an authentic scene capture ($y = 0$) or a re-photo of a display ($y = 1$). The constraint is that m is not an arbitrary attacker-chosen feature vector: under our threat model, any post-hoc modification to m that breaks C2PA verification causes the asset to be rejected or downgraded in trust. ScreenSense therefore aims to learn a decision rule $f(x,m)$ that leverages both pixel-level signals and metadata while respecting this provenance constraint, and to characterize how well such a rule can generalize across displays and camera bodies within our dataset.

2 Background and Related Work

2.1 C2PA and Content Provenance

The Coalition for Content Provenance and Authenticity (C2PA) standardizes how devices and software can attach cryptographic attestations, called "content credentials", to media assets [2]. A C2PA-enabled camera signs a manifest describing the original capture, including selected EXIF metadata and possibly device information, using a device-specific signing key anchored in a public key infrastructure. Subsequent editing applications are expected to record edits and new manifests, preserving a chain of signed transformations from capture to publication.

For a verifier, checking provenance involves validating these signatures and ensuring that the manifest chain is intact. If metadata fields that are covered by the manifest signature are altered in an ad-hoc way (for example, by manually editing EXIF tags in a JPEG), the resulting content credentials will fail to verify or will be flagged as untrusted. Our work deliberately operates in this setting: we assume the verifier only accepts images whose C2PA signatures validate, and therefore treats the presented metadata as authentic outputs of the capture pipeline rather than attacker-controlled inputs.

2.2 VerITAS: zk-SNARKs for Image Editing

Datta et al. present VerITAS, a system for "verifying image transformations at scale" using succinct zero-knowledge arguments (zk-SNARKs) [1]. VerITAS addresses a core limitation of the raw C2PA proposal: in practice, images are almost always edited (cropped, resized, blurred) before publication, and a naive signature on the original capture cannot be verified given only the edited image. VerITAS introduces a zk-SNARK proof system in which the editor proves that the published image results from applying only whitelisted transformations to a signed original whose signature verifies under the C2PA public key.

The key innovation is a proof system that supports large witness data: the prover knows a high-resolution original image and a valid signature, and proves that the edited image is a correct transformation without revealing the original. The resulting proofs are succinct and can be verified quickly by end users. However, VerITAS explicitly assumes that the original capture is a faithful recording of the scene. It does not attempt to detect whether the camera was pointed at a physical scene or at an untrusted display showing arbitrary content. Our project can be seen as an orthogonal "liveness" layer that plugs into the verifier side of VerITAS or similar C2PA workflows.

2.3 Moiré Pattern Detection in Digital Photos

Moiré patterns arise when a discrete sampling grid (such as a camera sensor) records another discrete periodic structure (such as an LCD sub-pixel layout). Prior forensic work on moiré detection in digital photos, including "Doing More with Moiré Pattern Detection in Digital Photos" [3], studies how to detect these structured aliasing artifacts as a way to infer that an image contains a display. These methods often analyze frequency-domain statistics, looking for peaks and regularities aligned with expected screen pixel pitches and stripe layouts, and sometimes feed these features into classifiers.

Our design borrows this intuition for our moiré detection model, in which we build a wavelet + CNN branch that takes as input wavelet-decomposed grayscale tensors and spatial RGB crops, allowing the network to learn non-linear combinations of high-frequency patterns, including moiré-like structures. Unlike prior work, we combine these signal-level cues with a strong EXIF-based prior inside a single hybrid pipeline.

3 Approach and Methodology

3.1 Threat Model

Our goal is to detect re-photos of screens at verification time, assuming that the verifier has access only to the decoded JPEG and its associated content credentials. We explicitly

restrict attention to images whose C2PA signatures validate: if the manifest chain is invalid or untrusted, we treat the image as out of scope or already suspicious.

Within this setting, we consider adversaries who:

- Control the content shown on a display (for example, they can render AI-generated images or replay older photos on an LCD or OLED screen),
- Can position a C2PA-signing camera arbitrarily with respect to that display (choosing angle, distance, and focus),
- And then submit the resulting signed capture as supposed evidence of a real-world scene.

Crucially, in our threat model attackers *cannot* freely alter EXIF metadata or other fields covered by the C2PA manifest signatures without detection. Signing binds the metadata to the image content: if an attacker edits EXIF tags after capture using commodity tools, the C2PA signature over the manifest will no longer verify, and a compliant verifier will either reject the asset or show its content credentials as invalid or untrusted. Similarly, if an attacker attempts to forge a new manifest and signature without access to a trusted signing key, they will fail the usual digital-signature unforgeability guarantees.

This assumption lets us treat EXIF fields such as focal length, aperture, exposure time, ISO, and device model as trustworthy observations of the capture process. The attacker can still indirectly influence these fields by choosing how they take the re-photo (for example, shooting in a dark room at a particular distance from the screen), but they cannot arbitrarily rewrite them post hoc. Our detector is therefore designed to flag patterns *induced* by typical re-photo workflows rather than defending against metadata spoofing.

In addition, we envision this verifier as a complement to human inspection: in many re-photos the moiré patterns, sub-pixel cues, and even individual pixels may be highly visible, making automated detection unnecessary. For this reason, we constructed our dataset to emphasize cases in which these cues are subtle and not immediately obvious to a human observer.

3.2 Data and Indexing Pipeline

Our dataset consists of physical captures partitioned into authentic photos and re-photos of displays. Re-photos include both AI-generated content shown on screens (AI-LCD, AI-OLED) and authentic images re-displayed for capture, while authentic images are ordinary scenes with no display present. All images are stored under a common project root, with EXIF and label metadata recorded in CSV files (for training, `data/exif_metadata.csv`; for testing, `data/test/test_exif_metadata.csv`).

The pipeline starts by building an index over images using a small data module. This index resolves absolute file paths, assigns a stable `image_id` per file, and constructs several derived attributes:

- **label**: a binary label where 0 indicates an authentic photo and 1 indicates a re-photo of a screen.
- **screen_group**: a combined descriptor of screen type and source (for example, LCD-vs-OLED and AI-vs-authentic).
- **camera_body**: a concatenation of device make and model (e.g. `Apple iPhone 17 Pro`).

This index is cached as an artifact and serves as the backbone for all downstream feature extraction and evaluation. Using it ensures that all feature tables and tensors remain aligned row-by-row.

3.3 EXIF Features

The EXIF prior is built on a dense feature table engineered from camera metadata. We parse EXIF fields into:

- **Numeric features**: image width and height, focal length in 35mm-equivalent units, f-number (aperture), exposure time and shutter speed value, and ISO.
- **Categorical features**: device make and model, screen type and source (for example, AI-LCD vs authentic-OLED), and metering mode.

We further derive engineered features that better capture exposure and geometry:

- **Aspect ratio** = width / height.
- **Log ISO** = $\log(1+\text{ISO})$, which compresses the dynamic range of ISO values.
- **Exposure value**, a brightness-related term combining aperture and exposure time.
- **Focal-per-aperture**, the focal length normalized by the aperture.

Categorical fields are one-hot encoded and concatenated with normalized numeric features to form a fixed-length EXIF vector per image. These vectors feed a random-forest classifier, which serves as a low-cost but empirically very strong prior about how real capture settings differ between authentic scenes and typical re-photo setups.

3.4 Moiré and Sub-pixel Features

To capture signal-level cues of screens, we extract two families of features from the image pixels.

Wavelet + spatial tensors. For each indexed image we load a grayscale version, resize it to a fixed spatial resolution, and compute a two-level 2D discrete wavelet transform (using the db2 wavelet). We stack approximation and detail coefficients into a multi-channel "wavelet tensor." In parallel, we resize the RGB image to a fixed size and arrange it as a `(channels, height, width)` tensor. These tensors are stored in a compressed feature artifact and used as inputs to a dual-branch CNN that specializes in moiré and other high-frequency artifacts.

Sub-pixel statistics. The sub-pixel branch focuses on cross-channel periodicity consistent with LCD/OLED sub-pixel layouts. For each image we extract overlapping RGB tiles using a sliding window with configured tile size, stride, and maximum tiles per image. For each tile:

- We apply a 2D Hann window to each color channel and compute 2D FFTs of R, G, and B.
- We compute cross-power spectra between channel pairs (R–G, G–B, R–B) and analyze them along a radial frequency grid.
- We estimate peak ratios (peak cross-power over mean), dominant radial periods, and a period-consistency score across channel pairs, combined with a Sobel-based edge-strength estimate.

These tile-level metrics are summarized into global image features: means, standard deviations, maxima, percentiles, top-k statistics, and fractions of tiles with strong peaks or consistent periods. The resulting feature vector is intended to fire when the image contains regular color stripe patterns characteristic of display sub-pixels.

3.5 Models and Fusion

Our hybrid model stack consists of three base learners and a fusion head:

- **EXIF Prior.** We train a `RandomForestClassifier` over the engineered EXIF feature table, with configuration controlling tree count, depth, and regularization. This model outputs a probability that each image is a re-photo based solely on metadata.
- **Sub-pixel head.** We feed sub-pixel features into a scikit-learn pipeline combining `StandardScaler` and `LogisticRegression`. The logistic head learns a linear decision boundary over the high-dimensional sub-pixel statistics, producing a re-photo probability.
- **Moiré wavelet CNN.** The CNN comprises two convolutional branches: one over the wavelet tensor and one over RGB spatial content. Each branch consists of stacked conv–batchnorm–ReLU–max-pooling blocks, followed by global average pooling. The branch outputs are concatenated and passed to a small MLP classifier. We train this network with `BCEWithLogitsLoss` and Adam, using early stopping based on validation loss and automatically selecting CPU, CUDA, or MPS depending on availability.
- **Hybrid fusion head.** Finally, we train a single-layer linear head over the three base probabilities $(p_{\text{moire}}, p_{\text{sub-pixel}}, p_{\text{exif}})$. Implemented in Torch, this head is equivalent to logistic regression but integrates cleanly with our existing training infrastructure. It outputs a fused logit that we pass through a sigmoid to obtain the final hybrid probability.

3.6 Training and Evaluation Protocol

We implement the end-to-end pipeline as a reusable module. Training proceeds as follows:

1. We materialize a *feature store* that contains the index table, EXIF feature table, sub-pixel feature table, and wavelet/spatial tensors, with explicit lists of feature columns per signal.
2. For each base learner we perform stratified k-fold cross validation on the training data, fitting the model and recording out-of-fold (OOF) probabilities for every image. For the wavelet CNN we train a full model per fold; for the EXIF and sub-pixel models we reuse standard scikit-learn cross-validation.
3. We stack the three OOF probability columns into a 3-dimensional representation and train the hybrid fusion head on this stacked matrix, reserving a held-out validation split for early stopping.
4. We then refit final versions of each base model on the full training data and store them, along with the fusion head and metadata about feature groups and tiling parameters, in an artifact bundle.

We evaluate the system along three axes:

- **Overall test-set performance**, using a held-out test CSV and corresponding image folders.
- **Leave-one-display-type-out (LOD)**, where we hold out all re-photos from a given screen type (LCD or OLED) and a balanced subset of authentic images, retrain on the remaining data, and evaluate on the held-out group.
- **Leave-one-camera-out (LOC)**, where we similarly hold out all images from a given camera body (for example, iPhone 12 Pro Max) and evaluate transfer to that device.

For each setting we report area under the ROC curve (AUC) and the false positive rate at 95% true positive rate (FPR@95 TPR) for each signal and for the hybrid fusion. AUC summarizes how well a model ranks re-photos above authentic images across all possible thresholds, which is helpful for comparing signals with different score scales. In contrast, FPR@95 TPR focuses on the high-sensitivity operating regime that matters for security: we would like to catch almost all re-photos (high TPR) while keeping the fraction of authentic images that are incorrectly flagged (FPR) as low as possible.

4 Results

4.1 Overall Test-set Metrics

Table 1 summarizes performance on the held-out test set. The EXIF prior and hybrid fusion achieve perfect AUC and zero false positives at 95% TPR, while the moiré CNN is strong but imperfect and the sub-pixel branch lags.

Signal	AUC	FPR@95 TPR
Moiré	0.9855	0.15
Sub-pixel	0.8421	0.50
EXIF	1.0000	0.00
Hybrid	1.0000	0.00

Table 1: Test-set performance for each signal and the hybrid fusion.

The wavelet CNN's AUC of about 0.99 indicates that moiré-related cues allow reasonably good ranking of re-photos above authentic images, but the non-zero FPR@95 TPR shows that at strict operating points the model still misclassifies a non-trivial fraction of authentic scenes as re-photos. The sub-pixel head exhibits both lower AUC and very high false positive rates, suggesting that its features respond to generic high-frequency texture rather than screen-specific structure. In contrast, the EXIF prior nearly perfectly separates the two classes, and the hybrid fusion essentially learns to trust EXIF almost exclusively.

4.2 Leave-one-display-type-out

Table 2 shows LOD performance when holding out LCD and OLED screen types in turn. Again, EXIF and the hybrid remain near-perfect, while moiré and sub-pixel generalize only partially.

The moiré CNN maintains AUC in the 0.82–0.87 range but suffers high false positive rates, especially when holding out LCD displays. The sub-pixel head performs worse, with AUC values in the low 0.7s and FPR@95 TPR often at or near 1.0. EXIF generalizes almost perfectly across display types within this dataset, and the hybrid fusion closely tracks EXIF.

Held-out	Signal	AUC	FPR@95 TPR
LCD	Moiré	0.8249	0.67
LCD	Sub-pixel	0.7191	1.00
LCD	EXIF	1.0000	0.00
LCD	Hybrid	0.9988	0.02
OLED	Moiré	0.8726	0.42
OLED	Sub-pixel	0.7684	0.75
OLED	EXIF	1.0000	0.00
OLED	Hybrid	1.0000	0.00

Table 2: Leave-one-display-type-out results by signal.

4.3 Leave-one-camera-out

Table 3 reports LOC results when holding out each iPhone model in turn.

Held-out camera	Signal	AUC	FPR@95 TPR
iPhone 12 Pro Max	Moiré	0.7541	0.51
iPhone 12 Pro Max	Sub-pixel	0.6575	0.99
iPhone 12 Pro Max	EXIF	1.0000	0.00
iPhone 12 Pro Max	Hybrid	1.0000	0.00
iPhone 14 Pro Max	Moiré	0.9078	0.50
iPhone 14 Pro Max	Sub-pixel	0.7022	1.00
iPhone 14 Pro Max	EXIF	1.0000	0.00
iPhone 14 Pro Max	Hybrid	1.0000	0.00
iPhone 17 Pro	Moiré	0.8560	0.29
iPhone 17 Pro	Sub-pixel	0.8966	0.46
iPhone 17 Pro	EXIF	1.0000	0.00
iPhone 17 Pro	Hybrid	1.0000	0.00

Table 3: Leave-one-camera-out results by signal.

Here the moiré CNN's AUC varies from about 0.75 to 0.91 with moderately high false positive rates, while sub-pixel AUCs span 0.66–0.90 but often with FPR@95 TPR near 1.0. The iPhone 17 Pro split is a mild exception, where both moiré and sub-pixel FPRs are noticeably lower than for the other two cameras; this could reflect that, for this newer sensor and optics, our capture protocol happened to produce clearer screen artifacts and less confusing texture in authentic scenes. EXIF remains effectively perfect across camera bodies, and the hybrid again mirrors EXIF performance.

5 Analysis and Discussion

5.1 Interpreting EXIF's Dominance

The most striking outcome is the strength of the EXIF prior. Across the test set, LOD, and LOC evaluations, EXIF alone achieves AUC essentially equal to 1.0 with zero false positives at 95% TPR. This suggests that, within our dataset, camera metadata strongly encodes whether a photo is a re-photo.

There are at least two plausible readings. On the positive

side, in realistic workflows re-photos may indeed exhibit distinctive EXIF profiles: they are often taken at closer focus distances, with particular aperture and shutter combinations (especially indoors), and with characteristic ISO and metering choices that differ from varied authentic scenes. If these patterns hold broadly, EXIF-based priors could be a powerful and inexpensive tool for re-photo detection.

On the other hand, our data collection followed a relatively consistent re-photo protocol: we used a small set of devices and capture styles when photographing screens. This consistency makes it easy for a random forest to latch onto dataset-specific quirks such as particular focal-length and ISO combinations that happen to correlate with re-photos in this environment. Because the LOD and LOC splits still share the same overall acquisition protocol, EXIF appears to generalize, but this may overestimate robustness in the wild.

5.2 Limitations and Attacker Adaptation

Our current results should be interpreted with care. We explicitly do not consider attackers who can compromise a trusted signing key or fully control the C2PA capture pipeline; such adversaries could produce arbitrarily forged provenance and fall outside our threat model. Even within our model, a motivated attacker could attempt to mimic EXIF statistics of benign photos by carefully tuning capture distance, exposure, and lighting. Whether such mimicry can be made both effective against ScreenSense and visually convincing to human viewers remains an open empirical question, and our current dataset is not rich enough to answer it.

5.3 Moiré and Sub-pixel Cues

The moiré wavelet CNN achieves reasonably high AUC values but less impressive false positive rates, especially under cross-domain splits. This aligns with the intuition that moiré is a useful but fragile cue: it depends sensitively on display pixel pitch, sub-pixel layout, camera distance and angle, focus, motion, and the underlying image content. For many re-photos, especially those taken with care to avoid obvious artifacts, visible moiré is weak or absent, limiting this signal's standalone reliability.

The sub-pixel branch performs worst, with moderate AUC but very high false positive rates. From the feature design, a likely explanation is that our sub-pixel statistics respond strongly to generic high-frequency structure (for example, fabric textures, foliage, building facades) that also appear in authentic scenes. Modern camera pipelines, including demosaicing, denoising, sharpening, and downscaling, can smear crisp display sub-pixel grids into more generic texture that is hard to distinguish from natural patterns. As a result, the current sub-pixel model behaves more like a generic "high-frequency texture detector" than a screen-specific indicator.

5.4 Differences from Prior Work

Our system sits alongside, rather than replacing, C2PA and VerITAS. Unlike VerITAS, we do not attempt to prove anything about the relationship between a published edited image and a signed original; we assume that this provenance pipeline is already in place and that the verifier only accepts C2PA-validated assets. Instead, we add a liveness-like check on top: given a C2PA-verified image, is it likely to be a re-photo of a display rather than a direct capture of the scene?

Compared to prior moiré-detection work [3], we adopt a more hybrid approach. We still exploit moiré cues, but we fuse them with an EXIF prior and evaluate generalization across displays and cameras. Our results indicate that, under our current conditions, metadata is actually the dominant signal, with moiré providing secondary help and sub-pixel features underperforming. This stands in mild contrast to some earlier forensic work that emphasized pixel-level cues as primary; it suggests that in modern camera/display ecosystems and under adversarial capture strategies, metadata-based reasoning may play a larger role than previously appreciated.

5.5 Future Steps

The current prototype leaves several open questions and opportunities:

- **Stress-testing EXIF generalization.** The most urgent next step is to collect additional data in which re-photos are taken with more diverse operators, devices, focal lengths, and exposure settings, and where authentic images share similar EXIF profiles. Rerunning our pipeline on such data would clarify whether EXIF remains as strong as observed here or whether its performance collapses once protocol-specific shortcuts are removed.
- **Revisiting sub-pixel modeling.** Instead of relying solely on handcrafted cross-power summaries, we could train a small CNN directly on high-resolution crops around edges or suspected screen regions, allowing the model to learn more discriminative representations of sub-pixel structure (if present).

Overall, our results suggest that hybrid re-photo detection at verification time is feasible and that metadata can be extremely informative in controlled settings. At the same time, they highlight the need for more diverse data and improved signal-level modeling if we want detectors that remain reliable across capture styles and adversaries.

References

[1] T. Datta, B. Chen, and D. Boneh. VerITAS: Verifying Image Transformations at Scale. In *IEEE Symposium on Security and Privacy*, pp. 4606-4623, 2025.

[2] Coalition for Content Provenance and Authenticity (C2PA). C2PA technical specification. Online specification, 2024.

[3] C. Yang, Z. Yang, Y. Ke, T. Chen, M. Grzegorzek and J. See. Doing More With Moiré Pattern Detection in Digital Photos. In *IEEE Transactions on Image Processing*, vol. 32, pp. 694-708, 2023.