STEAL THE KNOWLEDGE, INHERIT THE MARK: TWINMARK FOR DISTILLATION WATERMARKING VIA SECOND MOMENTS AND CLASS MULTIPLEXING

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation can copy a deployed model by training a student on its logits or features. The student inherits a watermark only through the output it imitates. Both outputs therefore need a mark, yet most distillation watermarks cover only one. **TwinMark** instead writes one secret payload into what each distillation objective preserves, the second moment of normalized features and the class-mean logits. Both marks use linear readouts that give sufficient conditions for bit recovery on fixed audit inputs, while detection is tested separately under a declared null. In a ten-seed CIFAR-100 study, each mark is detected in every student that imitates its output and stays at chance otherwise. Marking costs the teacher 1.3 accuracy points and the audited embedding of feature-distilled students 4.4 nearest-neighbor points. Second-moment bounds certify 50 to 60 of 64 bits in every feature-distilled student, whereas pointwise bounds and the evaluated logit bounds certify none. Class multiplexing, which signs the payload per class, raises the logit decoder's rank ceiling and lowers bit errors at equal energy, and even at $10\times$ over-encoding, 1,024 bits on 100 classes stay detectable in every student. Detection extends to 39 of 40 students of four other architectures and to segmentation, object detection and satellite-navigation jammer detection and classification. These results establish output-matched inheritance under the evaluated protocols, not resistance to all utility-preserving transformations.

1 INTRODUCTION

Query access to a deployed model can be enough to copy its behavior (Tramèr et al., 2016; Orekondy et al., 2019). Knowledge distillation (KD) does so by training a new *student* network to imitate a classifier's probabilities (Hinton et al., 2015) or to match an embedding service's features (Romero et al., 2015; Liu et al., 2022). Watermarking lets the owner of the original model, the *teacher*, detect such copies by testing a suspect for a secret signal. Distillation makes this difficult. The student copies none of the teacher's trained weights and never sees the owner's key (Section 2), so any mark it inherits must travel through the outputs it imitates. A student that imitates the classifier's scores (logits) is never asked to reproduce a mark in the teacher's features, and one that matches features is never asked to reproduce a mark in the logits. A model served through both interfaces therefore needs a mark on each served output, which we call a *surface*.

Existing watermarks address parts of this problem. Weight watermarks (Uchida et al., 2017) and trigger-set watermarks (Adi et al., 2018; Zhang et al., 2018) are verified through the teacher's parameters or its responses to crafted inputs, and a student distilled on ordinary inputs is not trained to reproduce either. Watermarks designed for distillation instead place the mark in what the student learns. Some perturb the served probabilities or logits (Charette et al., 2022; Zhao et al., 2022; Lyu et al., 2026), others entangle the mark with task representations (Jia et al., 2021; Lv et al., 2024b), and encoder watermarks protect self-supervised representations (Cong et al., 2022; Lv et al., 2024a; Peng et al., 2023). Each of them, however, is read through a single surface. Reading two surfaces is not new in itself, since DeepSigns (Rouhani et al., 2019) and SEAL (Dai et al., 2025) already verify through both hidden layers and outputs. Formal guarantees, where they exist, answer different questions. Randomized smoothing covers bounded changes to the parameters (Bansal et al., 2022), and the closest analysis relates the single ownership statistic of the cosine watermark CosWM to the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

teacher–student output discrepancy (Charette et al., 2022) (Appendix A). We therefore ask which statistic of each output can carry one secret multi-bit message, or *payload*, into a distilled student, how many bits it can hold, and which of them the student's measured imitation error guarantees.

We introduce **TwinMark**, which writes one K-bit payload, derived from the owner's key, into the statistic that each distillation objective drives the student to reproduce (Figure 1). Feature-matching KD (FM-KD) trains the student to reproduce the teacher's normalized features, so the second-moment feature writer (SM-Feat) encodes each bit in the sign of a key-derived projection of their average outer product over fixed carrier inputs. Kullback–Leibler KD (KL-KD) instead matches temperature-softened class probabilities, which determine the logits up to a common shift. The class-multiplexed logit writer (Mux-Logit) therefore adds a P-dimensional code to the logits in directions that the probabilities fully determine (Section 3). It signs the payload differently in each of the m classes, which we call *class multiplexing*, so the class-mean logits give m distinct measurements of one payload. This raises the *rank ceiling*, an upper limit on the bits that a linear decoder can resolve, from $P < m$ to mP. To audit a suspect, the owner needs only the key and natural inputs, labeled for the logit channel. Each exposed surface is read as a separate *channel*, and calibrated tests at a fixed total significance level can detect the payload even when some bits are wrong (Section 4). Because both decoders are linear in averages, the measured imitation error also bounds how far each bit's score can move, so every bit whose teacher score clears that bound is certified (Section 5).

Our contributions, evaluated in Section 6, are:

1. *Surface-matched inheritance of one payload.* Each TwinMark channel is inherited by the students that imitate its surface and not by the others. Joint training leaves each channel about as recoverable as a single writer, and training-aware logit marking is inherited more often than a logit writer added after training.
2. *Certificates that trained students meet.* Because the feature certificate bounds the second moment after averaging, the student's deviations on different carriers can cancel. It certifies detection in every primary FM-KD student, where bounds applied input by input certify nothing, and a training target chosen from the key alone removes all feature-bit errors. The logit bounds certify no KL-KD student, losing every bit once cross-class cancellation is dropped.
3. *Signatures beyond the class count.* Class multiplexing raises the rank ceiling of the logit decoder. Key-derived signatures up to ten times the number of classes keep a full-rank decoder and remain detectable after distillation, and at equal injection energy multiplexing leaves fewer bits wrong.
4. *Breadth and boundaries.* The mark carries over to nearly all students of four other architectures and, because SM-Feat needs only an exposed embedding, to lesion segmentation (Codella et al., 2018), object detection (Everingham et al., 2010) and Global Navigation Satellite System (GNSS) jammer detection and classification (Heublein et al., 2025). Fixed removal recipes strip the logit mark only at an accuracy cost that clean models pay too, and a rotation or a selective change of the public carriers evades the feature verifier.

2 Problem Setting and Threat Model

We now make the setting of Section 1 precise. An owner serves a teacher T through its m-class logits, an embedding, or both. An adversary distills a student S from one of these surfaces, and the owner audits suspects with a secret key.

Models and objectives. For a model $F \in \{T, S\}$ and an input x with label $y(x) \in \{1, \dots, m\}$, a comparison head P_F maps the exposed hidden feature $h_F(x)$ into a d-dimensional comparison space declared by the owner. Write $v_F(x) = P_F(h_F(x))$ and $u_F(x) = v_F(x)/\|v_F(x)\|_2$ for nonzero projection outputs in this common space, $z_F(x) \in \mathbb{R}^m$ for the logits and $\pi_F^{(\tau)} = \mathrm{softmax}(z_F/\tau)$ at temperature $\tau > 0$, using the served logits for T (Appendix B lists the notation). KD trains S on transfer inputs from a distribution $\mathcal{D}$, possibly with a task loss, and FM-KD and KL-KD use the canonical objectives

$$\mathcal{L}_{\mathrm{FD}} = \frac{1}{d}\mathbb{E}_{\mathcal{D}}\|u_S - u_T\|_2^2, \qquad \mathcal{L}_{\mathrm{KD}} = \tau^2 \mathbb{E}_{\mathcal{D}}\mathrm{KL}(\pi_T^{(\tau)} \| \pi_S^{(\tau)}). \tag{1}$$

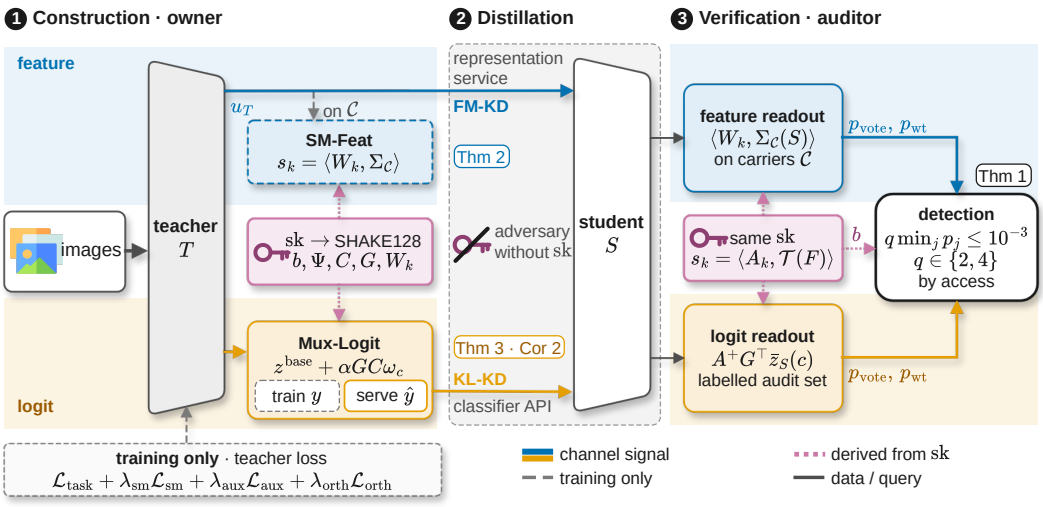

Figure 1: Overview. **(1)** SHAKE128 expands the secret key sk into the payload b and the readouts. The teacher writes b into its carrier feature second moment (SM-Feat) and logits (Mux-Logit, with y in training and $\hat{y}$ when served). **(2)** An adversary without sk distills a student by FM-KD or KL-KD. **(3)** The auditor reads each exposed channel with sk and detects when $q \min_j p_j \le 10^{-3}$ (q=2 p-values for one channel, 4 for both).

Here $1/d$ averages over feature coordinates, as implemented. Because probabilities fix the logits only up to a common shift, KL-KD targets the centered logits (Corollary 2, Appendix D.4).

Threat model. The adversary wants a useful student that cannot be linked to T. It lacks the key and the writers, never queries the verifier, and chooses the student's architecture, initialization and objective. Key probing, overwriting, and hard-label or top-k interfaces are out of scope. Each channel needs only the output its matching objective imitates: full logits, or full probabilities at a known temperature, for the logit channel, and the declared embedding for the feature channel. The inheritance claim for features concerns the declared embedding coordinates and the responses on fixed, public carriers. A rotated embedding, which the key alone cannot undo (Proposition 3, Appendix C), a replaced comparison head and an input-selective wrapper that alters only the carrier responses are evasion boundaries of the verifier. Task utility therefore does not imply that the mark survives, and the guarantees of Section 5 hold only under their output-fidelity conditions.

3 The TwinMark Construction

This section specifies how the key fixes the payload and how the two writers of Figure 1 place it in each surface.

Key derivation. SHAKE128, a SHA-3 extendable-output function (National Institute of Standards and Technology, 2015), expands the secret key sk with one domain tag per object into the payload $b \in \{\pm 1\}^K$, class signatures $\Psi \in \{\pm 1\}^{m \times K}$ with rows ψ_c, and a codebook $C \in \mathbb{R}^{P \times K}$ with unit-norm columns. The logit frame $G \in \mathbb{R}^{m \times P}$ satisfies $G^\top G = I_P$ and $G^\top \mathbf{1} = 0$, with $P \le m - 1$ and $\mathbf{1}$ the all-ones vector, along which softmax is invariant. The feature readouts are $W_k = V_k D_k V_k^\top$, where $V_k \in \mathbb{R}^{d \times r}$ has orthonormal columns, r is the readout rank and D_k is diagonal with entries ± 1. The per-class payload is $\omega_c = b \odot \psi_c$, with $\odot$ the elementwise product. The fixed carriers $\mathcal{C} = \{x_\ell\}_{\ell=1}^L$ are not secret (in classification, the first L training images), and the logit channel uses a separate labeled audit set (Appendix E).

Second-moment feature writer (SM-Feat). The feature summary and its bit scores are

$$\Sigma_C^T = \frac{1}{L} \sum_{\ell=1}^{L} u_T(x_\ell) u_T(x_\ell)^\top, \qquad s_k^{(\text{sm})}(T) = \operatorname{tr}(W_k \Sigma_C^T).$$

Here C marks the carriers, tr is the trace, and the same formulas define Σ_C^F and $s_k^{(\mathrm{sm})}(F)$ for any model F. The hinge objective

$$\mathcal{L}_{\mathrm{sm}} = \frac{1}{K}\sum_k \mathrm{ReLU}\big(m_{\mathrm{target}} - b_k s_k^{(\mathrm{sm})}(T)\big) \tag{2}$$

pushes each signed score toward a training target m_{target}, whereas certificates use the achieved margin m_{sm}. The key alone also brackets the largest margin that a trace-one positive semidefinite matrix can give all bits, and a positive lower bracket fixes a feasible target before training (Appendix F.3).

Class-multiplexed logit writer (Mux-Logit). During training, the teacher adds the label-conditioned payload with amplitude $\alpha > 0$ to the logits z_T^{base} of its base head, the ordinary classifier:

$$z_T^{\mathrm{lab}}(x) = z_T^{\mathrm{base}}(x) + \alpha G C \omega_{y(x)}. \tag{3}$$

After projection by $G^\top$, the class-c mean logit thus carries $\alpha C\omega_c$. The served teacher replaces $y(x)$ by the base head's prediction $\hat{y}(x) = \arg\max_i z_T^{\mathrm{base}}(x)_i$, giving $z_T^{\mathrm{serve}} = z_T^{\mathrm{base}} + \alpha G C \omega_{\hat{y}}$ (Definition 3, Appendix C), which differs from training only on misclassified inputs.

Training objective. The teacher minimizes

$$\mathcal{L}_T = \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{sm}}\mathcal{L}_{\mathrm{sm}} + \lambda_{\mathrm{aux}}\mathcal{L}_{\mathrm{aux}} + \lambda_{\mathrm{orth}}\mathcal{L}_{\mathrm{orth}}. \tag{4}$$

Here $\mathcal{L}_{\mathrm{task}}$ is the task loss (cross-entropy of the label-conditioned logits for classification), the auxiliary loss $\mathcal{L}_{\mathrm{aux}}$ fits keyed targets on carrier similarities to keyed training anchors, and the orthogonality loss $\mathcal{L}_{\mathrm{orth}}$ penalizes the leakage $G^\top z_T^{\mathrm{base}}$ of base logits into the code subspace (Appendix E).

4 VERIFICATION

Decoding. The auditor needs only sk, the saved frame G and natural inputs, with no triggers, learned decoder or fitted threshold. The feature channel decodes $\hat{b}_k = \mathrm{sign}(s_k^{(\mathrm{sm})}(S))$ on the carriers. For the logit channel, write $\mathbb{E}_c$ for the empirical mean over labeled audit inputs of class c, with at least one input per class, and M^+ for the Moore–Penrose pseudoinverse of a matrix M. The auditor projects each class mean onto the code subspace, $r_c = G^\top \mathbb{E}_c z_S(x) \in \mathbb{R}^P$, ideally $\alpha C\omega_c$. For $K < P$, each class determines its own signed copy $q_c = \alpha^{-1} C^+ r_c$, and the decoder undoes the class signs and averages them into the output score $s^{(o)} = m^{-1}\sum_c \psi_c \odot q_c$. For $K \geq P$, one class gives only P numbers for K bits, so the decoder combines all classes as $s^{(o)} = A^+ \mathrm{stack}_c(\alpha^{-1} r_c)$, where $A = \mathrm{stack}_c(C\,\mathrm{diag}(\psi_c)) \in \mathbb{R}^{mP\times K}$ stacks the m class blocks vertically, which is how class multiplexing enters the decoder. Both branches decode $\hat{b} = \mathrm{sign}(s^{(o)})$ with a fixed tie rule (Definition 4, Appendix C). In either channel, the bit error rate $\mathrm{BER}(S, b)$ is the fraction of bits with $\hat{b}_k \neq b_k$, and exact recovery means $\mathrm{BER}(S, b) = 0$.

Payload tests. Under the null hypothesis H_0, the suspect and every choice shaping its score vector s are independent of a uniformly random payload, as for a model trained without the key, so each decoded sign matches its bit with probability one half.

Theorem 1 (Conditional payload tests and detection)**.** *Condition on the suspect, evaluation inputs and labels, non-payload readout material, decoder, and all choices affecting the score construction or selection of the tested hypothesis. Under H_0, suppose b remains uniform on $\{\pm 1\}^K$ and independent of the resulting score vector s. Use a fixed $+1$ tie rule at zero.* ***(A) Bit agreement.*** *For $M = \sum_k \mathbf{1}\{\mathrm{sign}(s_k) = b_k\}$,*

$$p_{\mathrm{vote}} = \Pr[\mathrm{Binom}(K, 1/2) \geq M]$$

is an exact upper-tail p-value. ***(B) Weighted agreement.*** *For $s \neq 0$, let $t = \langle b, s\rangle/\|s\|_2$. The quantity $p_{\mathrm{wt}} = \exp[-\max(t, 0)^2/2]$ is a conservative p-value; set it to one when $s = 0$.* ***(C) Declared test family.*** *With q tests fixed in advance, $p_{\mathrm{family}} = \min(1, q\min_j p_j)$ controls Type-I error at the chosen total level, without independence between tests. If a recovery certificate guarantees at least r correct bits, then $p_{\mathrm{vote}} \leq \Pr[\mathrm{Binom}(K, 1/2) \geq r]$. This gives a sufficient detection condition even when the certificate does not cover every bit.*

The theorem's final clause lets a partial certificate (Proposition 1) imply detection. The auditor fixes its channels before querying and splits the total level α_{det} over $q = 2$ tests for one surface or $q = 4$ for both, valid even when channels share bits. The theorem is exact when the payload is independent of all other readout material, which SHAKE expansion of one finite seed does not guarantee, so we also calibrate the implemented detector over complete key generation (Appendix D.5) and check it on independently trained negative models (Section 6).

5 From Imitation Error to Certified Bits

The tests above hold for any suspect, so we now ask when a distilled student must make them reject. A *certificate* is a sufficient condition, evaluated on saved teacher and student outputs at the fixed audit inputs, that guarantees a decoded bit is correct. It assumes nothing about the student's architecture or training, but a small KD loss on the adversary's own inputs does not by itself meet it. It thus explains an observed recovery, not survival over a prespecified attack class (proofs in Appendix D, procedure in Appendix E.1).

Certified bits from linear readouts. Both scores are linear in an output statistic, the carrier second moment or the class-mean logits $\bar{z}_F(c) = \mathbb{E}_c z_F$ (Definition 1, Appendix C), so a discrepancy in the statistic bounds each score change. We write s_k for the kth score of either channel.

Proposition 1 (Certified partial recovery)**.** *For either channel, let $\mu_k = b_k s_k(T)$ be the signed teacher margin and suppose $|s_k(S) - s_k(T)| \le e_k$. Then*

$$\mathcal{J} = \{k : \mu_k > e_k\}, \qquad \mathrm{BER}(S, b) \le 1 - \frac{|\mathcal{J}|}{K}. \tag{5}$$

Indeed, every $k \in \mathcal{J}$ has $b_k s_k(S) > 0$. The bound needs neither independent errors nor a positive margin on every teacher bit. The following results supply the e_k.

SM-Feat: *averaging before bounding*. FM-KD keeps u_S close to u_T on average.

Theorem 2 (SM-Feat transfer certificate)**.** *Let u_T, u_S be unit vectors in a common comparison space and let $W_k = V_k D_k V_k^\top$ with $V_k^\top V_k = I_r$ and diagonal entries of D_k in $\{\pm 1\}$. Suppose the* measured *teacher margins satisfy $b_k s_k^{(\mathrm{sm})}(T) \ge m_{\mathrm{sm}} > 0$. Writing $d_x = \|u_S(x) - u_T(x)\|_2$, $\bar{\varepsilon}_h = \mathbb{E}_{\mathcal{C}} d_x$, and $\Delta\Sigma = \Sigma_C^S - \Sigma_C^T$, for every k,*

$$\begin{aligned} |s_k^{(\mathrm{sm})}(S) - s_k^{(\mathrm{sm})}(T)| &\le e_k^{\Sigma} := \|V_k^\top \Delta\Sigma V_k\|_* \\ &\le \mathbb{E}_{\mathcal{C}}[d_x \sqrt{4 - d_x^2}] \le 2\bar{\varepsilon}_h. \end{aligned} \tag{6}$$

Here $\|\cdot\|_*$ is the nuclear norm. Thus $m_{\mathrm{sm}} > \max_k e_k^{\Sigma}$ suffices for exact recovery; $m_{\mathrm{sm}} > 2\bar{\varepsilon}_h$ is a simpler, stronger requirement. A bit-specific measured margin may replace m_{sm} in Eq. 5. Keeping the action of each W_k before taking norms gives a tighter directional budget (Proposition 4); bounding the second-moment change after averaging uniformly refines that budget and retains cancellation across carriers (Eq. 13).

Proposition 4 and Eq. 13 appear in Appendix D.2, which also bounds $\bar{\varepsilon}_h$ by the FM-KD loss on the carriers. Averaging before bounding lets deviations of different carriers cancel inside $\Delta\Sigma$, whereas pointwise budgets add their magnitudes, so e_k^{Σ} can certify bits that pointwise budgets miss (Section 6).

Mux-Logit: *residuals and the KL bridge*. Write $\varepsilon_{\mathrm{top1}}(c)$ for the base head's error rate on audit inputs of class c, $\|\cdot\|_{\mathrm{op}}$ for the operator norm and $\sigma_{\min}$ for the smallest singular value.

Theorem 3 (Mux-Logit residual and recovery)**.** *Let $\Delta z = z_T^{\mathrm{serve}} - z_S$ and $\Delta\widetilde{z} = \Delta z - m^{-1}(\mathbf{1}^\top \Delta z)\mathbf{1}$. For each class define*

$$d_c = \overline{\|\Delta\widetilde{z}\|}_c := \mathbb{E}_c \|\Delta\widetilde{z}\|_2, \quad B_c = \|G^\top \mathbb{E}_c z_T^{\mathrm{base}}\|_2, \quad E_c = \alpha \|\mathbb{E}_c C(\omega_{\hat{y}} - \omega_c)\|_2. \tag{7}$$

Then $E_c \le 2\alpha \varepsilon_{\mathrm{top1}}(c) \|C\|_{\mathrm{op}} \sqrt{K}$ and

$$\|r_c - \alpha C \omega_c\|_2 \le R_c := B_c + d_c + E_c. \tag{8}$$

For $K < P$ and full-column-rank C, $\|s^{(o)} - b\|_\infty \le (m\alpha\sigma_{\min}(C))^{-1} \sum_c R_c$; a value strictly below one guarantees exact recovery. The stronger condition $R_c < \alpha\sigma_{\min}(C)$ for every class also recovers each ω_c by taking the signs of q_c. For full-column-rank A, the stacked decoder instead satisfies

$$\|s^{(o)} - b\|_\infty \le \frac{\sqrt{\sum_c R_c^2}}{\alpha\sigma_{\min}(A)}; \tag{9}$$

again, a value strictly below one suffices.

The residual R_c separates base-logit leakage B_c, which $\mathcal{L}_{\text{orth}}$ suppresses, the distillation error d_c, the only student-dependent term, and the serving mismatch E_c, while $\sigma_{\min}(C)$ or $\sigma_{\min}(A)$ sets its amplification. A rowwise budget against the served teacher needs no full rank and is often tighter (Proposition 5, Appendix D.3). KL-KD controls probabilities rather than logits, and Corollary 2 (Appendix D.4) bridges the two using only the served teacher. If every teacher probability at the audit temperature τ is at least δ_c on class c, then d_c is bounded by an increasing function of the mean unscaled divergence $\varepsilon_{\text{KL}}(c)$ of class c, approaching $\tau\sqrt{2\varepsilon_{\text{KL}}(c)/\delta_c}$ as it vanishes, with no assumption on the student.

Class multiplexing: *rank and conditioning*. The stacked bound of Theorem 3 divides by $\sigma_{\min}(A)$, and without class signatures every class would give the same P measurements, leaving A with rank at most P.

Theorem 4 (Mux-Logit rank and conditional concentration)**.** *For any fixed unit-column C, $\operatorname{rank}(A) \le \min(m \operatorname{rank}(C), K) \le \min(mP, K)$. Under independent uniform signs in Ψ, independent of C, set $H = C^\top C - I_K$, $v = \max_k (H^2)_{kk}$, and $\ell = \log(2K/\beta)$ for $0 < \beta < 1$. With probability at least $1 - \beta$,*

$$\|A^\top A - mI_K\|_{\text{op}} \le t_\beta := \sqrt{2mv\ell} + \tfrac{2}{3}\|H\|_{\text{op}}\ell. \tag{10}$$

If $t_\beta < m$, then A has full column rank and $\sigma_{\min}(A) \ge \sqrt{m - t_\beta}$.

The rank ceiling thus rises from P to mP, which is necessary but not sufficient. Under independent signatures, an idealization of the SHAKE expansion, cross-talk between codewords enters $A^\top A - mI_K$ as zero-mean terms that cancel on average (proof in Appendix D.6, via matrix Bernstein (Tropp, 2012)). As the bound need not be informative for large payloads, we certify each realized design by its measured rank and $\sigma_{\min}(A)$, computed from the key alone (Corollary 1, Appendix C).

6 Experimental Results

Setup. The primary study distills CIFAR-100 (Krizhevsky, 2009) ResNet-18 teachers into ResNet-18 students (He et al., 2016) over ten prespecified seeds, with one unscreened owner key per seed. Teachers train for 60 epochs and students for 30, with shared data, initialization and schedule. The joint writer uses $K = 64$ bits and a $P = 32$ logit code (Appendix E lists all settings). FM-KD trains the student's comparison head, whereas KL-KD, at $\tau = 4.538$, leaves it at initialization. Both add cross-entropy, and neither sees the key. Each access scenario is tested at total level $\alpha_{\text{det}} = 10^{-3}$. The recipe was inherited from runs selected on evaluation accuracy, so all comparisons are fixed-recipe comparisons.

Each objective transfers only the channel it imitates. The component study isolates the surface that carries each mark (Table 1). A feature-only mark survives FM-KD in every seed at 2.66% BER, but a KL-KD student, whose comparison head that objective never trains, reads it at 48.12% BER with no detection. A logit-only mark shows the converse (11.41% after KL-KD, 49.69% after FM-KD), although FM-KD students fit logits by cross-entropy. Every unmatched readout of the joint students also stays at chance, although all students start from their teacher's initialization, under which distillation can transmit traits that the imitated outputs do not express (Cloud et al., 2026).

Joint training retains both watermark channels. Joint training keeps feature BER at 3.44% after FM-KD and logit BER at 11.09% after KL-KD, detecting all ten seeds, within one standard deviation (SD) of the single writers. Random class signatures make the logit code $\psi_c \odot b$ independent of b, so an independent logit payload amounts to another signature draw, and on the first five seeds it keeps

Table 1: Component study on CIFAR-100 (ResNet-18, K=64, ten seeds, mean$_{\pm SD}$ in %). F/L: feature/logit readout, the unmatched one as a mean. Det.: detections under F/L/dual access at level 10^{-3}. Shading: the imitated readout. Last block: joint students after removal, from 58.15% (FM-KD) and 59.12% (KL-KD). †Fresh key per clean model. ‡Native SNR>8 rule (teachers 9/10), not a level-10^{-3} test.

Method	Teacher Acc.	**FM-KD** students (feature matching) Acc.	F BER	L BER	Det.	**KL-KD** students (logit matching) Acc.	F BER	L BER	Det.
Clean	$62.64_{\pm 0.29}$	$58.19_{\pm 0.27}$	–	–	0/0/0†	$59.51_{\pm 0.38}$	–	–	0/0/0†
Feature only	$61.65_{\pm 0.30}$	$58.03_{\pm 0.32}$	$2.66_{\pm 3.46}$	48.91	10/0/10	$59.37_{\pm 0.27}$	48.12	$50.00_{\pm 7.37}$	0/0/0
Logit only	$61.57_{\pm 0.40}$	$57.96_{\pm 0.46}$	$50.00_{\pm 6.55}$	49.69	0/0/0	$59.23_{\pm 0.27}$	48.91	$11.41_{\pm 4.04}$	0/10/10
Feature + posthoc logit	$61.63_{\pm 0.26}$	$58.03_{\pm 0.32}$	$2.66_{\pm 3.46}$	48.91	10/0/10	$59.40_{\pm 0.31}$	49.06	$36.41_{\pm 6.12}$	0/2/1
Joint, no aux. loss	$61.63_{\pm 0.40}$	$57.79_{\pm 0.30}$	$7.66_{\pm 3.86}$	49.38	10/0/10	$59.02_{\pm 0.35}$	49.53	$9.53_{\pm 4.06}$	0/10/10
TwinMark (joint)	$61.35_{\pm 0.41}$	$58.15_{\pm 0.36}$	$3.44_{\pm 2.74}$	49.53	10/0/10	$59.12_{\pm 0.36}$	49.84	$11.09_{\pm 5.02}$	0/10/10
TwinMark, key-only target	$61.19_{\pm 0.36}$	$58.16_{\pm 0.43}$	$0.00_{\pm 0.00}$	49.06	10/0/10	$58.95_{\pm 0.36}$	48.59	$10.16_{\pm 4.56}$	0/10/10
CosWM	$62.44_{\pm 0.22}$	$57.91_{\pm 0.23}$	SNR>8: 0/10‡			$59.38_{\pm 0.29}$	SNR>8: 0/10‡		
Removal after extraction									
WRT weight shifting	–	$43.24_{\pm 0.75}$	$3.91_{\pm 2.88}$	48.59	10/0/10	$43.27_{\pm 1.35}$	49.06	$47.66_{\pm 5.76}$	0/0/0
Neural Dehydration	–	$40.07_{\pm 0.93}$	$3.44_{\pm 3.52}$	48.75	10/0/10	$41.31_{\pm 0.66}$	49.38	$22.19_{\pm 5.84}$	0/10/9
FTAL	–	$40.76_{\pm 1.51}$	$3.75_{\pm 2.68}$	47.66	10/0/10	$40.23_{\pm 0.83}$	47.81	$45.16_{\pm 6.77}$	0/0/0
RTAL	–	$44.06_{\pm 1.03}$	$4.84_{\pm 3.25}$	49.53	10/0/10	$43.78_{\pm 1.31}$	48.91	$46.09_{\pm 7.84}$	0/0/0
Fine-pruning	–	$47.07_{\pm 0.77}$	$10.16_{\pm 3.55}$	47.81	10/0/10	$47.40_{\pm 0.95}$	47.50	$45.47_{\pm 6.61}$	0/0/0

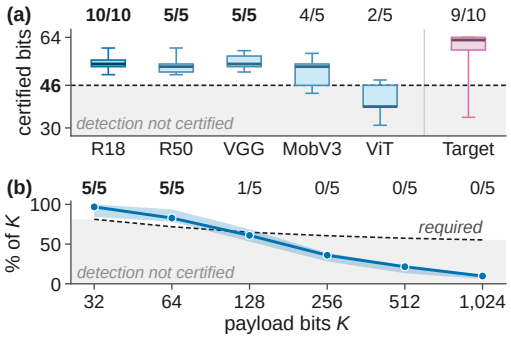

Figure 2: Second-moment certificates. (a) Certified bits per FM-KD student (boxes: quartiles and range) against the 46 that force vote rejection. Counts: certified detections. Pointwise and logit bounds certify no bit. (b) Certified share of K (median, range) against the share required.

Table 2: Post-processing attacks on the ten joint CIFAR-100 teachers (%). Rows test the feature channel, since the logit writer acts at serving time. Det./Cert.: detected/certified seeds at level 10^{-3}. No clean teacher is flagged. Other datasets and parameters: Appendix F.4.

Attack on the teacher	Acc.	F BER	Det.	Cert.
None	61.36	$3.28_{\pm 3.33}$	10/10	10/10
Weights of a stolen teacher				
INT8 quantization	61.36	$3.28_{\pm 3.33}$	10/10	10/10
INT4 quantization	51.04	$3.91_{\pm 2.78}$	10/10	10/10
Pruning (50%)	57.57	$3.44_{\pm 2.83}$	10/10	10/10
Weight noise	60.15	$3.28_{\pm 2.90}$	10/10	10/10
Fine-tuning (10 ep.)	57.23	$3.12_{\pm 2.85}$	10/10	10/10
Served embedding (logits unchanged)				
Orthogonal rotation	61.36	$51.41_{\pm 7.81}$	0/10	0/10
Non-orthogonal warp	61.36	$3.28_{\pm 3.72}$	10/10	10/10
Feature noise	61.36	$5.47_{\pm 2.47}$	10/10	10/10
ZCA whitening	61.36	$6.56_{\pm 2.64}$	10/10	0/10

every student detected. Marking costs the teacher 1.29 accuracy points and moves student classifiers by under half a point against their clean-teacher pairs, but a fixed cosine 5-nearest-neighbor probe of the verified embedding loses 4.36 points in FM-KD students (Appendix F.1). A logit writer added after training a feature-only teacher reaches only two of ten KL-KD students.

A matched distillation watermark is not inherited. With its published probability writer and spectral signal-to-noise ratio (SNR) detector, and our data, architectures, initializations and budgets, CosWM (Charette et al., 2022) flags nine of ten teachers but no student after either attack (Table 1). At a common budget of 100 held-out queries and conditional level 10^{-3}, TwinMark detects nine of ten KL-KD students and CosWM none, even among the six whose teacher it detects (Appendix Table 8). CosWM reports an ownership signal rather than a payload and draws different query classes, so this compares these implementations and queries, not robustness to distillation.

Averaging before bounding certifies every primary feature student. Evaluated on the saved outputs, the second-moment bound of Section 5 certifies 50 to 60 of 64 bits in each primary FM-KD student, above the 46 that force the analytic vote test to reject, so detection is certified in all ten

Table 3: Payload size against class count m and code dimension P (five seeds, mean$_{\pm\mathrm{SD}}$ in %). CIFAR-100 varies only K (K=64: first five seeds of Table 1). Rank and mean $\sigma_{\min}$: stacked logit decoder, rank $\leq \min(K, mP)$. Teacher accuracy: mean. Δ: paired student accuracy change. Det.: matching single access. Shaded: $K > m$.

					Teacher	**FM-KD**: feature channel			**KL-KD**: logit channel		
Data (m, P)	K	K/m	Rank	$\sigma_{\min}$	Acc.	Δ	F BER	Det.	Δ	L BER	Det.
CIFAR-100 (100, 32)	32	0.32×	32	9.10	61.60	$-0.23_{\pm 0.66}$	$1.25_{\pm 1.71}$	5/5	$-0.71_{\pm 0.25}$	$5.62_{\pm 3.42}$	5/5
	64	0.64×	64	8.66	61.40	$-0.09_{\pm 0.31}$	$3.75_{\pm 3.76}$	5/5	$-0.55_{\pm 0.58}$	$9.69_{\pm 2.04}$	5/5
	128	1.28×	128	8.06	61.37	$-0.15_{\pm 0.19}$	$9.84_{\pm 3.65}$	5/5	$-0.31_{\pm 0.65}$	$9.22_{\pm 2.02}$	5/5
	256	2.56×	256	7.24	61.04	$+0.05_{\pm 0.44}$	$20.23_{\pm 2.43}$	5/5	$-0.52_{\pm 0.30}$	$9.84_{\pm 0.97}$	5/5
	512	5.12×	512	6.03	61.17	$-0.31_{\pm 0.61}$	$32.03_{\pm 2.03}$	5/5	$-0.62_{\pm 0.64}$	$13.67_{\pm 2.09}$	5/5
	1024	10.24×	1024	4.36	60.82	$-0.22_{\pm 0.46}$	$39.34_{\pm 1.15}$	5/5	$-0.63_{\pm 0.45}$	$16.45_{\pm 1.37}$	5/5
MiniImageNet (100, 16)	128	1.28×	128	–	59.51	$-0.37_{\pm 0.84}$	$17.97_{\pm 4.94}$	5/5	$+0.05_{\pm 0.69}$	$0.00_{\pm 0.00}$	5/5
CIFAR-10 (10, 2)	32	3.20×	**20** $< K$	–	87.19	$-0.05_{\pm 0.35}$	$2.50_{\pm 4.07}$	5/5	$-0.29_{\pm 0.26}$	$31.25_{\pm 5.85}$	0/5

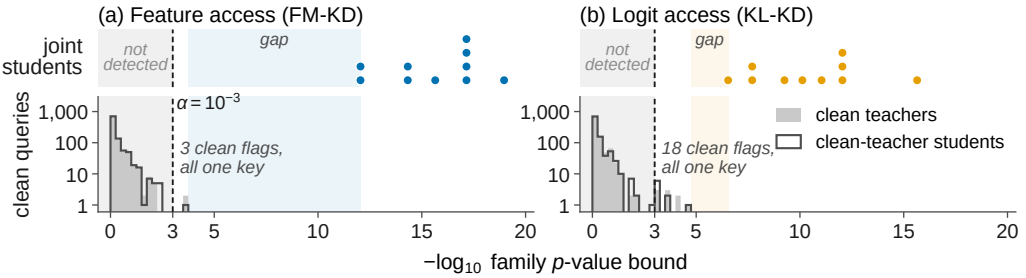

Figure 3: Clean battery queries (histograms) and joint students (dots) under the conditional null.

(Figure 2a), whereas pointwise norm and directional bounds certify no bit (Appendix Table 9). A hinge target computed from the key alone before training (Appendix F.3) removes every feature-bit error (Table 1), and its certificates cover all 64 bits in four seeds and detection in nine.

The evaluated logit bounds remain vacuous. The rowwise and teacher-floor KL bounds certify no bit, although the stacked design is well conditioned and every teacher margin is positive. The observed score changes would certify 569 of 640 bits, but dropping cross-class cancellation alone certifies none, before any KL step (Appendix F.3). KL-KD detection thus rests on the calibrated tests alone.

Post-processing a stolen teacher leaves the feature mark detectable. Nine fixed attacks on the ten CIFAR-100 teachers probe the mark once the weights leave the owner (Table 2). All five weight attacks keep detection and certification in every seed, even at a 10-point accuracy cost for four-bit (INT4) quantization. A warp, feature noise and whitening by zero-phase component analysis (ZCA) of the served embedding keep detection, although ZCA removes the certificate. An orthogonal rotation evades the key-only verifier, as Appendix Proposition 3 predicts, and so does a key-independent replacement of the carrier embeddings of FM-KD students. On CIFAR-10 and MiniImageNet, only the rotation and, in two MiniImageNet teachers, ZCA prevent detection (Appendix F.4).

Class multiplexing extends the signature beyond the class count. At fixed amplitude α, the stacked decoder keeps full column rank up to $K = 1{,}024$ while its conditioning degrades (Table 3). Logit BER after KL-KD rises only to 16.45%, and every student is detected at every size, but feature certificates cover detection only up to $K = 64$ (Figure 2b). With the injection energy $\alpha^2 K$ held at its primary level, every student is still detected up to $K = 1{,}024$, at 40.5% logit BER. At that energy and $K = 128$, removing class multiplexing ($\psi_c = \mathbf{1}$) cuts the rank to 32 and raises logit BER from 17.7% to 30.2% in every seed, while both arms stay detected (Appendix F.5). As detection power grows with K, these results show detectable signatures longer than the class count, not capacity at a fixed distortion. On CIFAR-10, whose 32-bit payload exceeds $mP = 20$, KL-KD gives 31.25% logit BER and no detections, although rank alone does not explain this (Appendix F.5).

Table 4: Other students and tasks (five seeds, mean$_{\pm \mathrm{SD}}$). CIFAR-100 students distill the first five teacher pairs of Table 1. Δ: paired student utility change in points. Det.: matching single access, or dual (^{D}). No clean model is flagged in any task. †Clean-student SD 14.3, no utility claim.

		Feature KD: feature channel			**KL-KD**: logit channel		
Student model	Data	Δ	F BER	Det.	Δ	L BER	Det.
ResNet-50	CIFAR-100	$-0.17_{\pm 0.70}$	$3.44_{\pm 3.56}$	5/5	$-0.19_{\pm 0.66}$	$5.62_{\pm 2.61}$	5/5
VGG-16-BN		$+0.11_{\pm 0.30}$	$3.44_{\pm 3.39}$	5/5	$-0.62_{\pm 0.21}$	$0.62_{\pm 0.86}$	5/5
MobileNetV3-L		$0.00_{\pm 0.54}$	$4.06_{\pm 3.60}$	5/5	$-0.50_{\pm 0.54}$	$22.81_{\pm 4.07}$	5/5
ViT-S/16 (pretrained)		$-0.15_{\pm 0.32}$	$4.06_{\pm 3.60}$	5/5	$-0.67_{\pm 0.22}$	$27.81_{\pm 7.44}$	4/5
ViT-S/16 (pretrained)	MiniImageNet	$-0.61_{\pm 0.46}$	$13.44_{\pm 3.24}$	5/5^{D}	$-0.49_{\pm 0.32}$	$2.19_{\pm 0.65}$	5/5^{D}
EffNet-B4 U-Net (Dice)	ISIC 2018	$-0.04_{\pm 0.18}$	$3.44_{\pm 3.39}$	5/5	no logit writer		
Swin-T Faster R-CNN (mAP)	VOC 07+12	$-0.35_{\pm 0.74}$	$7.81_{\pm 6.54}$	5/5	no logit writer		
ResNet-18, 4-ch. (episodic acc.)	GNSS/CRPA	$+15.94_{\pm 11.47}{}^{\dagger}$	$4.06_{\pm 3.24}$	5/5	no logit writer		

Inheritance extends to other students and tasks. Distilling the first five teacher pairs into four other architectures gives 20 of 20 feature detections after FM-KD and 19 of 20 logit detections after KL-KD, with paired accuracy changes from -0.67 to $+0.11$ points (Table 4), although these students share no initial weights with their teachers. Logit BER is highest for MobileNetV3 and the pretrained Vision Transformer (ViT), whose initialization and batch size also differ, and one ViT seed escapes detection. MiniImageNet ViT students are all detected, but at 36.5–39.4% accuracy this is a detection result only. The feature writer also transfers unchanged to lesion segmentation, object detection and GNSS jammer detection and classification with a controlled reception pattern antenna (CRPA), where every student is detected. Segmentation students lose 0.04 Dice points, while the object-detection students stay too weak in mean average precision (mAP) to show more than transfer (Appendix G).

Fixed removal recipes erase the logit mark while degrading accuracy. Five fixed attacks target the extracted students of all ten seeds (Table 1, last block). Weight shifting from the Watermark Robustness Toolbox (WRT) (Lukas et al., 2022), fine-tuning all layers (FTAL), retraining all layers after a classifier reset (RTAL) and fine-pruning remove every KL-KD logit detection, and basic Neural Dehydration (Lu et al., 2024) none. WRT costs 14.7–15.9 accuracy points and the others 11 to 19, as much in clean-teacher students (Appendix Table 17), so these recipes are not utility-preserving and show vulnerability to fixed attacks, not a utility cost that removal must pay. The feature mark survives all five, none of which retrains the verified head, and fine-pruning gives the only false flag among 100 attacked clean students.

Unattacked clean models are not flagged. Ten independently trained clean teachers and their students give no false detection with one fresh, prespecified key per model, in any cohort or access scenario. Ten models bound a population false-positive rate only by 25.89% (one-sided 95%), so they check the executed procedure rather than estimate that rate. A battery of 100 keys per model probes the null more finely (Figure 3). Each dot there is one joint student, and dots stack where students recover the same number of payload bits. Every dot lies beyond every clean query, most narrowly under logit access. The few clean flags come from two keys, so the level holds on average over keys rather than per key. Rank tests over 1,999 regenerated keys, which need no exact payload independence, reproduce every analytic decision (Appendix F.8).

7 Conclusion

TwinMark encodes one keyed signature in feature second moments and class-conditioned logits, and under the evaluated distillation protocols each channel is inherited by the students that imitate it. Its feature certificates certify most bits on the fixed audit inputs, while the evaluated logit bounds certify none. Class multiplexing expands the ideal linear measurement rank and lowers bit errors at equal injection energy, but detectable long signatures do not establish information capacity. The feature verifier is coordinate dependent and evadable through its public carriers, and the removal experiments establish no intrinsic utility cost of erasure. These results characterize when watermark evidence follows output imitation and where it stops protecting.

LLM USAGE DISCLOSURE

Large language models (Claude Opus 5) were used to (i) assist as a general-purpose writing aid for grammar, wording, and clarity, and (ii) draft initial portions of the experiment code, which the authors independently reviewed, tested, and validated. LLMs were not credited as authors and made no autonomous scientific decisions. The authors take full responsibility for the correctness, integrity, and originality of all claims and results.

REFERENCES

Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning Your Weakness into a Strength: Watermarking Deep Neural Networks by Backdooring. In *Proc. of the USENIX Conf. on Security Symposium (SEC)*, pp. 1615–1631, August 2018. doi: 10.5555/3277203.3277324. URL https://dl.acm.org/doi/10.5555/3277203.3277324.

Arpit Bansal, Ping yeh Chiang, Michael Curry, Rajiv Jain, Curtis Wigington, Varun Manjunatha, John P Dickerson, and Tom Goldstein. Certified Neural Network Watermarks with Randomized Smoothing. In *Intl. Conf. on Machine Learning (ICML)*, volume 162, pp. 1450–1465, 2022. URL https://proceedings.mlr.press/v162/bansal22a.html.

Laurent Charette, Lingyang Chu, Yizhou Chen, Jian Pei, Lanjun Wang, and Yong Zhang. Cosine Model Watermarking Against Ensemble Distillation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 36, pp. 9512–9520, June 2022. doi: 10.1609/aaai.v36i9.21184. URL https://ojs.aaai.org/index.php/AAAI/article/view/21184.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Sören Mindermann, Jacob Hilton, Samuel Marks, and Owain Evans. Language models transmit behavioural traits through hidden signals in data. *Nature*, 652(8110):615–621, 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10319-8. URL https://doi.org/10.1038/s41586-026-10319-8.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172, 2018. doi: 10.1109/ISBI.2018.8363547.

Tianshuo Cong, Xinlei He, and Yang Zhang. SSLGuard: A Watermarking Scheme for Self-Supervised Learning Pre-trained Encoders. In *Proc. of the ACM SIGSAC Conf. on Computer and Communications Security (CCS)*, pp. 579–593, November 2022. doi: 10.1145/3548606.3559355. URL https://dl.acm.org/doi/10.1145/3548606.3559355.

Yanbo Dai, Zongjie Li, Zhenlan Ji, and Shuai Wang. SEAL: Subspace-Anchored Watermarks for LLM Ownership, 2025. URL https://arxiv.org/abs/2511.11356.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Intl. Conf. on Learning Representations (ICLR)*, January 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. In *International Journal of Computer Vision*, volume 88(2), pp. 303–338, 2010. doi: 10.1007/s11263-009-0275-4. URL https://doi.org/10.1007/s11263-009-0275-4.

Teng Fei, Xinjing Liu, Yanjun Zhang, Haizhuan Yuan, Tianqing Zhu, and Leo Yu Zhang. MarkErase: Defeating Entangled Watermarks in Model Extraction Attacks. In *Advances in Knowledge Discovery and Data Mining (PAKDD)*, pp. 78–90, June 2025. doi: 10.1007/978-981-96-8183-9_7. URL https://link.springer.com/chapter/10.1007/978-981-96-8183-9_7.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, June 2016. doi: 10.1109/CVPR.2016.90. URL https://ieeexplore.ieee.org/document/7780459.

Lucas Heublein, Tobias Feigl, Thorsten Nowak, Alexander Rügamer, Christopher Mutschler, and Felix Ott. Evaluating ML Robustness in GNSS Interference Classification, Characterization & Localization. In *IEEE Intl. Conf. on Localization and GNSS (ICL-GNSS)*, Rome, Italy, June 2025. doi: 10.1109/ICL-GNSS65520.2025.11046126. URL https://ieeexplore.ieee.org/document/11046126.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, March 2015. URL https://arxiv.org/abs/1503.02531.

Andrew Howard, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Hartwig Adam, and Quoc V. Le. Searching for MobileNetV3. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, pp. 1314–1324, Seoul, South Korea, October 2019. doi: 10.1109/ICCV.2019.00140. URL https://ieeexplore.ieee.org/document/9008835.

Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High Accuracy and High Fidelity Extraction of Neural Networks. In *Proc. of the USENIX Conf. on Security Symposium (SEC)*, volume 76, pp. 1345–1362, August 2020. doi: 10.5555/3489212.3489288. URL https://dl.acm.org/doi/10.5555/3489212.3489288.

Hengrui Jia, Christopher A. Choquette-Choo, Varun Chandrasekaran, and Nicolas Papernot. Entangled Watermarks as a Defense Against Model Extraction. In *Proc. of the USENIX Conf. on Security Symposium (SEC)*, pp. 1937–1954, August 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/jia.

Nikola Jovanović, Robin Staab, and Martin Vechev. Watermark Stealing in Large Language Models. In *Intl. Conf. on Machine Learning (ICML)*, volume 907, pp. 22570–22593, July 2024. doi: 10.5555/3692070.3692977. URL https://dl.acm.org/doi/10.5555/3692070.3692977.

Byungjoo Kim, Suyoung Lee, Seanie Lee, Sooel Son, and Sung Ju Hwang. Margin-Based Neural Network Watermarking. In *Intl. Conf. on Machine Learning (ICML)*, July 2023. doi: 10.5555/3618408.3619093. URL https://proceedings.mlr.press/v202/kim23o.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017. doi: 10.1109/CVPR.2017.106.

Yupei Liu, Jinyuan Jia, Hongbin Liu, and Neil Zhenqiang Gong. StolenEncoder: Stealing Pre-trained Encoders in Self-supervised Learning. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, pp. 2115–2128, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394505. doi: 10.1145/3548606.3560586. URL https://doi.org/10.1145/3548606.3560586.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, pp. 9992–10002, Montreal, QC, Canada, October 2021. doi: 10.1109/ICCV48922.2021.00986. URL https://ieeexplore.ieee.org/document/9710580.

Yifan Lu, Wenxuan Li, Mi Zhang, Xudong Pan, and Min Yang. Neural Dehydration: Effective Erasure of Black-box Watermarks from DNNs with Limited Data. CCS '24, pp. 675–689, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706363. doi: 10.1145/3658644.3690334. URL https://doi.org/10.1145/3658644.3690334.

Nils Lukas, Yuxuan Zhang, and Florian Kerschbaum. Deep Neural Network Fingerprinting by Conferrable Adversarial Examples. In *Intl. Conf. on Learning Representations (ICLR)*, January 2021. URL https://openreview.net/forum?id=VqzVhqxkjH1.

Nils Lukas, Edward Jiang, Xinda Li, and Florian Kerschbaum. SoK: How Robust is Image Classification Deep Neural Network Watermarking? In *IEEE Symposium on Security and Privacy (SP)*, pp. 787–804, 2022. doi: 10.1109/SP46214.2022.9833693. URL https://www.computer.org/csdl/proceedings-article/sp/2022/131600a052/1FlQuMr0vq8.

Peizhuo Lv, Pan Li, Shenchen Zhu, Shengzhi Zhang, Kai Chen, Ruigang Liang, Chang Yue, Fan Xiang, Yuling Cai, Hualong Ma, Yingjun Zhang, and Guozhu Meng. SSL-WM: A Black-Box Watermarking Approach for Encoders Pre-trained by Self-Supervised Learning. In *Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, March 2024a. doi: 10.14722/ndss.2024.24374. URL https://www.ndss-symposium.org/wp-content/uploads/2024-374-paper.pdf.

Peizhuo Lv, Hualong Ma, Kai Chen, Jiachen Zhou, Shengzhi Zhang, Ruigang Liang, Shenchen Zhu, Pan Li, and Yingjun Zhang. MEA-Defender: A Robust Watermark Against Model Extraction Attack. In *IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, May 2024b. doi: 10.1109/SP54263.2024.00099. URL https://ieeexplore.ieee.org/document/10646835.

Ning Lyu, Yuntao Liu, Yonghong Bai, and Zhiyuan Yan. PDF: PUF-based DNN Fingerprinting for Knowledge Distillation Traceability, 2026. URL https://arxiv.org/abs/2602.23587.

National Institute of Standards and Technology. SHA-3 Standard: Permutation-Based Hash and Extendable-Output Functions. Federal Information Processing Standards Publication 202, National Institute of Standards and Technology, Gaithersburg, MD, August 2015. URL https://doi.org/10.6028/NIST.FIPS.202.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff Nets: Stealing Functionality of Black-Box Models. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 4949–4958, Long Beach, CA, June 2019. doi: 10.1109/CVPR.2019.00509. URL https://ieeexplore.ieee.org/document/8953839.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational Knowledge Distillation. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971, Long Beach, CA, June 2019. doi: 10.1109/CVPR.2019.00409. URL https://ieeexplore.ieee.org/document/8954416.

Wenjun Peng, Jingwei Yi, Fangzhao Wu, Shangxi Wu, Bin Bin Zhu, Lingjuan Lyu, Binxing Jiao, Tong Xu, Guangzhong Sun, and Xing Xie. Are You Copying My Model? Protecting the Copyright of Large Language Models for EaaS via Backdoor Watermark. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7653–7668, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.423. URL https://aclanthology.org/2023.acl-long.423/.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, 2016. URL https://arxiv.org/abs/1506.01497.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for Thin Deep Nets. In *Intl. Conf. on Learning Representations (ICLR)*, March 2015. URL https://arxiv.org/abs/1412.6550.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. DeepSigns: An End-to-End Watermarking Framework for Ownership Protection of Deep Neural Networks. In *Proc. of the Intl. Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pp. 485–497, April 2019. doi: 10.1145/3297858.3304051. URL https://dl.acm.org/doi/10.1145/3297858.3304051.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Intl. Conf. on Learning Representations (ICLR)*, San Diego, CA, April 2015. URL https://arxiv.org/abs/1409.1556.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4080–4090, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Intl. Conf. on Machine Learning (ICML)*, volume 97, pp. 6105–6114, 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Representation Distillation. In *Intl. Conf. on Learning Representations (ICLR)*, December 2019. URL https://openreview.net/forum?id=SkgpBJrtvS.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing Machine Learning Models via Prediction APIs. In *Proc. of the USENIX Conf. on Security Symposium (SEC)*, pp. 601–618, August 2016. doi: 10.5555/3241094.3241142. URL https://dl.acm.org/doi/10.5555/3241094.3241142.

Joel A. Tropp. User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. ISSN 1615-3383. doi: 10.1007/s10208-011-9099-z. URL https://doi.org/10.1007/s10208-011-9099-z.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. *Scientific Data*, 5 (1):180161, 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.161. URL https://doi.org/10.1038/sdata.2018.161.

Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding Watermarks Into Deep Neural Networks. In *Proc. of the ACM Intl. Conf. on Multimedia Retrieval (ICMR)*, pp. 269–277, June 2017. doi: 10.1145/3078971.3078974. URL https://dl.acm.org/doi/10.1145/3078971.3078974.

Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *arXiv preprint arXiv:1606.04080*, volume abs/1606.04080, 2016. URL http://arxiv.org/abs/1606.04080.

Sergey Zagoruyko and Nikos Komodakis. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *Intl. Conf. on Learning Representations (ICLR)*, February 2017. URL https://openreview.net/forum?id=Sks9_ajex.

Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph. Stoecklin, Heqing Huang, and Ian Molloy. Protecting Intellectual Property of Deep Neural Networks with Watermarking. In *Proc. of the Asia Conf. on Computer and Communications Security (AsiaCCS)*, pp. 159–172, May 2018. doi: 10.1145/3196494.3196550. URL https://dl.acm.org/doi/10.1145/3196494.3196550.

Xuandong Zhao, Lei Li, and Yu-Xiang Wang. Distillation-Resistant Watermarking for Model Protection in NLP. In *Intl. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5044–5055, Abu Dhabi, United Arab Emirates, December 2022. doi: 10.18653/v1/2022.findings-emnlp.370. URL https://aclanthology.org/2022.findings-emnlp.370.