ECAI 2025
I. Lynce et al. (Eds.)

doi:10.3233/FAIA250907

Stealing Knowledge from Auditable Datasets

Hongyu Zhu[a], Sichu Liang[b], Wentao Hu[a], Wenwen Wang[c], Fangqi Li[a], Shilin Wang[a,*] and Zhuosheng Zhang[a]

[a]School of Computer Science, Shanghai Jiao Tong University, Shanghai, China
[b]School of Computer Science and Engineering, Southeast University, Nanjing, China
[c]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, USA

Abstract. The success of modern deep learning hinges on vast training data, much of which is scraped from the web and may include copyrighted or private content—raising serious legal and ethical concerns when used without authorization. Dataset provenance seeks to identify whether a model has been trained on specific data collections, thus protecting copyright holders while preserving data utility. Existing techniques either *watermark* datasets to embed distinctive behaviors, or directly *infer* usage from discrepancies in model outputs between seen and unseen samples. These approaches exploit the fundamental problem of empirical risk minimization to overfit to seen features. Hence, provenance signals are considered inherently hard to erase, while the adversary's perspective remains largely overlooked, limiting our ability to assess reliability in real-world scenarios. In this work, we present a unified framework that interprets both watermarking and inference-based provenance as manifestations of output divergence, modeling the interaction between auditor and adversary as a min-max game over such divergences. This perspective motivates DivMin, a simple yet effective learning strategy that minimizes the relevant divergence to suppress provenance cues. Experiments across diverse image datasets demonstrate that, starting from a pretrained vision-language model, DivMin retains over 93% of the full fine-tuning performance gain relative to a zero-shot baseline, while evading all six state-of-the-art auditing methods. Our findings establish divergence minimization as a direct and practical path to obfuscating provenance, offering a realistic simulation of potential adversary strategies to guide the development of more robust auditing techniques. Code and Appendix will be available at https://github.com/GradOpt/DivMin.

1 Introduction

Modern deep learning is largely defined by a core paradigm: massive, overparameterized models first learn general representations from web-scale data, before adaptation on curated datasets for specialized downstream tasks [25]. The relentless expansion of data has consistently fueled paradigm shifts, from early breakthroughs like the ImageNet visual recognition challenge [7] to recent advances in cross-modal alignment exemplified by CLIP [40] and GPT-4 [2]. The central role of data, empirically reinforced by scaling laws [21], underscores its indisputable significance for the future progress of deep learning and broader scientific discovery [15].

Yet beyond laboratory settings, real-world model development raises pressing ethical and legal concerns over data transparency. Proprietary datasets and personal information are vulnerable to leakage [46, 14], often ending up on illicit markets [49]. Such data fuels unauthorized model training, resulting in public scandals from privacy violations [45] to political manipulation [54].

Despite growing attention, preventing unauthorized training remains fundamentally challenging. Privacy-Preserving Machine Learning (PPML) [35] enables training on encrypted data to prevent leakage but falls short when data must remain publicly accessible—e.g., open-source datasets restricted to academic usage or personal images shared online. Unlearnable Examples [18, 41] perturb protected data to obstruct model learning, yet render it unusable for *any* training, undermining legitimate use by authorized parties.

Since directly blocking unauthorized training is often infeasible, a more practical solution lies in **dataset provenance**—enabling auditors to detect whether a protected dataset was used in training a suspect model via black-box queries and confidence responses. *Dataset watermarks* embed subtle triggers into a small subset of protected samples and detect misuse through behavioral shifts in suspicious models—e.g., trigger-induced misclassification via backdoor watermarks. *Fingerprint inference*, in contrast, preserves dataset integrity and identifies trained models by analyzing output distributions, applying hypothesis tests to reveal nontrivial generalization gaps between protected and hold-out data from the same distribution.

Dataset provenance capitalizes on the fundamental flaw of models trained via empirical risk minimization (ERM): their propensity to overfit, manifesting either as shortcut prediction on spurious features (e.g., watermarks) or as a discernible generalization gap between seen and unseen inputs. Such underlying mechanisms have fostered the belief that provenance is inherently robust, as unauthorized training invariably leaves identifiable traces.

Evaluations of this presumed robustness, however, are often superficial. They typically rely on generic countermeasures: *regularizations* to reduce memorization; *input perturbations* to disrupt watermark patterns; or standard *backdoor defenses*. Crucially, to our knowledge, no systematic investigation has been explored from the perspective of a dedicated adversary. The genuine resilience of provenance against adaptive threats remains largely speculative, hindering fair comparisons and further progress. This raises critical questions: *Can a sophisticated adversary nonetheless devise strategies to neutralize provenance signals? And if so, what principles should guide the design of more robust auditing techniques?*

To investigate these questions, we establish that provenance signals, whether from watermarks or fingerprints, ultimately stem from *divergence in the model's output distributions*. Such divergence arises either from learning watermark patterns—causing discrepancies between clean and triggered inputs—or from overfitting, leading

* Corresponding Author. Email: wsl@sjtu.edu.cn.

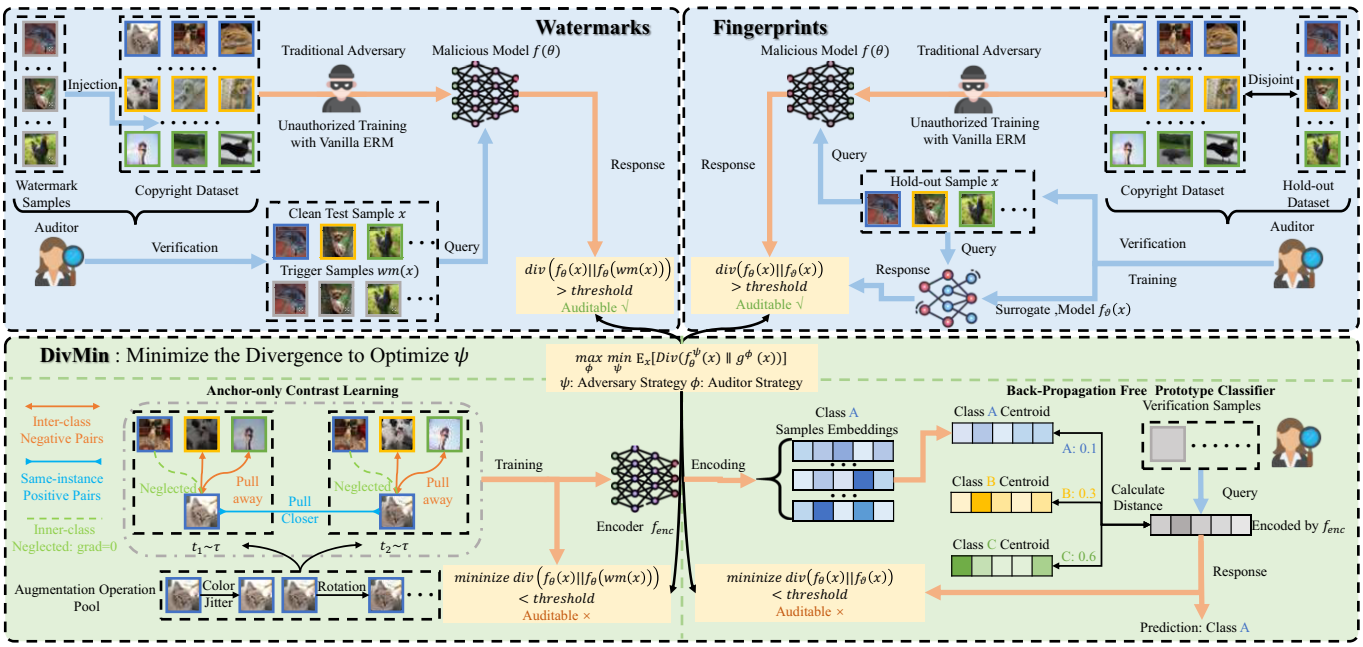


Figure 1. **Top**: the auditor–adversary min-max game over output divergence, instantiated with both watermarks and fingerprints. **Bottom**: our DivMin attack minimizes this divergence to suppress provenance signals, with anchor-only contrastive learning and a backpropagation-free prototype classifier.

to distinct behaviors on seen versus unseen data. We thus conceptualize the competition between auditor and adversary as a min-max game over this divergence: the adversary minimizes it to suppress provenance signals, while the auditor maximizes it to elicit informative evidence from the suspect model. This perspective furnishes a principled framework to systematically analyze the auditor and adversary strategies, moving beyond heuristic evaluations.

Building on this framework, we introduce the **DivMin attack** to instantiate the adversary's inner minimization strategy. DivMin decouples the model into a feature encoder and a label mapping. For the encoder, we design an Anchor-Only Contrastive Learning (AoCL) objective that pulls together different views of the same instance while pushing apart samples from different classes. Crucially, AoCL's meticulous design ensures non-paired, intra-class samples exert no gradient, allowing the encoder to learn task-relevant representations while suppressing sensitivity to subtle watermark transformations. On top of this robust encoder, we construct a backpropagation-free prototype classifier that performs stable classification under training sample variations, reducing output divergence between seen and unseen data.

Starting from a vision-language model (e.g., CLIP) on diverse downstream datasets, DivMin requires neither auxiliary data nor assumptions about the provenance scheme. It evades a broad suite of state-of-the-art (SOTA) watermarks and fingerprints while retaining over 93% of the task performance gain from full fine-tuning over the zero-shot baseline. DivMin thus provides a critical reference for analyzing adversary strategies and optimizing data provenance. Furthermore, as a preliminary study in Appendix C shows, solving the auditor's outer maximization yields an effective adaptive defense, corroborating the practical significance of our min-max framework. Our auditor-adversary game framework and the DivMin attack are illustrated in Figure 1, and main contributions are summarized as follows:

- We present the first systematic analysis of data provenance from the adversary perspective, yielding key insights and crucial guidance for improving provenance techniques.
- We formulate the auditor-adversary competition as a min-max game over output divergence, providing a principled framework to understanding adversarial strategies.
- We propose the DivMin attack, featuring Anchor-Only Contrastive Learning and a backpropagation-free prototype classifier, to solve the inner adversary problem.
- Extensive experiments show that DivMin achieves a remarkable trade-off between task performance and provenance signal suppression; our preliminary adaptive defense further validates the practical value of the min-max formulation.

2 Related Work

2.1 Data Provenance against Unauthorized Training

Inspired by digital watermarking used to protect multimedia content from duplication, remixing, or exploitation [34], dataset watermarks embed carefully crafted triggers into protected datasets to elicit identifiable behaviors in suspicious models, providing evidence of unauthorized training. The central challenge is to balance stealth against data sanitization, performance degradation on authorized models, and reliable identification of unauthorized ones. Early approaches leveraged well-studied *backdoors* [28, 27], measuring induced misclassification rates as provenance signals. However, such artificial shortcuts often impair authorized models, notably undermining adversarial robustness [47, 59]. To mitigate such drawbacks, recent work has introduced *benign watermarks* that avoid incorrect predictions. They shift selected samples into hard-to-generalize domains [12], specific color spaces [62], or apply constrained perturbations [19, 3]. Provenance is then assessed via hypothesis testing whether suspect models show elevated confidence on watermarked inputs.

Another line of research seeks to preserve dataset integrity by avoiding any data modification, instead characterizing the behavioral fingerprints of models trained on the protected dataset. Inspired by

membership inference [44], these fingerprints estimate per-sample membership by exploiting the model's differential responses to seen versus unseen data, then aggregate results across multiple queries to statistically test for unauthorized training [33, 30]. Despite their non-intrusive nature, fingerprints are often criticized for high false positive rates [29, 47, 48]. Beyond image classification, recent work has explored scaling data provenance to domains such as large language models [43], though these efforts remain preliminary and have yet to yield reliable results under rigorous evaluation [8, 61].

2.2 Robustness Evaluation in Dataset Provenance

Despite rapid advances in dataset provenance, their robustness under adversarial settings has received limited attention. These methods exploit fundamental limitations of ERM—overfitting to spurious correlations [9] and memorizing seen data [50]—and are often perceived as inherently difficult to suppress [33, 53]. However, to the best of our knowledge, no systematic study has been conducted on their robustness, leaving the reliability in real-world scenarios uncertain. Existing work has only heuristically adapted known techniques, falling into three main categories: (1) *regularizations* such as label smoothing [36] and differential privacy [1] to reduce overfitting [62, 30]; (2) input perturbations such as Gaussian blur and adversarial training [32] to obscure watermark features [62, 53]; and (3) classical backdoor defenses [27, 3] including fine-pruning [31] and data sanitization [13]. Yet, most of these methods prove ineffective against SOTA provenance techniques. Following the acceptance of this paper, concurrent work on adaptive attacks against dataset auditing emerged [60, 42], which we discuss in Appendix D.

3 Preliminary Study

3.1 A Unified Perspective of Dataset Provenance

Consider a protected dataset $\mathcal{D} \subset \mathcal{X}$ and a suspicious model $f_\theta : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$. The auditor A aims to determine, under black-box access, whether f_θ was trained on $\mathcal{D}$. To this end, A issues a set of queries $Q = \{q_j\}_{j=1}^m \subset \mathcal{X}$ and collects the corresponding responses $E_j = f_\theta(q_j)$. The aggregated evidence $E = (E_1, \ldots, E_m)$ is then used to infer a binary random variable $Z \in \{0, 1\}$, where $Z = 1$ indicates that f_θ was trained on $\mathcal{D}$, and $Z = 0$ otherwise.

In *dataset watermarking*, the auditor A designs a transformation function $\tau(x)$ (e.g., $\tau(x) = x + \delta$, where $\delta \in \mathcal{X}$ is a predefined trigger), and applies it to a small subset of $\mathcal{D}$. During verification, A checks whether $f_\theta(\tau(x))$ exhibits the intended behavior: for instance, in backdoor watermarking, the Verification Success Rate (VSR) serves as a proxy for the confidence in $Z = 1$. In *dataset fingerprinting*, auditor A samples a hold-out set $\mathcal{D}'$, drawn independently from the same distribution as $\mathcal{D}$ but disjoint from it. A calibration model $f_\varnothing$ is trained on $\mathcal{D}'$, and a one-sided t-test is performed to assess whether f_θ yields higher confidence on $\mathcal{D}$ than $f_\varnothing$. A statistically significant difference supports the hypothesis $Z = 1$.

Under the above setting, we introduce a unified perspective to quantify the effectiveness of dataset provenance. Given a sample-level divergence measure $\mathrm{Div}(\cdot \,\|\, \cdot)$, e.g., the α-order f-divergence, the *informativeness* of each query x is subject to:

$$S(x; \tau) = \mathrm{Div}\left(f_\theta(x) \,\|\, g(x; \tau)\right) \tag{1}$$

where the reference distribution $g(x; \tau)$ is given by:

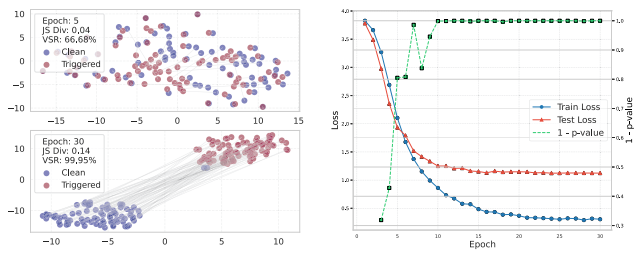
$$g(x; \tau) = \begin{cases} f_\theta(\tau(x)), & \text{for watermarks} \\ f_\varnothing(x), & \text{for fingerprints.} \end{cases} \tag{2}$$

Intuitively, larger values of $S(x; \tau)$ indicate stronger evidence that f_θ was trained on $\mathcal{D}$. In this view, the information about the provenance variable Z from evidence E is governed by $S(x; \tau)$, with the mutual information satisfying $I(Z; E) \propto \mathbb{E}_{x \sim Q}\left[S(x; \tau)\right]$. The following proposition (proved in Appendix A.1) provides an upper bound on this mutual information:

Proposition 1 (Upper Bound on Mutual Information)**.** *Let $Z \in \{0, 1\}$ be a binary random variable with uniform prior. Assume that, conditioned on $Z = z$, the responses $(E_1, \ldots, E_m)$ are independent with marginal distributions $P_{z,j}$, and the joint distribution factorizes as $P_z = \bigotimes_{j=1}^m P_{z,j}$. Then there exists a constant $c > 0$ such that*

$$\boxed{I(Z; E) \;\le\; c \cdot \bar{S}} \quad \textit{with} \quad \bar{S} = \mathbb{E}_{x \sim Q}\left[S(x; \tau)\right]. \tag{3}$$

To empirically validate the above analysis, Figure 2 illustrates how increasing divergence correlates with stronger verification signals in both watermarks and fingerprints. Figure 2(a) shows the divergence $\mathrm{Div}(f_\theta(x), f_\theta(\tau(x)))$ and the corresponding VSR at epochs 5 and 30, during full fine-tuning on the DTD dataset embedded with BadNets watermark. As training progresses, the representations of x and $\tau(x)$ drift apart, resulting in a clear increase in divergence—and consequently, a higher VSR. In Figure 2(b), f_θ is fine-tuned on the clean DTD dataset. The divergence between losses on training and hold-out samples steadily grows, while the p-value in fingerprint-based inference decreases accordingly. Once again, the rising divergence strengthens the evidence in favor of $Z = 1$.



(a) Representation shift in BadNets model's embedding space (b) Loss curves and detection confidence (1 - p-value) of fingerprints

Figure 2. Divergence as the source of provenance signals.

3.2 The Min-Max Game over Output Divergence

Building upon the analysis, we formulate the competition between the auditor and the adversary as a zero-sum game. The adversary aims to suppress provenance signals by minimizing divergence, while the auditor seeks to recover informative evidence by maximizing it. This interplay is captured by the following min-max objective:

$$\max_{\phi} \min_{\psi} \; \mathbb{E}_{x \sim Q} \; \mathrm{Div}\left(f_\theta^{\psi}(x) \,\|\, g^{\phi}(x; \tau)\right) \tag{4}$$

where ψ denotes the adversary's strategy for altering the model f_θ, and ϕ denotes the auditor's strategy for eliciting reliable evidence via query design or reference modeling.

Prior work has focused on optimizing the auditor's strategy ϕ, leaving the adversary's side ψ largely underexplored. Analogous to adversarial training, meaningful robustness against adaptive threats can only emerge when both players are actively engaged, enabling the system to approach a potential Nash equilibrium. In this work, we take a first step toward this goal by introducing the DivMin attack to solve the inner minimization problem, thereby providing a solid starting point for modeling the adversary's behavior.

4 The DivMin Attack

4.1 Overview

Since the divergence $\bar{S}$ upper-bounds the provenance information $I(Z;E)$ available to the auditor, minimizing $S(x;\tau)$ during training emerges as a natural strategy to suppress verification cues. However, such divergence reflects the fundamental limitation of empirical risk minimization (ERM) in statistical learning: models tend to memorize spurious features [9]—including those injected by watermarks—that serve as predictive shortcuts, leading to output discrepancies between clean and trigger samples. Likewise, fingerprints exploit the well-known gap between memorization and generalization [50, 55], wherein models consistently assign higher confidence to training data than to unseen samples. Therefore, without deviating from the ERM paradigm, simultaneously mitigating both watermark and fingerprint signals has long been considered impractical [33, 53], rendering data provenance ostensibly secure by design.

However, we propose a simple training framework, DivMin, that effectively suppresses provenance signals without requiring additional clean data or prior knowledge of the auditing method. By *explicitly minimizing divergence during training*, DivMin reduces the informativeness of queries and pushes the model into an ambiguous regime for provenance inference. We decouple the model f_θ into a feature encoder f_{enc} and a linear projector f_w. To mitigate divergence caused by *spurious features* (e.g., watermarks), we introduce an *anchor-only contrastive loss* that encourages f_{enc} to learn task-relevant representations while minimizing $\mathrm{Div}(f_\theta(x), f_\theta(\tau(x)))$. To suppress divergence stemming from *overfitting* to training data, we forgo gradient-based optimization of f_w and instead construct a *prototype classifier* that maps features to labels by aggregating class-wise centroids, thus limiting the influence of individual training samples and reducing $\mathrm{Div}(f_\theta(x), f_\varnothing(x))$. We next elaborate on the design of these two components.

4.2 Anchor-Only Contrastive Learning

The objective of this part is to adapt the feature encoder f_{enc} to the target task by learning meaningful representations from domain-specific inputs, while minimizing $\mathrm{Div}(f_\theta(x), f_\theta(\tau(x)))$ to suppress watermarks. Although the specific auditing scheme is unknown, effective verification requires the watermark distribution $P(\tau(X), Y)$ to be sufficiently distinct or orthogonal to the task distribution $P(X, Y)$, ensuring trigger specificity and avoiding false positives. As a result, watermark features are typically uninformative for the primary task. This motivates a learning objective that discards task-irrelevant signals while preserving only features essential for downstream performance. This principle naturally aligns with Contrastive Predictive Coding (CPC) [38], which promotes semantic consistency across different views of the same instance to retain task-relevant information while filtering out nuisance variation.

However, existing contrastive objectives based on the InfoNCE framework [38] are not well suited for this purpose. In self-supervised SimCLR [6], each input is augmented into two views forming a positive pair, while all other samples in the batch serve as negatives. This neglects label information, and hard negatives from the same class hinder the model's ability to learn discriminative, task-relevant features. In contrast, the supervised SupCon [22] treats all same-class pairs as positives and all different-class pairs as negatives. While it leverages labels effectively, its learning dynamics resemble those of cross-entropy, encouraging same-class representations to collapse toward the vertices of a regular simplex [10], which can lead to overfitting and memorization of poisoned labels introduced by watermarks.

Building on the minimal design of InfoNCE-based contrastive learning, we propose an *Anchor-Only Contrastive Loss (AoCL)* that leverages label information to learn task-relevant representations while suppressing provenance signals. Given a mini-batch of B samples, two independent random augmentations $t \sim \mathcal{T}$ are applied to each input, yielding $2B$ augmented views. Let $i \in \mathcal{I} = \{1, \ldots, 2B\}$ index an arbitrary view, and let $p(i)$ denote the index of the paired view originating from the same input. The corresponding normalized embedding is given by $\mathbf{z}_i = f_{\text{enc}}(t(x_i))/\|f_{\text{enc}}(t(x_i))\| \in \mathbb{R}^D$, with class label $y_i \in \{1, \ldots, C\}$. For each *anchor* i, AoCL selects the paired view $p(i)$ as the sole *positive*, treats all samples from different classes as *negatives*, and explicitly *neglects* other same-class samples. Formally, we define: the positive set $\mathcal{P}(i) = \{p(i)\}$, the ignored inner-class set $\mathcal{I}(i) = \{j \neq i \mid y_j = y_i,\ j \notin \mathcal{P}(i)\}$, and the negative set $\mathcal{D}(i) = \mathcal{I} \setminus (\{i\} \cup \mathcal{I}(i))$. The Anchor-Only Contrastive Loss is then defined as:

$$\mathcal{L}_{\text{AoCL}} = -\frac{1}{2B} \sum_{i=1}^{2B} \log \frac{\exp\left(\mathbf{z}_i^\top \mathbf{z}_{p(i)}/\tau\right)}{\sum_{j \in \mathcal{D}(i)} \exp\left(\mathbf{z}_i^\top \mathbf{z}_j/\tau\right)}, \tag{5}$$

where $\tau > 0$ is a temperature parameter.

Intuitively, minimizing $\mathcal{L}_{\text{AoCL}}$ encourages the encoder f_{enc} to become invariant to random transformations $t \sim \mathcal{T}$, reducing its sensitivity to low-level input perturbations. In fact, optimizing $\mathcal{L}_{\text{AoCL}}$ implicitly maximizes the expected feature similarity between an input and its transformed view:

$$\mathbb{E}_{x,\, t \sim \mathcal{T}} \left[f_{\text{enc}}(x)^\top f_{\text{enc}}(t(x)) \right], \tag{6}$$

which in turn reduces the output discrepancy between a clean sample and its augmented counterpart. We provide a formal derivation of this connection in Appendix A.2. This view-consistency objective plays a critical role in suppressing watermark signals. Since watermark triggers are designed to evade both human scrutiny and automated sanitization, the trigger function $\tau(x)$ typically introduces imperceptible modifications to the input. Such subtle patterns are inherently brittle and can be disrupted by standard data transformations [58]. Thus, maximizing the invariance of f_{enc} under stochastic augmentations effectively reduces the divergence $\mathrm{Div}(f_\theta(x), f_\theta(\tau(x)))$, without requiring any prior knowledge of the watermark generation process.

In practice, we instantiate the transformation pool $\mathcal{T}$ via TrivialAugment [37], which applies a diverse set of randomized operations, including spatial warping, color shifts, geometric distortion, and additive noise. Empirically, we observe this strategy consistently reduces the divergence signal available for watermark verification.

Another nice property of AoCL lies in its treatment of label information. By exclusively using samples from different classes as negatives, AoCL avoids pushing apart inner-class representations when watermarked samples constitute a small portion of the dataset. Moreover, for each anchor i, all inner-class views except the paired positive (i.e., elements in $\mathcal{I}(i)$) are explicitly neglected and contribute *zero gradient* to the loss. This conservative contrastive signal prevents the encoder from over-collapsing inner-class representations, which could otherwise lead to overfitting or unintended memorization of watermark artifacts as class-specific features. As demonstrated in Section 5, AoCL strikes a favorable balance between learning discriminative, task-relevant features and suppressing provenance signals, consistently outperforming widely used contrastive losses such as SimCLR and SupCon.

4.3 Backpropagation-Free Prototype Classifier

Once the encoder f_{enc} has been trained with AoCL to extract task-relevant representations, a mapping function is required to perform classification on top of these features. However, training a classifier via ERM with gradient-based optimization is inherently prone to overfitting [50, 57], often resulting in substantial output discrepancies between seen and unseen samples [44], which can leak provenance information to inference-based auditing methods. To mitigate this, we introduce a simple yet effective *prototype classifier* that requires only a single forward pass and avoids any gradient-based parameter updates. This classifier leverages feature centroids to map inputs to class labels, limiting the influence of individual training samples and reducing generalization gaps.

Formally, given an auditing dataset $\{(x_i, y_i)\}_{i=1}^N$ with $y_i \in \{1, \dots, C\}$, and a feature extractor f_{enc} trained via AoCL, the class prototype $\boldsymbol{\mu}_c$ for each class c is computed by averaging the normalized feature vectors of all training samples in that class:

$$\mathbf{z}_i = f_{\text{enc}}(x_i), \quad \boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{i:y_i=c} \mathbf{z}_i, \tag{7}$$

where N_c denotes the number of samples in class c.

For any test input x, we compute its cosine similarity to each class prototype and assign it to the most similar one. This is equivalent to selecting the prototype with the least cosine distance:

$$\hat{y} = \arg\min_{c\in\{1,\dots,C\}} d_c = \arg\max_{c\in\{1,\dots,C\}} \hat{\mathbf{z}}^\top \hat{\boldsymbol{\mu}}_c, \tag{8}$$

where $d_c = 1 - \hat{\mathbf{z}}^\top \hat{\boldsymbol{\mu}}_c$, and $\hat{\mathbf{z}}, \hat{\boldsymbol{\mu}}_c$ denote the ℓ_2-normalized test feature and class prototype, respectively.

The prototype classifier is functionally equivalent to a linear mapping: the prediction score for class c is computed as $f(\mathbf{z})_c = \mathbf{w}_c^\top \mathbf{z}$, where the weight $\mathbf{w}_c$ is given by the centroid of training features in class c, rather than learned via backpropagation. This non-parametric construction inherently limits the influence of any individual samples, improving the model's uniform stability [24, 4]. The following result formalizes this effect (see Appendix A.3 for proof):

Proposition 2 (Output Stability of Prototype Classifier)**.** *Let the training set consist of normalized features $D = \{(\mathbf{z}_i, y_i)\}_{i=1}^N$, where $\|\mathbf{z}_i\| \le 1$ and $y_i \in \{1, \dots, C\}$. For each class c, let $\boldsymbol{\mu}_c$ be the prototype with norm $\lambda_c = \|\boldsymbol{\mu}_c\| > 0$, and let $\hat{\boldsymbol{\mu}}_c = \boldsymbol{\mu}_c/\lambda_c$ be its normalized direction. For any normalized test input $\hat{\mathbf{z}}$, if D' differs from D by a single sample in class $c^\star$, then the output variation is bounded by*

$$\boxed{\Delta(\mathbf{z}) := \left|\hat{\mathbf{z}}^\top \hat{\boldsymbol{\mu}}_{c^\star} - \hat{\mathbf{z}}^\top \hat{\boldsymbol{\mu}}'_{c^\star}\right| \le \frac{4}{\lambda_{c^\star} N_{c^\star}}} \tag{9}$$

This bound reveals that the output sensitivity to any single training point decays inversely with class size. As a result, the membership advantage [55] exploited by inference-based auditors vanishes as $N_{c^\star}$ grows, reducing the divergence $\text{Div}(f_\theta(x), f_\varnothing(x))$ and weakening the strength of provenance signals.

Note that Proposition 2 implicitly assumes that the normalized training and test features $\mathbf{z}$ are independently and identically distributed. While this assumption may be violated—since the f_{enc} used to compute prototypes is trained on the same data—Section 5.3 and 5.4.1 shows that AoCL avoids strong same-class attraction, yielding a well-structured feature space that supports stable classification without overfitting to spurious patterns.

Despite its simplicity, our vanilla DivMin already strikes a favorable balance between task performance and provenance suppression. Building on this foundation, DivMin admits various extensions—e.g., learnable perturbations in AoCL to capture watermark patterns [23], or noise injection in prototypes to satisfy notions like differential privacy [52]. We leave these directions to future work.

5 Experiments

5.1 Experimental Settings

Datasets and Models. We evaluate DivMin on four high-resolution, fine-grained image classification datasets: Caltech-101 (101 natural object classes), DTD (47 texture classes), EuroSAT (10 land cover classes), and FGVCAircraft (100 aircraft classes). While Caltech-101 represents generic natural image tasks, the latter three are domain-specific datasets that require dedicated data collection and annotation, closely reflecting real-world model development scenarios. Our primary experiments begin with a lightweight CLIP [51], with its zero-shot classification performance as the baseline, enabling a clear evaluation of how DivMin extracts knowledge from protected datasets while evading detection. In Section 5.4.4, we further evaluate the image-only self-supervised DINOv2 [39].

Data Provenance Methods. We benchmark DivMin against SOTA methods from three major categories of dataset provenance: *Backdoor watermarks*, including the seminal BadNets [11], and stealthier UBW [27] with dynamic target classes; *Benign watermarks*, including MLAuditor [19] and Data Taggants [3], offering superior stealth, minimal interference, and robustness; *Fingerprint inference*, including Dataset Inference [33] and MeFA [30], which aggregate per-sample membership estimates for set-level hypothesis testing. For backdoor watermarks, we report the Verification Success Rate (VSR)—the misclassification rate on trigger-injected test samples. A higher VSR indicates stronger evidence of unauthorized training, with VSR $> 20\%$ typically used as the detection threshold. For benign watermarks and fingerprint inference, we use the p-value from hypothesis testing, with $p < 0.01$ indicating strong rejection of the null hypothesis that no unauthorized use has occurred.

Attack Baselines. Since DivMin builds upon a pretrained vision-language model, we first assess whether standard parameter-efficient fine-tuning (PEFT) can sufficiently learn task-specific knowledge without triggering detection. These include classic techniques such as linear probing, fully-connected (MLP) probing [39], LoRA adaptation [17], and the SOTA visual prompting LoR-VP [20]. We then consider robustness evaluation techniques commonly employed in provenance literature, categorized into three groups: *Regularization*, including ℓ_2 regularizers, label smoothing, and a SOTA implementation of DP-SGD [5]; *Input perturbations*, such as Gaussian blur and adversarial training; SOTA *Backdoor defenses*, including IBD-PSC [16], which filters trigger samples at inference time, and AIBD [56], which sanitizes the training data.

5.2 Evading SOTA Provenance Techniques

Table 1 presents evaluations of a pretrained CLIP model under three settings—zero-shot classification, full finetuning (FF), and our proposed DivMin—across four datasets and against all SOTA provenance methods. Zero-shot classification, which does not adapt to any protected data, is never flagged by provenance method. However, it relies solely on pretrained knowledge, leading to suboptimal performance, especially on domain-specific datasets like EuroSAT and

Table 1. Performance of Zero-Shot, Full Finetuning (FF), and DivMin across datasets. Red indicates detection, while green highlights successful evasion.

Dataset	Method	Badnets		UBW		MLAuditor		Data Taggants		Dataset Inference		MeFA	
		ACC (↑)	VSR (↓)	ACC (↑)	VSR (↓)	ACC (↑)	p-value (↑)	ACC (↑)	p-value (↑)	ACC (↑)	p-value (↑)	ACC (↑)	p-value (↑)
Caltech101	Zero-Shot	91.30	0.00	91.30	1.07	91.30	4.67E-01	91.30	8.62E-01	91.30	6.14E-01	91.30	1.00E+00
	FF	95.97±0.56	97.03±1.22	96.37±0.78	98.03±1.28	95.51±1.02	1.32E-03	95.79±0.58	3.22E-08	94.76±0.53	2.44E-04	94.76±0.53	4.23E-04
	DivMin	96.31±0.38	0.06±0.03	95.74±0.35	0.73±0.11	96.31±0.56	6.99E-01	96.03±0.26	6.70E-01	96.37±0.19	3.38E-01	96.37±0.19	1.00E+00
DTD	Zero-Shot	54.84	3.42	54.84	3.79	54.84	5.25E-01	54.84	7.02E-01	54.84	4.51E-01	54.84	1.00E+00
	FF	66.17±1.45	99.95±0.00	66.60±1.93	71.96±5.44	70.96±0.90	1.30E-05	71.01±0.33	4.25E-08	72.61±0.78	3.31E-08	72.61±0.78	9.05E-08
	DivMin	69.47±0.13	3.97±0.08	71.06±0.40	2.77±0.53	70.90±0.28	5.23E-01	71.86±0.28	4.83E-01	71.38±0.15	2.32E-01	71.38±0.25	9.72E-01
EuroSAT	Zero-Shot	45.22	0.10	45.22	11.81	45.22	1.00E+00	45.22	8.23E-01	45.22	3.53E-01	45.22	1.00E+00
	FF	98.85±0.45	100.00±0.00	98.50±0.36	99.98±0.00	98.83±0.78	3.70E-03	98.85±0.45	9.47E-03	98.76±0.71	1.01E-03	98.76±0.71	2.25E-03
	DivMin	95.69±0.48	4.97±0.35	95.30±0.22	4.57±0.15	95.35±0.49	9.49E-01	96.96±0.29	3.24E-01	95.65±0.45	2.28E-01	95.65±0.45	9.56E-01
FGVCAircraft	Zero-Shot	21.09	4.06	21.09	8.43	21.09	8.47E-01	21.09	9.98E-01	21.09	6.62E-01	21.09	1.00E+00
	FF	43.80±3.59	100.00±0.00	43.29±5.53	99.03±0.04	46.47±5.83	1.22E-11	43.92±4.67	5.63E-08	46.08±3.89	4.17E-08	46.08±3.89	1.76E-03
	DivMin	36.36±1.13	5.97±0.11	35.73±1.29	6.88±0.04	36.27±1.41	7.28E-01	35.91±1.46	2.37E-01	36.87±0.71	4.46E-01	36.87±0.71	1.00E+00

FGVCAircraft. As such, it serves as an *evasion oracle*—guaranteed to evade detection, but offering a lower bound on task performance. In contrast, full finetuning significantly boosts accuracy by exploiting rich task-specific information from the protected data. Yet, it is consistently detected by all provenance methods, due to the vanilla ERM's tendency to learn shortcut features and memorize training data. This setting thus serves as a *task oracle*—achieving maximal task performance at the cost of complete visibility to auditing.

Remarkably, DivMin achieves a favorable trade-off between utility and stealth, despite also updating all model parameters. Across all datasets, it substantially outperforms the zero-shot baseline—nearly doubling accuracy on domain-specific datasets like EuroSAT and FGVCAircraft—while remaining undetected. On average, DivMin recovers over 93% of the accuracy gain achieved by full finetuning relative to zero-shot. Its *anchor-only contrastive learning* and *backpropagation-free prototype classifier* effectively suppress the divergence signals exploited by watermarks and fingerprints. As a result, DivMin consistently evades detection, often yielding verification scores close to those of zero-shot and well below decision thresholds. These results demonstrate that DivMin can successfully extract downstream knowledge from protected datasets without being flagged, providing strong empirical evidence for divergence minimization as a viable and versatile adversarial strategy.

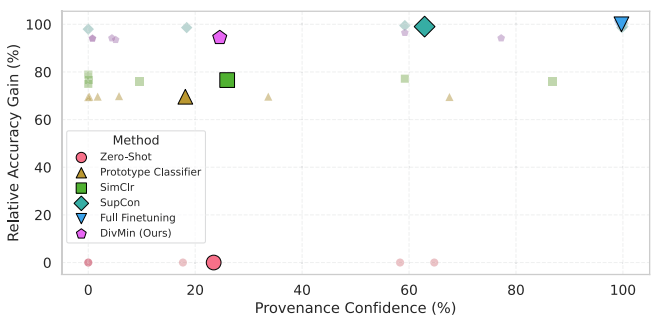


Figure 3. Relative performance gains (↑) and provenance confidence (↓) of different design choices in DivMin.

5.3 Comparison with Baseline Attacks

Table 2 compares DivMin with existing attack strategies on DTD, evaluating both task accuracy and provenance suppression. PEFT methods are less prone to fingerprint inference due to reduced overfitting. However, they consistently fall into shortcuts introduced by watermark patterns, such as BadNets and Data Taggants. Notably, LoR-VP introduces learnable visual prompts in the input space, which partially disrupt watermarks but fails to achieve complete evasion. Moreover, due to limited parameter updates, PEFT methods also underperform DivMin in task accuracy.

Regularizations such as ℓ_2 regularizer, label smoothing, and DP-SGD can weaken provenance signals but never fully suppress any watermark or fingerprint. Input perturbations like Gaussian blur and adversarial training occasionally affect invisible watermarks, yet are ineffective against poison-label backdoors and fingerprints. Advanced backdoor defenses scale poorly to vision-language models. They cannot reliably eliminate shortcuts and are entirely ineffective against methods do not induce misclassification. In contrast, DivMin yields nontrivial success by explicitly minimizing output divergence, providing a principled attack strategy.

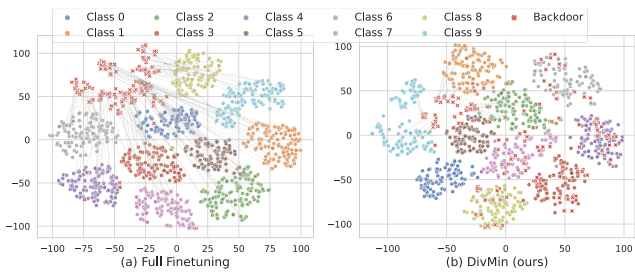


Figure 4. Features from BadNets models. Clean and trigger samples are connected by dashed lines. Output JS Div: 0.174 (FF), 0.069 (DivMin).

5.4 Ablations and Discussions

5.4.1 Contribution of Components in DivMin

We analyze the individual contributions of *anchor-only contrastive learning* (AoCL) and the *prototype classifier*, focusing on both task performance and provenance signal suppression. We also compare AoCL with other contrastive learning objectives, including the self-supervised SimCLR and the supervised SupCon, as shown in Figure 3, with detailed results in Appendix B. The prototype classifier alone incurs minimal detection by dataset provenance but offers limited accuracy gains over the zero-shot baseline. AoCL further improves task performance—approaching that of full finetuning—while inducing negligible risk of detection.

Compared to AoCL, SimCLR yields significantly lower performance gains; for instance, AoCL achieves nearly 10% higher absolute accuracy on EuroSAT. SupCon, by contrast, overfits to backdoor shortcuts and is susceptible to fingerprints. These results highlight the effectiveness of our tailored AoCL and prototype classifier, which together strike a favorable balance between learning task-relevant features and reducing divergence. In contrast, standard contrastive losses fail to replicate AoCL's role within the DivMin attack.

Table 2. Comparison of DivMin with baseline attacks on the DTD dataset. **Bold** marks the best result, and underline indicates the second best.

Method	Badnets		UBW		MLAuditor		Data Taggants		Dataset Inference		MeFA	
	ACC (↑)	VSR (↓)	ACC (↑)	VSR(↓)	ACC (↑)	p-value (↑)	ACC (↑)	p-value (↑)	ACC (↑)	p-value (↑)	ACC (↑)	p-value (↑)
Linear Prob	66.41±4.24	41.58±1.83	68.62±3.76	3.85±0.43	68.57±0.25	4.13E-01	68.66±0.40	7.29E-04	68.78±1.81	2.85E-02	68.78±1.81	4.60E-01
FC Prob	67.71±2.71	60.16±1.93	69.73±1.91	4.84±0.30	69.20±1.28	4.48E-01	69.31±1.03	1.81E-04	69.15±0.90	1.08E-02	69.15±0.90	2.42E-01
LoRA	68.72±1.76	99.41±0.18	70.05±0.93	50.06±5.87	69.57±1.13	3.03E-04	71.12±1.40	8.00E-06	69.26±1.78	1.42E-03	69.26±1.78	1.36E-01
LOR-VP	66.81±4.69	31.70±2.73	**71.06±2.23**	3.67±0.73	69.41±1.33	1.83E-01	69.36±2.08	7.29E-04	70.37±0.43	2.52E-02	70.37±0.43	3.09E-01
L2 Reg	67.29±1.91	98.56±0.33	66.91±2.83	67.33±3.89	68.72±2.83	2.19E-03	69.41±1.18	2.00E-06	70.27±1.65	1.93E-04	70.27±1.65	2.35E-06
LS	66.70±3.91	96.52±0.93	69.79±0.93	64.48±3.74	69.95±1.93	1.74E-05	70.53±1.50	4.07E-05	**72.93±1.91**	2.41E-06	**72.93±1.91**	3.07E-11
Gauss Blur	63.72±5.42	86.68±1.86	62.13±0.23	45.38±8.43	65.80±0.95	9.43E-03	64.31±1.15	4.07E-05	65.42±1.93	9.41E-08	65.41±1.93	6.29E-06
AT	58.19±6.34	100.00±0.00	56.44±7.30	59.47±12.41	58.46±6.39	3.09E-01	59.15±7.47	1.81E-04	58.30±4.44	6.76E-06	58.30±4.44	1.05E-02
DP-SGD	65.11±5.84	100.00±0.00	65.85±3.91	72.62±1.33	68.72±1.86	2.89E-03	69.84±1.91	2.00E-06	68.24±0.88	1.49E-06	68.24±0.88	4.37E-04
IBD-PSC	59.10±2.86	16.49±4.31	56.17±2.38	36.36±2.31	54.68±1.91	2.60E-05	54.14±2.46	6.29E-08	56.97±1.40	6.95E-06	56.97±1.40	8.34E-07
AIBD	66.06±3.39	31.65±2.78	64.79±1.91	22.25±3.28	64.89±2.86	2.40E-05	65.27±4.39	8.00E-06	62.29±2.16	3.66E-04	62.29±2.16	1.61E-05
DivMin (Ours)	**69.47±0.13**	**3.97±0.08**	**71.06±0.40**	**2.77±0.53**	**70.90±0.28**	**5.23E-01**	**71.86±0.28**	**4.83E-01**	71.38±0.15	**2.32E-01**	71.38±0.25	**9.72E-01**

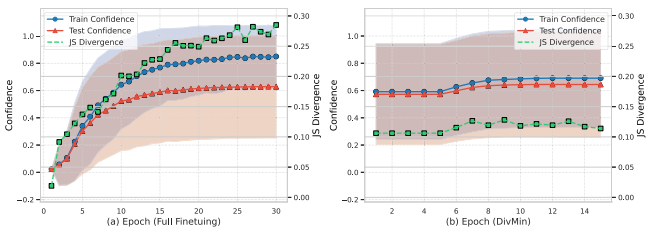


Figure 5. Dynamics of confidence (train/test) and divergence during training of Full Finetuning and DivMin.

5.4.2 *Divergence as the Source of Provenance Signals*

Motivated by the min-max game over divergence, we examine whether DivMin's evasion capability arises from minimizing divergence. We visualize the feature spaces of models trained with full finetuning and DivMin on EuroSAT with BadNets watermarks. As shown in Figure 4, the fully finetuned model exhibits substantial feature shifts when test samples are modified with triggers, indicating high sensitivity to the watermark. In contrast, DivMin shows minimal changes between clean and triggered inputs. This suggests that minimizing $\mathrm{Div}(f_\theta(x), f_\theta(\tau(x)))$ effectively suppresses watermark-based provenance signals. We further track divergence between seen and unseen samples during training. For full finetuning, divergence steadily increases, providing strong fingerprint signals. In contrast, DivMin maintains low and stable divergence, indicating its ability to disrupt fingerprint inference by minimizing $\mathrm{Div}(f_\theta(x), f_\varnothing(x))$.

5.4.3 *Adaptive Defense against DivMin*

Given the theoretical link between AoCL and the InfoNCE framework, we investigate whether CTRL (ICCV 2023) [26], a SOTA backdoor on contrastive learning, can serve as an adaptive defense. CTRL introduces a frequency trigger intended to be robust to data augmentation, encouraging contrastive loss to cluster trigger samples in feature space. However, CTRL proves inconsistent on high-resolution images and pretrained CLIP models. On the relatively vulnerable EuroSAT, CTRL achieves a 58.48% VSR against SimCLR, but only 4.7% against DivMin. We attribute this to AoCL's selective use of negatives, which avoids forcing the model to compress all discriminative features of a single image. While CTRL seeks watermark features resilient under divergence minimization, this strategy does not generalize well. In Appendix C, we adopt the auditor's perspective and explicitly solve the outer maximization problem, demonstrating a promising direction to optimize dataset provenance.

5.4.4 *Generalization Across Backbones*

We evaluate DivMin on DTD using DINOv2 as the backbone. Although the vision-only, self-supervised DINOv2 lacks zero-shot capability, DivMin still achieves competitive task performance while fully evading detection, as shown in Table 3.

Table 3. Full Finetuning vs. DivMin on DTD with DINOv2 backbone.

Method		Full Finetuning	DivMin
Badnets	**ACC**	74.15±1.86	72.98±0.35
	VSR	100.00±0.00	1.20±0.18
UBW	**ACC**	72.87±1.35	73.30±0.40
	VSR	58.03±1.88	2.71±0.20
MLAuditor	**ACC**	76.28±2.31	72.98±0.90
	p-value	1.64E-04	9.44E-01
Taggants	**ACC**	75.27±2.41	73.14±0.85
	p-value	5.90E-11	6.95E-01
DI	**ACC**	76.60±1.91	73.88±0.43
	p-value	2.49E-07	1.46E-01
MeFA	**ACC**	76.60±1.91	73.88±0.43
	p-value	1.27E-16	3.82E-01

6 Conclusion

In this work, we present a systematic robustness evaluation of dataset provenance from the adversary perspective. We show that provenance signals exploited by both watermarks and fingerprints fundamentally arise from divergences in model outputs, and establish the auditor-adversary competition as a min-max game over divergence. Building upon this framework, we propose DivMin, an attack with anchor-only contrastive learning and a prototype classifier, which achieves favorable trade-off between task utility and provenance evasion, providing a principled reference to improve auditing.

Acknowledgements

The work was supported by the National Natural Science Foundation of China under Grant 62271307 and 61771310.

Ethics Statement

Our work aims to advance the robustness and reliability of dataset provenance by offering a principled adversarial perspective. While our proposed method can evade existing auditing tools, it is intended as a diagnostic tool to better understand the auditor-adversary competition and inspire stronger defenses. We also explore adaptive defense strategies in Section 5.4.3 and Appendix C.

References

[1] M. Abadi, A. Chu, et al. Deep learning with differential privacy. In *ACM CCS 2016*, pages 308–318, 2016.

[2] J. Achiam, S. Adler, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] W. Bouaziz, N. Usunier, et al. Data taggants: Dataset ownership verification via harmless targeted data poisoning. In *ICLR*, 2025.

[4] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[5] Z. Bu, Y.-X. Wang, et al. Automatic clipping: Differentially private deep learning made easier and stronger. *NeurIPS*, 36:41727–41764, 2023.

[6] T. Chen, S. Kornblith, et al. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PmLR, 2020.

[7] J. Deng, W. Dong, et al. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[8] M. Duan, A. Suri, et al. Do membership inference attacks work on large language models? In *First Conference on Language Modeling*, 2024.

[9] R. Geirhos, J.-H. Jacobsen, et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[10] F. Graf et al. Dissecting supervised contrastive learning. In *ICML*, pages 3821–3830. PMLR, 2021.

[11] T. Gu, B. Dolan-Gavitt, et al. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

[12] J. Guo, Y. Li, et al. Domain watermark: Effective and harmless dataset copyright protection is closed at hand. *NeurIPS*, 36:54421–54450, 2023.

[13] J. Hayase, W. Kong, et al. Spectre: Defending against backdoor attacks using robust statistics. In *ICML*, pages 4129–4139. PMLR, 2021.

[14] M. Hogan, Y. Michalevsky, et al. Dbreach: Stealing from databases using compression side channels. In *IEEE S&P*, pages 182–198. IEEE, 2023.

[15] N. Hollmann, S. Müller, et al. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

[16] L. Hou, R. Feng, et al. Ibd-psc: input-level backdoor detection via parameter-oriented scaling consistency. In *Proceedings of the 41st ICML*, pages 18992–19022, 2024.

[17] E. J. Hu, yelong shen, et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[18] H. Huang, X. Ma, et al. Unlearnable examples: Making personal data unexploitable. In *ICLR*, 2021.

[19] Z. Huang, N. Z. Gong, et al. A general framework for data-use auditing of ml models. In *ACM CCS 2024*, pages 1300–1314, 2024.

[20] C. Jin, Y. Li, et al. Lor-VP: Low-rank visual prompting for efficient vision model adaptation. In *ICLR*, 2025.

[21] J. Kaplan, S. McCandlish, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[22] P. Khosla, P. Teterwak, et al. Supervised contrastive learning. *NeurIPS*, 33:18661–18673, 2020.

[23] M. Kim, J. Tack, et al. Adversarial self-supervised contrastive learning. *NeurIPS*, 33:2983–2994, 2020.

[24] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *ICML*, pages 1885–1894. PMLR, 2017.

[25] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

[26] C. Li, R. Pang, et al. An embarrassingly simple backdoor attack on self-supervised learning. In *ICCV*, pages 4367–4378, 2023.

[27] Y. Li, Y. Bai, et al. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. *NeurIPS*, 35:13238–13250, 2022.

[28] Y. Li, M. Zhu, et al. Black-box dataset ownership verification via backdoor watermarking. *IEEE TIFS*, 18:2318–2332, 2023.

[29] Y. Li, L. Zhu, et al. Move: Effective and harmless ownership verification via embedded external features. *IEEE TPAMI*, 2025.

[30] G. Liu, T. Xu, et al. Your model trains on my data? protecting intellectual property of training data via membership fingerprint authentication. *IEEE TIFS*, 17:1024–1037, 2022.

[31] K. Liu, B. Dolan-Gavitt, et al. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *RAID*, pages 273–294. Springer, 2018.

[32] A. Madry, A. Makelov, et al. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[33] P. Maini, M. Yaghini, et al. Dataset inference: Ownership resolution in machine learning. In *ICLR*, 2021.

[34] N. Memon and P. W. Wong. Protecting digital media content. *Communications of the ACM*, 41(7):35–43, 1998.

[35] P. Mohassel and Y. Zhang. Secureml: A system for scalable privacy-preserving machine learning. In *IEEE S&P*, pages 19–38. IEEE, 2017.

[36] R. Müller, S. Kornblith, and G. E. Hinton. When does label smoothing help? *NeurIPS*, 32, 2019.

[37] S. G. Müller and F. Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *ICCV*, pages 774–782, 2021.

[38] A. v. d. Oord, Y. Li, et al. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[39] M. Oquab, T. Darcet, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Featured Certification.

[40] A. Radford, J. W. Kim, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PmLR, 2021.

[41] S. Shan, W. Ding, et al. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *IEEE S&P*, pages 807–825. IEEE, 2024.

[42] S. Shao, Y. Li, M. Zheng, et al. Databench: Evaluating dataset auditing in deep learning from an adversarial perspective. *arXiv preprint arXiv:2507.05622*, 2025.

[43] W. Shi, A. Ajith, et al. Detecting pretraining data from large language models. In *ICLR*, 2024.

[44] R. Shokri, M. Stronati, et al. Membership inference attacks against machine learning models. In *IEEE S&P*, pages 3–18. IEEE, 2017.

[45] O. Solon. Facial recognition's 'dirty little secret': Millions of online photos scraped without consent, 2019.

[46] S. J. Stolfo, M. B. Salem, et al. Fog computing: Mitigating insider data theft attacks in the cloud. In *2012 IEEE symposium on security and privacy workshops*, pages 125–128. IEEE, 2012.

[47] S. Szyller and N. Asokan. Conflicting interactions among protection mechanisms for machine learning models. In *AAAI*, volume 37, pages 15179–15187, 2023.

[48] S. Szyller, R. Zhang, et al. On the robustness of dataset inference. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

[49] R. Van Wegberg, S. Tajalizadehkhoob, et al. Plug and prey? measuring the commoditization of cybercrime via online anonymous markets. In *USENIX security*, pages 1009–1026, 2018.

[50] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[51] P. K. A. Vasu, H. Pouransari, et al. Mobileclip: Fast image-text models through multi-modal reinforced training. In *CVPR*, pages 15963–15974, 2024.

[52] D. Wahdany, M. Jagielski, et al. Differentially private prototypes for imbalanced transfer learning. In *AAAI-25, Philadelphia, PA, USA*, pages 20991–20999. AAAI Press, 2025.

[53] E. Wenger, X. Li, et al. Data isotopes for data provenance in dnns. *Proceedings on Privacy Enhancing Technologies*, 2024.

[54] Wikipedia contributors. Cambridge analytica, 2025.

[55] S. Yeom, I. Giacomelli, et al. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.

[56] J.-L. Yin, W. Wang, et al. Adversarial-inspired backdoor defense via bridging backdoor and adversarial attacks. In *AAAI*, volume 39, pages 9508–9516, 2025.

[57] C. Zhang, S. Bengio, et al. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

[58] X. Zhao, K. Zhang, et al. Invisible image watermarks are provably removable using generative ai. *NeurIPS*, 37:8643–8672, 2024.

[59] H. Zhu, S. Liang, et al. Reliable model watermarking: Defending against theft without compromising on evasion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10124–10133, 2024.

[60] H. Zhu, S. Liang, W. Wang, et al. Evading data provenance in deep neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[61] H. Zhu, S. Liang, et al. Revisiting data auditing in large vision-language models. *arXiv preprint arXiv:2504.18349*, 2025.

[62] Z. Zou, B. Gong, et al. Anti-neuron watermarking: Protecting personal data against unauthorized neural networks. In *ECCV*, pages 449–465. Springer, 2022.