

IrisFP: Adversarial-Example-based Model Fingerprinting with Enhanced Uniqueness and Robustness

Ziye Geng¹, Guang Yang², Yihang Chen¹, and Changqing Luo¹,

¹University of Houston, ²Virginia Commonwealth University

{zgeng2, ychen165, cluo3}@uh.edu, yangg2@vcu.edu

Abstract

We propose IrisFP, a novel adversarial-example-based model fingerprinting framework that enhances both uniqueness and robustness by leveraging multi-boundary characteristics, multi-sample behaviors, and fingerprint discriminative power assessment to generate composite-sample fingerprints. Three key innovations make IrisFP outstanding: 1) It positions fingerprints near the intersection of all decision boundaries—unlike prior methods that target a single boundary—thus increasing the prediction margin without placing fingerprints deep inside target class regions, enhancing both robustness and uniqueness; 2) It constructs composite-sample fingerprints, each comprising multiple samples close to the multi-boundary intersection, to exploit collective behavior patterns and further boost uniqueness; and 3) It assesses the discriminative power of generated fingerprints using statistical separability metrics developed based on two reference model sets, respectively, for pirated and independently-trained models, retains the fingerprints with high discriminative power, and assigns fingerprint-specific thresholds to such retained fingerprints. Extensive experiments show that IrisFP consistently outperforms state-of-the-art methods, achieving reliable ownership verification by enhancing both robustness and uniqueness.

1. Introduction

Adversarial-example-based model fingerprinting has recently gained attention as a promising approach to protect intellectual property (IP) of deep neural networks (DNNs) developed for many tasks like image classification. This type of model fingerprinting builds on the adversarial example technique to generate input-output pairs as fingerprints that exhibit distinctive model-specific behavior. More specifically, such a fingerprinting approach introduces subtle perturbations to clean inputs to elicit unique model-specific responses from a protected model: the protected model produces outputs that deviate from those induced

by the original clean inputs, whereas independently-trained models tend to produce outputs consistent with the clean inputs and significantly different from those of the protected model [3, 15, 21, 23, 25, 38]. These crafted fingerprints are then utilized for ownership verification [5, 37].

So far, researchers have developed many adversarial-example-based model fingerprinting methods [3, 6, 19, 20, 23, 30, 32, 33], which typically generate individual fingerprints near a single decision boundary, without considering their relative distances to other boundaries. For example, IPGuard [3] crafts fingerprints close to the decision boundaries of the protected model to capture model-specific behaviors; UAP [23] leverages universal perturbations to induce model-dependent outputs over a broad set of inputs; ADV-TRA [30] constructs adversarial trajectories composed of samples that traverse boundary regions to capture rich model characteristics; and AKH [6] constructs fingerprints by generating adversarial examples that originate from naturally misclassified samples.

However, adversarial-example-based model fingerprinting is inherently prone to model modification attacks like fine-tuning and pruning, as such attacks can shift the decision boundaries of protected models, consequently circumventing ownership verification [4, 9, 18]. To address this issue, prior methods propose placing fingerprints deep inside the regions of target classes—far away from decision boundaries—to enhance robustness against model modifications [3, 20]. Nevertheless, this strategy may in turn undermine uniqueness, as such fingerprints are less sensitive to model-specific decision boundaries. Consequently, existing methods have weak uniqueness or robustness, or even both. This observation has been corroborated by our experimental results in Figure 1. The experiments, performed on CIFAR-10 and Fashion-MNIST, assess uniqueness by measuring true negative rates (TNRs) under varying false negative rate (FNR) levels, and robustness by measuring true positive rates (TPRs) under different false positive rate (FPR) levels. A higher TNR indicates better uniqueness in distinguishing the protected model from independently-trained ones, while a higher TPR indicates stronger robust-

ness to model modifications. The results show that conventional methods can achieve either weak uniqueness or robustness, which highlights the inherent difficulty of generating effective fingerprints for ownership verification.

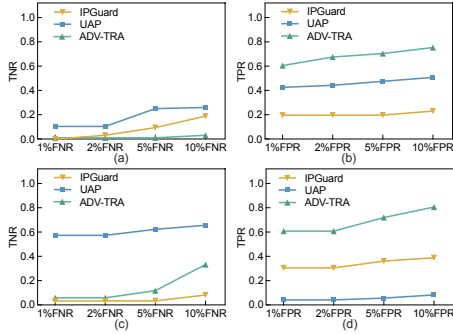


Figure 1. The uniqueness and robustness achieved by prior adversarial-example-based fingerprinting methods.

Main question. Based on the discussion so far, we have the following research question: *Is it possible to design a new adversarial-example-based model fingerprinting method that enhances both uniqueness and robustness?*

1.1. Our Answer

In this paper, we give an affirmative answer to the above question. Recent studies [1, 2] show that compared to fingerprints positioned on a single decision boundary, those placed at the intersection of multi-class decision boundaries has high sensitivity, i.e., achieving enhanced uniqueness. By leveraging this property, IBSF [1] and SDBF [2] were developed to detect model tampering. While the fingerprints placed at the intersection of multi-class decision boundaries can enhance uniqueness, they are extremely vulnerable to model modifications, i.e., struggling with weak robustness. Therefore, our question is: *Can we exploit decision boundary intersection to design a new fingerprinting method with enhanced uniqueness and robustness?*

To positively answer this question, we first provide an intuitive illustration by Figure 2 to show that proximity to all decision boundaries is crucial for enhancing robustness. As illustrated in Figure 2(a), sample s_1 lies within the region of class c_2 , positioned near the boundary between c_2 and c_4 , but is far from the other regions, while s_2 , which also belongs to c_2 , is located near the intersection of multiple boundaries, making it simultaneously close to all surrounding class regions. The corresponding output distributions for s_1 and s_2 , shown in Figure 2(b) and Figure 2(c), reveal that s_2 has a larger prediction margin—that is, a greater confidence gap between the top two predicted classes. This increased margin means enhancing robustness at no cost of uniqueness. This key observation motivates our design: to construct fingerprints that are not only model-specific but

also strategically placed near the multi-boundary intersection to improve uniqueness and robustness.

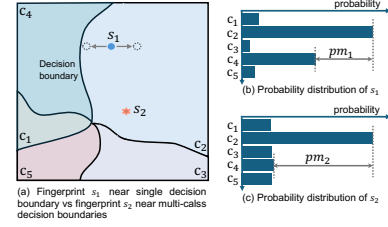


Figure 2. (a): The placement of input samples s_1 and s_2 in a region; (b): The prediction margin pm_1 of fingerprint s_1 ; and (c): The prediction margin pm_2 of fingerprint s_2 .

Our Design: We propose IrisFP, a novel model fingerprinting framework that generates **unI**que and **rob**ust composite-sample **F**inger**P**rints (IrisFP) by exploiting multi-boundary characteristics, multi-sample behaviors, and fingerprint quality assessment. Unlike conventional methods that place each fingerprint near a single decision boundary, IrisFP first generates fingerprint seeds by crafting adversarial samples to reside near the intersection of all decision boundaries of a protected model. This strategy increases the prediction margin without placing the sample deep inside the target class region, thus enhancing both robustness and uniqueness. To further boost uniqueness, IrisFP applies subtle, diverse perturbations to each fingerprint seed, generating multiple variants which, together with the seed, collectively form a composite-sample fingerprint. Compared to the original sample, the seed and its variants are crafted to elicit different responses from the protected model, while tending to produce consistent predictions by independently-trained ones, resulting in a clear behavioral gap that enhances the fingerprint’s discriminative capability. Moreover, IrisFP addresses a key limitation in prior adversarial-example-based fingerprinting methods—their lack of consideration for model modifications and independently trained models during fingerprint construction. To overcome this, IrisFP employs a fingerprint refinement strategy that evaluates the discriminative power of each fingerprint using statistical separability metrics to retain the fingerprints with high discriminative power, and assigns sample-specific thresholds to the retained fingerprints. Finally, IrisFP performs model ownership verification through a two-step process involving ownership matching and decision aggregation, specifically designed for composite-sample fingerprints.

2. Preliminary

2.1. Model Fingerprinting

An adversarial-example-based fingerprinting approach typically allows a model owner to craft inputs that elicit unique

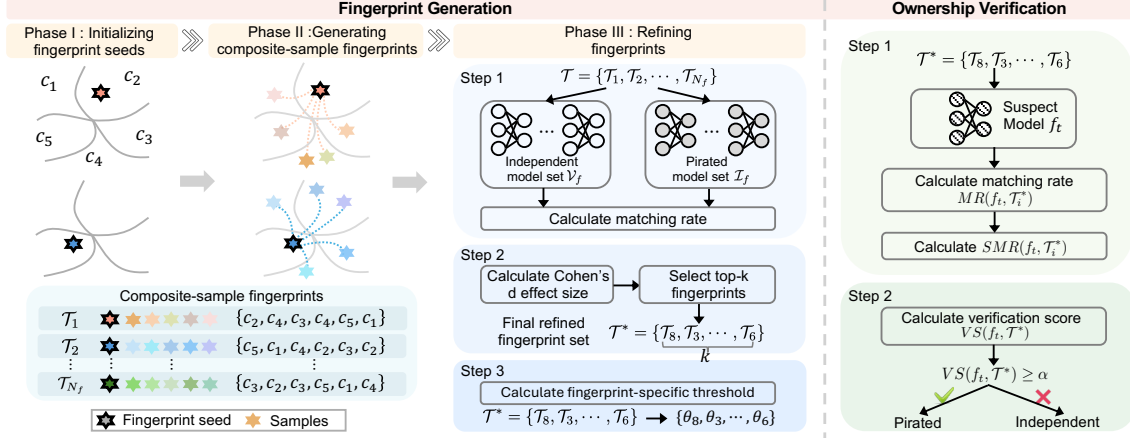


Figure 3. The overview of IrisFP.

model-specific responses. Given a protected model f and a clean input x_i with ground-truth label y_i , a small perturbation δ_i , which is constrained to capture model-specific boundary characteristics, is added to produce a perturbed input $\hat{x}_i = x_i + \delta_i$, such that f outputs a target label $\hat{y}_i \neq y_i$, resulting in fingerprint $(\hat{x}_i, \hat{y}_i)$. We mathematically formulate this process by

$$\min \|x_i - \hat{x}_i\|, \quad \text{s.t.} \quad f(\hat{x}_i) \neq y_i, \quad (1)$$

where $\|\cdot\|$ represents a distance metric (e.g., ℓ_2 or ℓ_∞ norm). This process can generate a set of fingerprints whose inputs are typically located near the decision boundaries of model f , making the fingerprints' behavior highly model-specific. The resulting fingerprints are used to query a target model f_t to produce query results that are compared to the expected outputs for determining ownership.

2.2. Model Modification Attacks

In real-world scenarios, an adversary may alter a protected model f through model modification techniques to create a pirated variant f_p in order to evade ownership verification. Common modification techniques include fine-tuning (FT) [17, 41], pruning (PR) [8, 16], adversarial training (AT) [24, 40], and knowledge distillation (KD) [27]. In practice, adversaries may also combine these techniques—such as injecting noise or pruning before fine-tuning—to amplify the degree of modification. Such alterations can shift the model's decision boundaries, potentially invalidating existing fingerprints and thereby enabling evasion of verification [42]. For example, given a fingerprint $(\hat{x}_i, \hat{y}_i)$ for the protected model f , querying f_p with $\hat{x}_i$ might yield an output differing from the expected one, i.e., $f_p(\hat{x}_i) \neq \hat{y}_i = f(\hat{x}_i)$, thus undermining the reliability of ownership verification.

3. Methodology

3.1. Overview

We propose IrisFP, a novel model fingerprinting framework that enhances both uniqueness and robustness. As shown in Figure 3, IrisFP comprises two main processes: fingerprint generation and ownership verification. The fingerprint generation process includes three phases: fingerprint seed initialization, composite-sample fingerprint generation, and fingerprint set refinement. The first phase places fingerprint seeds near the intersection of all decision boundaries to maximize prediction margins while keeping model-specific behavior intact. The second phase then minimally perturbs each seed to produce a set of variants with diverse outputs that, along with the seed, form a composite-sample fingerprint. The third phase further exploits statistical separability to select the most discriminative fingerprints from the generated ones to construct a fingerprint set, and assigns a specific threshold to each selected fingerprint.

The ownership verification process involves two steps: ownership matching and decision aggregation. At the first step, the matching rate across samples of each composite-sample fingerprint is computed and compared with its corresponding threshold to determine if the fingerprint is matched. At the second step, the matching outcomes across all fingerprints are aggregated to make a final decision. If the proportion of matched fingerprints exceeds a predefined threshold, the target model is deemed pirated; otherwise, it is identified as independently-trained.

3.2. The Fingerprint Generation Process

3.2.1. Phase I: Initializing Fingerprint Seeds

Our proposed model fingerprinting framework starts with generating initial fingerprint seeds. Given a protected model f trained on a dataset $\mathcal{D}$ with C classes, we first randomly select N_f input-label pairs $\{(x_i^0, y_i^0)\}_{i=1}^{N_f}$ from $\mathcal{D}$, where x_i^0

is an input with ground-truth label y_i^0 . For each x_i^0 , our adversarial-example-based model fingerprinting method introduces a trainable perturbation δ_i^0 to produce a fingerprint seed $\hat{x}_i^0 = x_i^0 + \delta_i^0$. Unlike traditional fingerprinting methods that directly push a fingerprint toward a single decision boundary associated with a specific target class, our method instead guides $\hat{x}_i^0$ to be located near the intersection of f 's decision boundaries. This design enhances robustness: positioning samples close to such an intersection enables f to predict the target class $\hat{y}_i^0$ with high confidence, while distributing the remaining probability almost evenly across other classes. Thus, model-specific output behavior is preserved, and the prediction margin is increased, improving robustness to modifications without sacrificing uniqueness.

To obtain such seeds, we define a biased probability distribution $p_i \in \mathbb{R}^C$ for each sample with respect to a randomly chosen target class $\hat{y}_i^0$, with $p_i(y)$ denoting the probability assigned to class y :

$$p_i(y) = \begin{cases} \frac{1}{C} + \tau, & \text{if } y = \hat{y}_i^0, \\ \frac{1 - (\frac{1}{C} + \tau)}{C-1}, & \text{otherwise,} \end{cases} \quad (2)$$

where $0 < \tau < 1$ is a tunable hyperparameter controlling the degree of bias toward $\hat{y}_i^0$. Then, we define $f_o(x)$ as the softmax output of the model f , which produces a probability distribution over C classes. The fingerprint seed $\hat{x}_i^0$ is then optimized to ensure that the model's output distribution $f_o(\hat{x}_i^0)$ aligns with the predefined p_i . This is achieved by minimizing the Kullback-Leibler (KL) divergence between $f_o(\hat{x}_i^0)$ and p_i , with an L_1 regularization term applied to minimize the perturbation. The loss function is given by $\mathcal{L}_{\text{phase1}} = \text{KL}(f_o(\hat{x}_i^0) || p_i) + \lambda_1 \|\delta_i^0\|_1$, where λ_1 is a regularization coefficient. This optimization process yields a set of fingerprint seeds $\{(\hat{x}_i^0, \hat{y}_i^0)\}_{i=1}^{N_f}$, each positioned near the intersection of multi-class decision boundaries.

3.2.2. Phase II: Generating Composite-sample Fingerprints

To enhance uniqueness, we extend each fingerprint seed into a composite-sample fingerprint comprising multiple minimally perturbed variants designed to elicit diverse responses from the protected model. Specifically, each fingerprint seed $\hat{x}_i^0$ is perturbed by a set of small, trainable perturbations $\{\delta_i^1, \delta_i^2, \dots, \delta_i^T\}$ to generate a set of variants $\{\hat{x}_i^1, \hat{x}_i^2, \dots, \hat{x}_i^T\}$, where $\hat{x}_i^t = \hat{x}_i^0 + \delta_i^t$, for $1 \leq t \leq T$. Each variant is assigned a target class uniformly sampled from $\{1, 2, \dots, C\}$. To position each variant near the intersection of multiple decision boundaries, we define a biased target probability distribution p_i^t for each variant toward a randomly chosen target class $\hat{y}_i^t$, and $p_i^t(y)$ is the probability assigned to class y :

$$p_i^t(y) = \begin{cases} \frac{1}{C} + \tau, & \text{if } y = \hat{y}_i^t, \\ \frac{1 - (\frac{1}{C} + \tau)}{C-1}, & \text{otherwise.} \end{cases} \quad (3)$$

The perturbations $\{\delta_i^1, \delta_i^2, \dots, \delta_i^T\}$ are jointly optimized by minimizing the average KL divergence between the model's output distribution and the biased targets. The loss function is formulated as $\mathcal{L}_{\text{phase2}} = \frac{1}{T} \sum_{t=1}^T [\text{KL}(f_o(\hat{x}_i^t) || p_i^t) + \lambda_2 \|\delta_i^t\|_1]$, where λ_2 is a regularization coefficient. This process yields a fingerprint set with N_f composite-sample fingerprints $\{\mathcal{T}_i\}_{i=1}^{N_f}$, where $\mathcal{T}_i = \{(\hat{x}_i^0, \hat{y}_i^0), (\hat{x}_i^1, \hat{y}_i^1), \dots, (\hat{x}_i^T, \hat{y}_i^T)\}$.

By applying minimal, model-specific perturbations to each seed, IrisFP captures subtle variations in the model's decision landscape. These perturbations are optimized to induce diverse responses, leveraging the fact that adversarial perturbations are highly sensitive to a model's decision boundaries. Since independently-trained models—even those with architectural similarity—often exhibit different decision boundaries, their outputs across all samples in a fingerprint are typically inconsistent with those of the protected model, failing to replicate the behavioral pattern elicited by the protected model. This behavioral discrepancy enhances the discriminative capability of composite-sample fingerprints.

3.2.3. Phase III: Refining Fingerprints

Some of the generated composite-sample fingerprints may not have strong and consistent discriminative behaviors, inevitably degrading ownership verification performance. Therefore, we propose a fingerprint refinement strategy comprising three steps: calculating the matching rate, selecting composite-sample fingerprints, and computing fingerprint-specific thresholds.

Prior to describing these steps, we define two reference model sets used throughout the refinement process:

- **Pirated Model Set ($\mathcal{V}_f$):** Models derived from the protected model via model modification techniques.
- **Independent Model Set ($\mathcal{I}_f$):** Models trained independently from scratch using different initializations, architectures, or training data.

Both sets are used to evaluate whether a composite-sample fingerprint elicits distinguishable responses between pirated and independently-trained models.

Step 1: Calculating the matching rate. To support fingerprint refinement, we first compute the fingerprint-level matching rate $MR(f, \mathcal{T}_i)$ for each composite-sample fingerprint $\mathcal{T}_i$ across models in both $\mathcal{V}_f$ and $\mathcal{I}_f$. Given $\mathcal{T}_i = \{(\hat{x}_i^0, \hat{y}_i^0), (\hat{x}_i^1, \hat{y}_i^1), \dots, (\hat{x}_i^T, \hat{y}_i^T)\}$ and a model f_j within $\mathcal{V}_f$ or $\mathcal{I}_f$, the fingerprint-level matching rate is defined as $MR(f_j, \mathcal{T}_i) = \frac{1}{T+1} \sum_{t=0}^T \mathbb{I}[f_j(\hat{x}_i^t) = \hat{y}_i^t]$, where $\mathbb{I}[\cdot]$ is an indicator function that returns “1” if f_j with input $\hat{x}_i^t$ outputs $\hat{y}_i^t$, and “0” otherwise. This metric quantifies the degree to which the model's behavior aligns with the protected model.

After computing the matching rate for each fingerprint-model pair, we aggregate these results across the models to compute statistics that describe each fingerprint's matching

behavior on the pirated model set $\mathcal{V}_f$ and the independent model set $\mathcal{I}_f$. For a given composite-sample fingerprint $\mathcal{T}_i$, we compute the mean matching rate $\mu_i^{\mathcal{V}}$ and standard deviation $\sigma_i^{\mathcal{V}}$ across all models in $\mathcal{V}_f$, and likewise obtain $\mu_i^{\mathcal{I}}$ and $\sigma_i^{\mathcal{I}}$ for $\mathcal{I}_f$. These results are subsequently used to select fingerprints and compute the fingerprint-specific thresholds.

Step 2: Selecting composite-sample fingerprints. To refine the fingerprint set, we assess the discriminative capability of each fingerprint by quantifying the behavioral separation it induces between $\mathcal{V}_f$ and $\mathcal{I}_f$. Thus, we compute Cohen’s d effect size for each $\mathcal{T}_i$ by $d_i = (\mu_i^{\mathcal{V}} - \mu_i^{\mathcal{I}}) / \sqrt{\frac{1}{2}((\sigma_i^{\mathcal{V}})^2 + (\sigma_i^{\mathcal{I}})^2)}$. Cohen’s d effect size is employed to quantify the standardized separation between the two matching rate distributions of pirated and independently-trained models, capturing both the magnitude of behavioral difference and the statistical stability of that difference. A larger d_i indicates that a fingerprint $\mathcal{T}_i$ elicits more divergent behaviors between pirated and independent model sets, while maintaining higher consistency within each group, making it a stronger candidate for reliable ownership verification. Based on d_i , we select the top- K fingerprints with the highest d_i values to construct a final fingerprint set $\mathcal{T}^* = \{\mathcal{T}_i^*\}_{i=1}^K$.

Step 3: Computing the fingerprint-specific threshold. To account for the varying discriminative power of individual fingerprints, we develop a fingerprint-specific thresholding strategy that assigns a specific threshold to each fingerprint. This design ensures that each fingerprint enforces a verification criterion consistent with its statistical separability between pirated and independently-trained models. The threshold defines the minimum matching rate required to assert a target model f_t to be pirated. Each threshold θ_i is computed based on the matching rate statistics of a fingerprint $\mathcal{T}_i^*$ across $\mathcal{V}_f$ and $\mathcal{I}_f$. More specifically, θ_i is calculated as a weighted average of their means, with weights inversely proportional to their standard deviations:

$$\theta_i = \begin{cases} \frac{\mu_i^{\mathcal{V}}/\sigma_i^{\mathcal{V}} + \mu_i^{\mathcal{I}}/\sigma_i^{\mathcal{I}}}{1/\sigma_i^{\mathcal{V}} + 1/\sigma_i^{\mathcal{I}}}, & \text{if } \sigma_i^{\mathcal{V}}, \sigma_i^{\mathcal{I}} > 0, \\ \frac{\mu_i^{\mathcal{V}} + \mu_i^{\mathcal{I}}}{2}, & \text{otherwise.} \end{cases} \quad (4)$$

This thresholding strategy assigns each fingerprint an individually optimized threshold, jointly considering the mean and standard deviation of matching-rate distributions to achieve maximal discriminative capability.

3.3. The Ownership Verification Process

To verify the ownership of a target model f_t , we assess how it responds to all samples in each composite-sample fingerprint within the fingerprint set $\mathcal{T}^*$. To this end, we design a two-step ownership verification scheme, comprising ownership matching and decision aggregation. The detailed design is presented below.

Step 1: Ownership matching. For each composite-sample fingerprint $\mathcal{T}_i^* \in \mathcal{T}^*$, its matching rate is computed by $MR(f_t, \mathcal{T}_i^*) = \frac{1}{T+1} \sum_{t=0}^T \mathbb{I}[f_t(\hat{x}_i^t) = \hat{y}_i^t]$. Then, the resulting matching rate is compared with the corresponding fingerprint-specific threshold θ_i , i.e., $SMR(f_t, \mathcal{T}_i^*) = \mathbb{I}[MR(f_t, \mathcal{T}_i^*) \geq \theta_i]$, where a result of “1” indicates that the target model’s behavior on $\mathcal{T}_i^*$ is consistent with the protected model f , and “0” otherwise.

Step 2: Decision aggregation. After obtaining the matching decision for each composite-sample fingerprint, a final ownership decision will be made. Since pirated models inherit decision behavior from the protected model f , they are expected to match the majority of fingerprints, whereas independently-trained models tend to diverge. We aggregate these decisions to produce a final verification score by $VS(f_t, \mathcal{T}^*) = \frac{1}{K} \sum_{i=1}^K SMR(f_t, \mathcal{T}_i^*)$, which quantifies the overall behavioral alignment of the target model f_t with f . The target model is classified as pirated if $VS(f_t, \mathcal{T}^*) \geq \alpha$, where α is a predefined decision threshold; otherwise, it is deemed independently-trained.

4. Experiments

4.1. Experimental Setting

To evaluate the effectiveness of IrisFP, we conduct comprehensive experiments on different DNN models trained on five widely-used datasets: CIFAR-10 [11], CIFAR-100 [12], Fashion-MNIST [28], MNIST [14], and Tiny-ImageNet [13]. We compare IrisFP with four representative fingerprinting methods: IPGuard [3], UAP [23], ADV-TRA [30], and AKH[6]. We also evaluate the robustness of IrisFP against model modification attacks by considering six types of attacks, including fine-tuning (FT), pruning (PR), knowledge distillation (KD), adversarial training (AT), prune-then-tune (PFT), and noise injection-then-fine-tuning (NFT). To ensure fair comparisons, we query 200 times in each experiment, i.e., 40 composite fingerprints with 5 samples each for our proposed IrisFP and 200 fingerprints for all baseline methods. In the following, we will summarize protected models, reference model sets, testing model set, and evaluation metrics, and their detailed settings can be found in Sec. 12 in the Supplementary Material.

Protected Model. We consider three different architectures, including ResNet-18, MobileNet-V2, and ViT-B/16. For ResNet-18 and MobileNet-V2, each protected model is trained from scratch across its corresponding different dataset, while for ViT-B/16, its protected model is fine-tuned from a pretrained ViT-B/16 model on Tiny-ImageNet.

Reference Model Sets. We construct two reference model sets: one for pirated models and the other for independently-trained models. Specifically, for each protected model, its corresponding pirated model set is con-

Table 1. AUCs achieved by different fingerprinting approaches across datasets and protected model architectures.

Protected Model	Method	CIFAR-10	CIFAR-100	Fashion-MNIST	MNIST	Tiny-ImageNet
ResNet-18	IPGuard	0.675 ± 0.095	0.654 ± 0.097	0.721 ± 0.061	0.471 ± 0.089	0.726 ± 0.099
	UAP	0.732 ± 0.014	0.761 ± 0.049	0.721 ± 0.061	0.789 ± 0.036	0.812 ± 0.045
	ADV-TRA	0.799 ± 0.003	0.806 ± 0.005	0.845 ± 0.010	0.753 ± 0.019	0.767 ± 0.073
	AKH	0.710 ± 0.052	0.785 ± 0.086	0.765 ± 0.042	0.820 ± 0.026	0.823 ± 0.043
	IrisFP	0.893 ± 0.015	0.916 ± 0.009	0.940 ± 0.031	0.854 ± 0.024	0.874 ± 0.052
MobileNet-V2	IPGuard	0.821 ± 0.047	0.823 ± 0.021	0.607 ± 0.035	0.634 ± 0.010	0.692 ± 0.019
	UAP	0.749 ± 0.028	0.836 ± 0.042	0.816 ± 0.019	0.743 ± 0.021	0.806 ± 0.039
	ADV-TRA	0.824 ± 0.051	0.795 ± 0.011	0.782 ± 0.028	0.720 ± 0.023	0.877 ± 0.045
	AKH	0.860 ± 0.067	0.867 ± 0.044	0.797 ± 0.054	0.805 ± 0.019	0.863 ± 0.034
	IrisFP	0.936 ± 0.011	0.937 ± 0.017	0.963 ± 0.017	0.876 ± 0.015	0.934 ± 0.023
ViT-B/16	IPGuard	—	—	—	—	0.778 ± 0.029
	UAP	—	—	—	—	0.803 ± 0.042
	ADV-TRA	—	—	—	—	0.832 ± 0.019
	AKH	—	—	—	—	0.806 ± 0.010
	IrisFP	—	—	—	—	0.887 ± 0.036

structed by modifying the protected model through three model modification techniques, including fine-tuning (FT), knowledge distillation (KD), and adversarial training (AT). Each attack type produces three pirated variants, leading to 9 pirated variants in the pirated model set. On the other hand, our independent model set is constructed by using three architectures: ResNet-18, MobileNet-V2, and DenseNet-121. The models in this set are independently trained from scratch with different seeds and hyperparameters from training the protected models and testing models. Each architecture will be instantiated with 3 different random seeds to create 3 independently-trained models, resulting in a total of 9 models in the independent model set.

Testing Model Set. We evaluate IrisFP using a testing model set comprising both pirated and independently-trained models. The pirated models are derived by applying six attacks (i.e., FT, PR, KD, AT, PFT, and NFT) to each protected model. For each attack type, we generate 20 variants using different random seeds and hyperparameters, resulting in 120 pirated models. The independently-trained models are built from scratch using six architectures: ResNet-18, ResNet-50, MobileNet-V2, MobileNet-V3 Large, EfficientNet-B2, and DenseNet-121. For each architecture, 20 models are trained with distinct seeds and hyperparameters, without any access to the protected models or the models in the reference model set, yielding 120 independently-trained models.

Metrics. We use the Area Under the ROC (Receiver Operating Characteristic) Curve (AUC) as the main metric to evaluate the performance of our proposed fingerprinting method. AUC quantifies the probability that a randomly selected pirated model receives a higher matching score than a randomly selected independently-trained model. Higher AUC values reflect stronger discriminative capability. All AUC results are reported as the mean and standard deviation over five independent runs.

4.2. Main Performance

4.2.1. The Effectiveness of IrisFP

Table 1 shows the superior effectiveness of IrisFP by reporting the AUCs achieved by IrisFP and four representative baselines (IPGuard, UAP, ADV-TRA, and AKH) across five benchmark datasets—CIFAR-10, CIFAR-100, Fashion-MNIST, MNIST, and Tiny-ImageNet. To ensure a comprehensive evaluation, we train two protected models using ResNet-18 and MobileNet-V2 architectures on each of the five datasets. In addition, we incorporate a larger and more complex architecture by training a protected model using ViT-B/16 on Tiny-ImageNet. From the table, we can observe that IrisFP consistently achieves the highest AUC across the models and datasets. For example, on CIFAR-100, IrisFP achieves an AUC of 0.916 when fingerprinting ResNet-18 and an AUC of 0.937 when fingerprinting MobileNet-V2, achieving the improvements by 13.7% and 8.1%, respectively, over the best-performing baselines. It is noteworthy that IrisFP maintains the highest AUC of 0.887 when fingerprinting ViT-B/16 trained on Tiny-ImageNet, demonstrating that IrisFP can achieve reliable ownership verification on larger and more complex models.

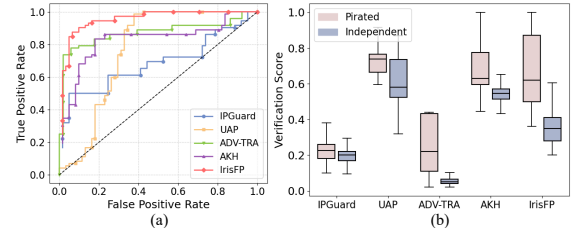


Figure 4. (a) The ROC curve and (b) the distribution of verification score on the protected model with ResNet-18 architecture trained on Fashion-MNIST.

To further validate IrisFP’s discriminative power, Figure 4 visualizes the ROC curves and the distribution of verification scores. These plots provide intuitive insights

Table 2. AUCs under six model modification attacks across datasets and methods on the protected model with ResNet-18 architecture.

Dataset	Method	FT	PR	KD	AT	PFT	NFT
CIFAR-10	IPGuard	0.656 $\pm$ 0.101	0.997 $\pm$ 0.005	0.515 $\pm$ 0.082	0.511 $\pm$ 0.265	0.687 $\pm$ 0.144	0.724 $\pm$ 0.142
	UAP	0.809 $\pm$ 0.011	0.998 $\pm$ 0.005	0.722 $\pm$ 0.013	0.346 $\pm$ 0.017	0.856 $\pm$ 0.016	0.868 $\pm$ 0.018
	ADV-TRA	1.000 $\pm$ 0.000	1.000 $\pm$ 0.000	0.805 $\pm$ 0.009	0.025 $\pm$ 0.006	0.959 $\pm$ 0.012	0.962 $\pm$ 0.009
	AKH	0.921 $\pm$ 0.005	0.876 $\pm$ 0.020	0.621 $\pm$ 0.042	0.531 $\pm$ 0.108	0.701 $\pm$ 0.029	0.733 $\pm$ 0.066
	IrisFP	0.954 $\pm$ 0.046	1.000 $\pm$ 0.000	0.616 $\pm$ 0.068	0.929 $\pm$ 0.042	0.965 $\pm$ 0.009	0.968 $\pm$ 0.014
CIFAR-100	IPGuard	0.617 $\pm$ 0.137	0.998 $\pm$ 0.003	0.593 $\pm$ 0.141	0.621 $\pm$ 0.106	0.576 $\pm$ 0.104	0.584 $\pm$ 0.129
	UAP	0.764 $\pm$ 0.069	0.900 $\pm$ 0.002	0.709 $\pm$ 0.071	0.677 $\pm$ 0.053	0.752 $\pm$ 0.052	0.770 $\pm$ 0.065
	ADV-TRA	0.863 $\pm$ 0.010	1.000 $\pm$ 0.000	0.899 $\pm$ 0.007	0.338 $\pm$ 0.008	0.864 $\pm$ 0.014	0.863 $\pm$ 0.012
	AKH	0.813 $\pm$ 0.076	0.670 $\pm$ 0.185	0.858 $\pm$ 0.061	0.221 $\pm$ 0.076	0.671 $\pm$ 0.050	0.775 $\pm$ 0.090
	IrisFP	0.953 $\pm$ 0.011	1.000 $\pm$ 0.000	0.927 $\pm$ 0.006	0.758 $\pm$ 0.031	0.933 $\pm$ 0.009	0.928 $\pm$ 0.019
Fashion-MNIST	IPGuard	0.688 $\pm$ 0.139	0.990 $\pm$ 0.001	0.592 $\pm$ 0.096	0.612 $\pm$ 0.115	0.632 $\pm$ 0.187	0.636 $\pm$ 0.216
	UAP	0.809 $\pm$ 0.055	0.869 $\pm$ 0.045	0.749 $\pm$ 0.012	0.716 $\pm$ 0.018	0.797 $\pm$ 0.061	0.781 $\pm$ 0.045
	ADV-TRA	0.980 $\pm$ 0.000	0.990 $\pm$ 0.000	0.830 $\pm$ 0.004	0.158 $\pm$ 0.008	0.981 $\pm$ 0.001	0.980 $\pm$ 0.000
	AKH	0.855 $\pm$ 0.118	0.990 $\pm$ 0.000	0.844 $\pm$ 0.006	0.203 $\pm$ 0.038	0.759 $\pm$ 0.078	0.737 $\pm$ 0.091
	IrisFP	0.982 $\pm$ 0.002	0.992 $\pm$ 0.000	0.853 $\pm$ 0.037	0.816 $\pm$ 0.190	0.982 $\pm$ 0.003	0.983 $\pm$ 0.002
MNIST	IPGuard	0.378 $\pm$ 0.022	0.733 $\pm$ 0.030	0.624 $\pm$ 0.108	0.067 $\pm$ 0.065	0.432 $\pm$ 0.045	0.456 $\pm$ 0.122
	UAP	0.966 $\pm$ 0.009	0.630 $\pm$ 0.011	0.455 $\pm$ 0.054	0.946 $\pm$ 0.088	0.892 $\pm$ 0.007	0.959 $\pm$ 0.005
	ADV-TRA	0.965 $\pm$ 0.002	0.753 $\pm$ 0.028	0.485 $\pm$ 0.023	0.475 $\pm$ 0.019	0.975 $\pm$ 0.003	0.976 $\pm$ 0.001
	AKH	0.980 $\pm$ 0.000	0.797 $\pm$ 0.013	0.745 $\pm$ 0.011	0.837 $\pm$ 0.069	0.915 $\pm$ 0.010	0.920 $\pm$ 0.030
	IrisFP	0.985 $\pm$ 0.001	0.966 $\pm$ 0.007	0.467 $\pm$ 0.091	0.704 $\pm$ 0.025	0.983 $\pm$ 0.002	0.985 $\pm$ 0.001
Tiny-ImageNet	IPGuard	0.958 $\pm$ 0.044	0.754 $\pm$ 0.129	0.525 $\pm$ 0.136	0.316 $\pm$ 0.118	0.867 $\pm$ 0.133	0.954 $\pm$ 0.034
	UAP	0.953 $\pm$ 0.038	0.981 $\pm$ 0.001	0.455 $\pm$ 0.089	0.411 $\pm$ 0.114	0.967 $\pm$ 0.012	0.952 $\pm$ 0.016
	ADV-TRA	0.970 $\pm$ 0.040	0.917 $\pm$ 0.056	0.466 $\pm$ 0.069	0.173 $\pm$ 0.143	0.980 $\pm$ 0.064	0.948 $\pm$ 0.036
	AKH	0.937 $\pm$ 0.040	0.875 $\pm$ 0.019	0.326 $\pm$ 0.055	0.488 $\pm$ 0.099	0.988 $\pm$ 0.024	0.944 $\pm$ 0.051
	IrisFP	0.986 $\pm$ 0.032	1.000 $\pm$ 0.000	0.661 $\pm$ 0.108	0.545 $\pm$ 0.106	0.992 $\pm$ 0.015	0.977 $\pm$ 0.027

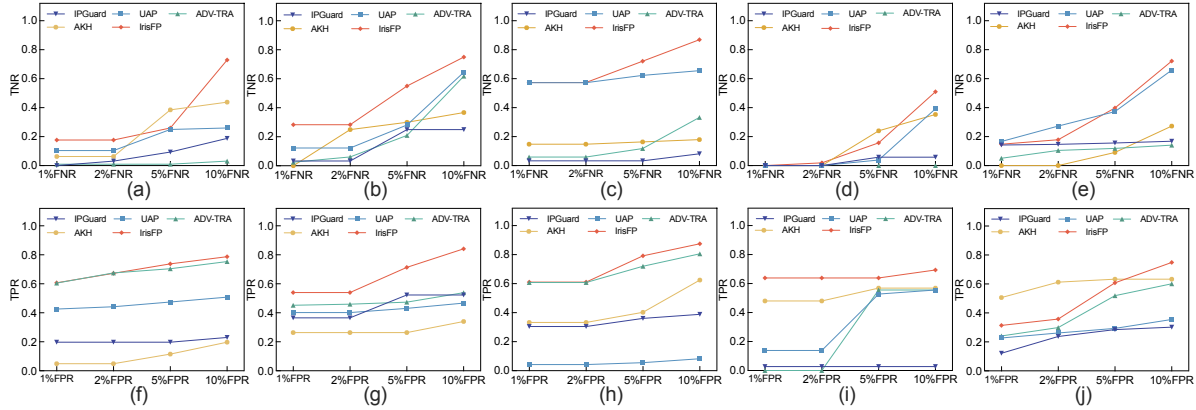


Figure 5. TNR-FNR and TPR-FPR curves on the protected models with ResNet-18 architecture trained on CIFAR-10 ((a) and (f)), CIFAR-100 ((b) and (g)), Fashion-MNIST ((c) and (h)), MNIST ((d) and (i)), and Tiny-ImageNet ((e) and (j)).

into how well each method distinguishes pirated models from independently-trained ones. Figure 4(a) shows the ROC curves for all five methods, illustrating the trade-off between TPR and FPR. IrisFP clearly demonstrates the strongest discriminative capability, with its curve close to the top-left corner, outperforming all baselines. In contrast, IPGuard’s and UAP’s curves lie near the diagonal, indicating weak uniqueness achieved by them, while ADV-TRA and AKH perform better but still fall short of IrisFP. Figure 4(b) presents two distributions of fingerprint verification scores achieved by each fingerprinting method under both pirated and independently-trained models. A wider separation between the two distributions reflects a stronger

discriminative capability. As shown in this figure, IrisFP achieves a clear separation, while baselines exhibit greater distributional overlap, reflecting their limited ability to differentiate pirated models from independently-trained ones.

4.2.2. Evaluation of IrisFP’s Uniqueness and Robustness

Figure 5 demonstrates the uniqueness and robustness achieved by IrisFP. The evaluation is performed under varying FNR and FPR levels, specifically 1%, 2%, 5%, and 10%. We can observe from the figure that IrisFP consistently outperforms all baselines, achieving the highest TNR (as shown by (a)-(e)) and the highest TPR (as shown by (f)-(j)), thereby demonstrating superior uniqueness and robustness. Among the baselines, ADV-TRA exhibits strong

robustness—achieving high TPR by effectively identifying pirated models—but suffers from low TNR, indicating poor uniqueness due to frequent misidentification of independently-trained models. In contrast, UAP maintains relatively stable TNR, indicating better uniqueness, but struggles with lower TPR, especially on Fashion-MNIST, revealing limited robustness. Moreover, AKH shows inconsistent behavior, performing well in either robustness or uniqueness but not both. Additionally, IPGuard performs the worst, failing to reliably distinguish pirated models from independently-trained ones under any setting. These results demonstrate that IrisFP is the only method achieving high robustness and uniqueness simultaneously, making it a reliable model fingerprinting solution.

4.2.3. Robustness to Model Modification Attacks

Table 2 shows the effectiveness of IrisFP against model modifications that aim to remove or distort the original fingerprints. Specifically, we compare IrisFP with four representative baselines using the protected model with ResNet-18 architecture trained across five datasets—CIFAR-10, CIFAR-100, Fashion-MNIST, MNIST, and Tiny-ImageNet—under six representative attacks: FT, PR, KD, AT, NFT, and PFT. The comparison is based on the AUCs achieved under each setting. We can observe that IrisFP consistently achieves the best performance under CIFAR-100, Fashion-MNIST, and Tiny-ImageNet across all six attacks. For instance, IrisFP attains AUCs of 0.953 (FT), 0.927 (KD), and 0.758 (AT) for CIFAR-100, and 0.982 (FT), 0.853 (KD), and 0.816 (AT) for Fashion-MNIST, substantially outperforming the baselines. For CIFAR-10, IrisFP achieves the highest AUCs under four attacks (PR, AT, PFT, NFT) and slightly lower AUC under FT and KD. These observations demonstrate IrisFP’s strong robustness to a wide range of model modification attacks.

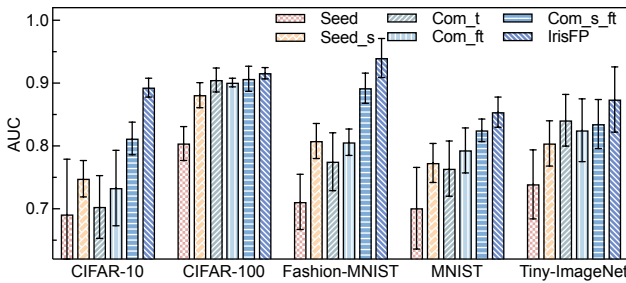


Figure 6. AUCs under different fingerprint configurations on the protected model with ResNet-18 architecture.

4.3. Ablation Studies

We conduct ablation studies to evaluate the effectiveness of each component in IrisFP, including fingerprint seed initialization, composite-sample fingerprint generation, fingerprint selection, and fingerprint-specific threshold:

- **Seed:** Directly using initial fingerprint seeds without further processing.

- **Seed_s:** Applying fingerprint selection to fingerprint seeds.
- **Com_ft:** Extending fingerprint seeds into composite-sample fingerprints and using a fixed global threshold.
- **Com_s_ft:** Extending fingerprint seeds into composite-sample fingerprints, using a fixed global threshold, and applying fingerprint selection.
- **Com_t:** Extending fingerprint seeds into composite-sample fingerprints and applying fingerprint-specific thresholds, without fingerprint selection.

Figure 6 illustrates the AUCs across five datasets under various configurations. Specifically, Seed alone yields the worst verification performance across all datasets, as it only reflects the initial fingerprinting step without other components. Seed_s leads to a clear improvement (e.g., from 0.691 to 0.748 on CIFAR-10), demonstrating the necessity of retaining only the most discriminative fingerprints. Furthermore, extending fingerprint seeds into composite-sample fingerprints (Com_ft) enhances uniqueness by capturing richer decision-region behavior through class-diverse outputs. Building upon this, applying fingerprint selection (Com_s_ft) further strengthens discriminative capability by filtering out less effective fingerprints. More importantly, using fingerprint-specific thresholds also results in performance improvement, for example, improving the AUC from 0.812 (Com_s_ft) to 0.893 (IrisFP) on CIFAR-10, demonstrating the advantage of adapting to a specific threshold for each composite-sample fingerprint. In addition, IrisFP consistently outperforms all ablated variants across datasets. These results highlight the complementary strengths of each component and the necessity of integrating all of them in our design for reliable ownership verification.

Further hyperparameter analysis and additional experimental results are presented in the supplementary material.

5. Conclusions

This paper studies the problem of enhancing both uniqueness and robustness for adversarial-example-based model fingerprinting. To address this issue, we propose IrisFP, a new fingerprinting method that leverages multi-boundary characteristics, multi-sample behavior, and fingerprint quality assessment. Specifically, IrisFP crafts fingerprint seeds positioned near the intersection of multi-decision boundaries to increase prediction margins. It then applies subtle and diverse perturbations to each seed to generate multiple variants, forming a composite-sample fingerprint. The resulting fingerprints are refined using a discriminative selection strategy to retain only those with the strongest separability. For ownership verification, IrisFP employs a two-step process comprising ownership matching and decision aggregation, tailored to the structure of composite-sample fingerprints. Extensive experiments validate the effectiveness of IrisFP, with results consistently showing superior performance over state-of-the-art methods.

References

- [1] Xiaofan Bai, Chaoxiang He, Xiaojing Ma, Bin Benjamin Zhu, and Hai Jin. Intersecting-boundary-sensitive fingerprinting for tampering detection of DNN models. In *Proceedings of Forty-first International Conference on Machine Learning (ICML'24)*, 2024. 2
- [2] Xiaofan Bai, Shixin Li, Xiaojing Ma, Bin Benjamin Zhu, Dongmei Zhang, and Linchen Yu. SDBF: Steep-decision-boundary fingerprinting for hard-label tampering detection of DNN models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR'25)*, pages 29278–29287, 2025. 2
- [3] Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. IP-Guard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security (AsiaCCS'21)*, 2021. 1, 5
- [4] Jialuo Chen, Jingyi Wang, Tinglan Peng, Youcheng Sun, Peng Cheng, Shouling Ji, Xingjun Ma, Bo Li, and Dawn Xiaodong Song. Copy, Right? A testing framework for copyright protection of deep learning models. In *Proceedings of IEEE Symposium on Security and Privacy (S&P'22)*, pages 824–841, 2022. 1
- [5] Yufei Chen, Chao Shen, Cong Wang, and Yang Zhang. Teacher model fingerprinting attacks against transfer learning. In *Proceedings of 31st USENIX Security Symposium (USESec'22)*, pages 3593–3610, Boston, MA, 2022. 1, 5
- [6] Augustin Godinot, Erwan Le Merrer, Camilla Penzo, François Taïani, and Gilles Trédan. Queries, representation & detection: The next 100 model fingerprinting schemes. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'25)*, pages 16817–16825, 2025. 1, 5, 6
- [7] Jiyang Guan, Jian Liang, and Ran He. Are you stealing my model? sample correlation for fingerprinting deep neural networks. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS'22)*, 2022. 5
- [8] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *ICLR'16*, 2016. 3
- [9] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High accuracy and high fidelity extraction of neural networks. In *29th USENIX security symposium (USESec'20)*, pages 1345–1362, 2020. 1
- [10] Hengrui Jia, Hongyu Chen, Jonas Guan, Ali Shahin Shamsabadi, and Nicolas Papernot. A zest of lime: Towards architecture-independent model distances. In *International Conference on Learning Representations (ICLR'21)*, 2021. 5
- [11] Alex Krizhevsky, Vinod Nair, Geoffrey Hinton, et al. The CIFAR-10 dataset. online: <http://www.cs.toronto.edu/kriz/cifar.html>, 55(5):2, 2014. 5
- [12] Alex Krizhevsky, Vinod Nair, Geoffrey Hinton, et al. The CIFAR-100 dataset. online: <http://www.cs.toronto.edu/kriz/cifar.html>, 55(5):2, 2014. 5
- [13] Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015. 5
- [14] Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>, 2, 2010. 5
- [15] Fangqi Li, Shilin Wang, and Lei Yang. Rethinking the fragility and robustness of fingerprints of deep neural networks. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'25)*, pages 1–5. IEEE, 2025. 1
- [16] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In *ICLR'17*, 2017. 3
- [17] Honglin Li, Chenglu Zhu, Yunlong Zhang, Yuxuan Sun, Zhongyi Shui, Wenwei Kuang, Sunyi Zheng, and Lin Yang. Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*, pages 7454–7463, 2023. 3
- [18] Jiacheng Liang, Ren Pang, Changjiang Li, and Ting Wang. Model extraction attacks revisited. In *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security (AsiaCCS'24)*, pages 1231–1245, 2024. 1
- [19] Zhangting Lin, Mingfu Xue, Kewei Chen, Wenmao Liu, Xiang Gao, Leo Yu Zhang, Jian Wang, and Yushu Zhang. Adversarial example based fingerprinting for robust copyright protection in split learning. *arXiv preprint arXiv:2503.04825*, 2025. 1
- [20] Weixing Liu and Shenghua Zhong. MarginFinger: Controlling generated fingerprint distance to classification boundary using conditional gans. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR'24)*, 2024. 1, 6
- [21] Nils Lukas, Yuxuan Zhang, and Florian Kerschbaum. Deep neural network fingerprinting by conferrable adversarial examples. In *International Conference on Learning Representations (ICLR'21)*, 2021. 1, 5, 6
- [22] Thibault Maho, Teddy Furon, and Erwan Le Merrer. Fingerprinting classifiers with benign inputs. *IEEE Transactions on Information Forensics and Security (TIFS)*, 18:5459–5472, 2023. 5
- [23] Zirui Peng, Shaofeng Li, Guoxing Chen, Cheng Zhang, Haojin Zhu, and Minhui Xue. Fingerprinting deep neural networks globally via universal adversarial perturbations. In *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, pages 13420–13429, 2022. 1, 5, 6
- [24] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems (NeurIPS'19)*, 32, 2019. 3
- [25] Shuo Shao, Haozhe Zhu, Yiming Li, Hongwei Yao, Tianwei Zhang, and Zhan Qin. Fit-print: Towards false-claim-resistant model ownership verification via targeted fingerprint. *arXiv preprint arXiv:2501.15509*, 2025. 1
- [26] Gowthami Somepalli, Liam Fowl, Arpit Bansal, Ping Yeh-Chiang, Yehuda Dar, Richard Baraniuk, Micah Goldblum,

- and Tom Goldstein. Can neural nets learn the same model twice? investigating reproducibility and double descent from the decision boundary perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, pages 13699–13708, 2022. 5
- [27] Yuzheng Wang, Zhaoyu Chen, Dingkang Yang, Pinxue Guo, Kaixun Jiang, Wenqiang Zhang, and Lizhe Qi. Out of thin air: Exploring data-free adversarial robustness distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'24)*, pages 5776–5784, 2024. 3
- [28] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *ArXiv*, abs/1708.07747, 2017. 5
- [29] Tianlong Xu, Chen Wang, Gaoyang Liu, Yang Yang, Kai Peng, and Wei Liu. United we stand, divided we fall: Fingerprinting deep neural networks via adversarial trajectories. In *Proceedings of Neural Information Processing Systems (NeurIPS'24)*, 2024. 6
- [30] Tianlong Xu, Chen Wang, Gaoyang Liu, Yang Yang, Kai Peng, and Wei Liu. United we stand, divided we fall: Fingerprinting deep neural networks via adversarial trajectories. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS'24)*, 2024. 1, 5
- [31] Anli Yan, Huali Ren, Kanghua Mo, Zhenxin Zhang, Shaowei Wang, and Jin Li. Enhancing model intellectual property protection with robustness fingerprint technology. *IEEE Transactions on Information Forensics and Security (TIFS)*, 2025. 5
- [32] Guang Yang, Ziyue Geng, Yihang Chen, and Changqing Luo. Fingerprinting deep neural networks for ownership protection: An analytical approach. In *The Fourteenth International Conference on Learning Representations (ICLR'26)*, 2026. 1
- [33] Guang Yang, Ziyue Geng, Yihang Chen, and Changqing Luo. Liteguard: Efficient task-agnostic model fingerprinting with enhanced generalization. In *The Fourteenth International Conference on Learning Representations (ICLR'26)*, 2026. 1
- [34] Kan Yang and Kunhao Lai. NaturalFinger: Generating natural fingerprint with generative adversarial networks. *arXiv preprint arXiv:2305.17868*, 2023. 6
- [35] Xiaoyu You, Youhe Jiang, Jianwei Xu, Mi Zhang, and Min Yang. GNNFingers: A fingerprinting framework for verifying ownerships of graph neural networks. In *Proceedings of the ACM on Web Conference (WWW'24)*, 2024. 5
- [36] Zhuomeng Zhang, Fangqi Li, Hanyi Wang, and Shi-Lin Wang. Boosting the uniqueness of neural networks fingerprints with informative triggers. In *Proceedings of The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS'25)*, 2025. 5
- [37] Boyao Zhao, Haozhe Chen, Jie Zhang, Weiming Zhang, and Nenghai Yu. Dual-verification-based model fingerprints against ambiguity attacks. *Cybersecurity*, 7(1):78, 2024. 1, 5
- [38] Jingjing Zhao, Qingyue Hu, Gaoyang Liu, Xiaoqiang Ma, Fei Chen, and Mohammad Mehedi Hassan. AFA: Adversarial fingerprinting authentication for deep neural networks. *Computer Communications*, 150:488–497, 2020. 1, 5
- [39] Yue Zheng, Si Wang, and Chip-Hong Chang. A dnn fingerprint for non-repudiable model ownership identification and piracy detection. *IEEE Transactions on Information Forensics and Security (TIFS)*, 17:2977–2989, 2022. 5
- [40] Xiaoling Zhou, Wei Ye, Zhemg Lee, Rui Xie, and Shikun Zhang. Boosting model resilience via implicit adversarial data augmentation. *arXiv preprint arXiv:2404.16307*, 2024. 3
- [41] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 3
- [42] Wei Zong, Yang-Wai Chow, Willy Susilo, Joonsang Baek, Jongkil Kim, and Seyit Camtepe. IPRemover: A generative model inversion attack against deep neural network fingerprinting and watermarking. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'24)*, pages 7837–7845, 2024. 3