Misclassification-driven Fingerprinting for DNNs Using Frequency-aware GANs

Weixing Liu[1,2] , **Shenghua Zhong**[1,*] ,
[1]College of Computer Science and Software Engineering, Shenzhen University
[2]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University
liuweixing2022@email.szu.edu.cn, csshzhong@szu.edu.cn

Abstract

Deep neural networks (DNNs) have become valuable assets due to their success in various tasks, but their high training costs also make them targets for model theft. Fingerprinting techniques are commonly used to verify model ownership, but existing methods either require training many additional models, leading to increased costs, or rely on GANs to generate fingerprints near decision boundaries, which may compromise image quality. To address these challenges, we propose a GAN-based fingerprint generation method that applies frequency-domain perturbations to normal samples, effectively creating fingerprints. This approach not only resists intellectual property (IP) threats, but also improves fingerprint acquisition efficiency while maintaining high imperceptibility. Extensive experiments demonstrate that our method achieves a state-of-the-art (SOTA) AUC of 0.98 on the Tiny-ImageNet dataset under IP removal attacks, outperforming existing methods by 8%, and consistently achieves the best ABP for three types of IP detection and erasure attacks on the GTSRB dataset. Our source code is available at https://github.com/wason981/Frequency-Fingerprinting.

1 Introduction

In recent years, DNNs have been successfully applied to a wide range of tasks, including image classification [Zhu *et al.*, 2024], image generation [Huang *et al.*, 2024], and object recognition [Yuan *et al.*, 2024]. However, high-performing DNNs typically rely on substantial amounts of high-quality data, computational resources, and expert knowledge, which makes them critical and valuable assets. Malicious attackers can exploit various stealing strategies, such as transfer learning [Zhuang *et al.*, 2021], to illicitly acquire these models (referred to as source models) for personal or financial gain. To protect the IP of the model owners, model IP protection technologies have been developed to enhance the trustworthiness of models.

*Corresponding author

Current IP protection methods can be broadly classified into watermarking and fingerprinting techniques [Sun *et al.*, 2023]. Watermarking techniques involve embedding watermark information into the source model through training or fine-tuning. However, they often face significant challenges to achieve a balance between the robustness of the watermark and the precision of the model [Jia *et al.*, 2021]. Unlike watermarking techniques, fingerprinting techniques extract unique properties of a model as its "fingerprint", enabling the differentiation of models based on the characters of the extracted fingerprint.

Recent advances have focused on extracting fingerprints from discrepancies in decision areas between a source model and irrelevant models [Yang *et al.*, 2022a; Lukas *et al.*, 2019]. However, these approaches require training a large number of post-processed versions of the source model and irrelevant models (commonly referred to as surrogate models), and the number of fingerprints that can be extracted is limited. Furthermore, some methods generate fingerprints near decision boundaries [Liu and Zhong, 2024; Yang and Lai, 2023], but the quality of these fingerprints is often difficult to ensure. Therefore, current fingerprinting techniques remain in their early stages and cannot meet the diverse requirements of practical applications, such as efficiency, imperceptibility, and robustness.

To address the issues discussed above, we propose a fingerprint generation method based on generative adversarial networks (GANs). This approach generates fingerprints by applying frequency-domain perturbations to the defender's datasets, effectively resisting IP threats while ensuring high imperceptibility. The core idea of our method stems from the observation that introducing perturbations in the frequency domain can change the representation of an image and subsequently influence the classifier's decision-making process, but without producing a noticeable change in the image that can be observed. Building on this insight, we utilize a frequency-aware GAN to generate perturbations that induce the misclassification of normal samples, which can effectively serve as fingerprints to verify the model ownership.

Our main contributions are summarized as follows.

1. We propose using the generated misclassified samples as fingerprints, addressing the scarcity of misclassified samples, and avoiding unnecessary model training overhead.

2. We propose to generate fingerprints by applying frequency-domain perturbations on the original dataset, and those perturbations can change the representation and subsequently influence the classifier's decision for the image, but without producing a noticeable change that can be observed. Thus, this approach effectively improves the robustness of fingerprints while maintaining strong imperceptibility. This is the first fingerprinting model protection method based on the frequency domain.
3. Extensive experiments demonstrate the robustness of our method against IP threats. Specifically, our method achieves an SOTA AUC of 0.98 in the complex Tiny-ImageNet dataset under IP removal attacks, exceeding current methods by 8%, and consistently achieves the best ABP in three types of IP detection and erasure attacks on the GTSRB dataset.

2 Related Work

In this section, we first briefly review the existing literature on model IP protection methods, which can be broadly categorized as watermarking and fingerprinting [Sun *et al.*, 2023]. Next, we review frequency-based GAN for different applications.

2.1 Deep IP Protection

Inspired by digital image copyright protection techniques, watermarking and fingerprinting are the main methods for deep model IP protection [Sun *et al.*, 2023].

Model Watermarking

Watermarking embeds unique identification information into DNNs via fine-tuning or retraining. Parameter-based watermarking is a white-box approach that embeds watermarks in the DNN original components using a regularizer as proposed by [Uchida *et al.*, 2017]. Parameter-based verification necessitates full local access to the model. Label-based watermarking is a black-box method that embeds watermarks by fine-tuning or retraining the source model on trigger sets, creating a backdoor that enables verification through model predictions alone. Some methods [Sun *et al.*, 2021; Adi *et al.*, 2018; Zhang *et al.*, 2018; Lao *et al.*, 2022] used specific samples from the original training set as the trigger set, while others attached trigger patterns to natural images [Zhang *et al.*, 2018; Guo and Potkonjak, 2018], and yet others synthesized trigger sets through generation [Li *et al.*, 2019] or gradient optimization techniques [Le Merrer *et al.*, 2020; Li *et al.*, 2022]. However, recent studies [Jia *et al.*, 2021] suggested that watermarking can potentially compromise the accuracy of the model due to its impact on the training process.

Model Fingerprinting

Fingerprinting is a non-invasive method that extracts model representations to compare the similarities between different DNNs. Techniques such as IPGuard [Cao *et al.*, 2021] and DFA [Wang and Chang, 2021] focused on extracting fingerprint samples near the decision boundary of the source model. This is based on the idea that a DNN classifier's unique traits lie within this boundary. However, these methods are vulnerable to model extraction attacks that modify the boundary. CAE [Lukas *et al.*, 2019] proposed extracting conferrable adversarial samples that are transferable only to stolen models and not to independently trained irrelevant models. This approach effectively enhanced robustness against model extraction attacks, but required significant training overhead. Despite these advancements, transfer learning remains a challenge. ModelDiff [Li *et al.*, 2021] proposed using the cosine similarity of the decision distance vectors (DDVs) between models on the same inputs to detect transfer learning and model compression attacks. However, it struggled with model extraction detection. Zest [Jia *et al.*, 2022] addressed this shortcoming by using the Local Interpretable Model-Agnostic Explanations (LIME) algorithm [Ribeiro *et al.*, 2016] to generate linear models that approximate local behavior, thus forming a global approximation. By comparing the cosine distances of the linear models' weights, Zest more accurately captured model similarities than direct prediction comparisons. Furthermore, SAC [Guan *et al.*, 2022] introduced a metric to assess the correlation of pairwise samples within the model to defend against adversarial training. In order to mitigate the impact of accuracy degradation, we propose a fingerprinting framework and design a method that avoids the dependence on a significant amount of additional model training. This approach can also effectively resist transfer learning and adversarial training.

2.2 Frequency-Based Generative Adversarial Networks

GANs have demonstrated considerable success in various computer vision tasks, such as quality enhancement [Zhang *et al.*, 2019], image inpainting [Yu *et al.*, 2021], few-shot image generation [Yang *et al.*, 2022b] and image reconstruction [Jiang *et al.*, 2021]. However, studies [Schwarz *et al.*, 2021] have shown that GANs exhibit a spectral bias in the fitting of frequency signals, often failing to capture high-frequency details, which can result in blur and noticeable artifacts. To address this issue, several methods have been developed to incorporate frequency information into GANs. For example, SSD-GAN [Chen *et al.*, 2021] addressed spectral information loss by incorporating a frequency-aware classifier into the discriminator, guiding the generator to learn high-frequency content and improve detail generation. Zhang *et al.* proposed a super-resolution reconstruction algorithm [Zhang *et al.*, 2019] that combined wavelet transform with GANs to enrich high-frequency details in low-resolution images by predicting wavelet coefficients. Furthermore, Huang *et al.* applied this technique to face hallucination, capturing rich contextual information from low-resolution face images captured in the wild [Huang *et al.*, 2019]. Their method achieved better PSNR and SSIM while significantly improving the accuracy of the identification. Jiang *et al.* introduced a focal frequency loss that improved image reconstruction by adaptively focusing on hard-to-synthesize frequencies [Jiang *et al.*, 2021], complementing spatial losses. Meanwhile, WaveFill *et al.* enhanced image inpainting by filling corrupted areas with decomposed frequency components [Yu *et al.*, 2021]. Additionally, Yang *et al.* introduced WaveGAN [Yang

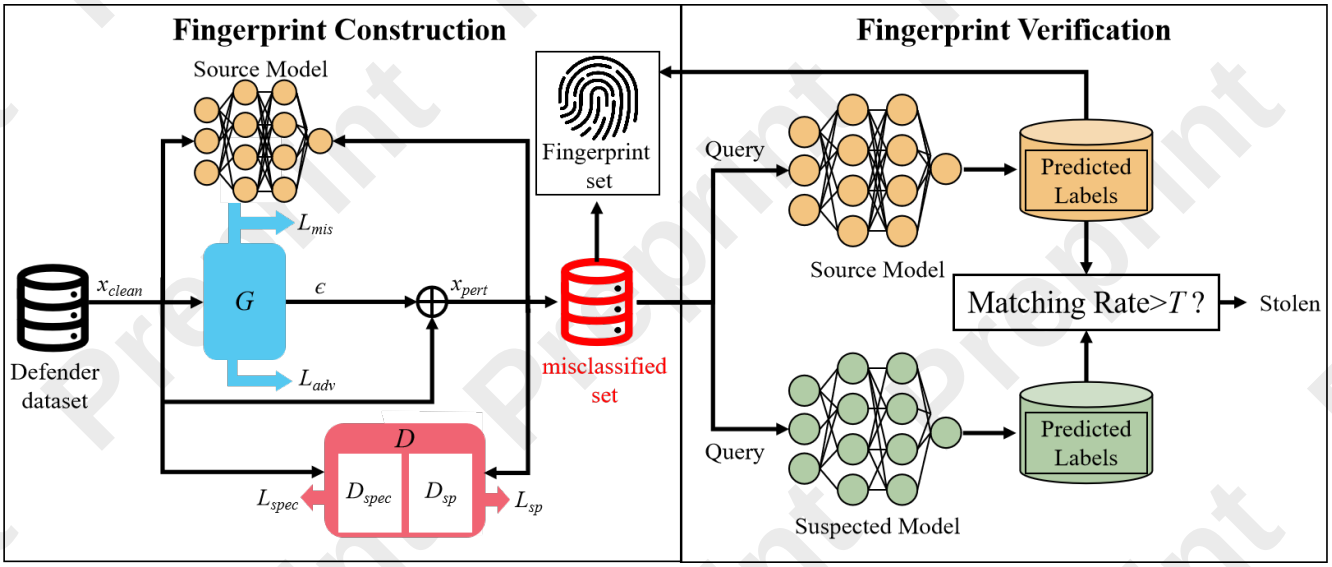

Figure 1: The illustration of our framework: We first generate misclassified set using a frequency-aware GAN. Then, we calculate the matching rate between the predicted labels of the source model $\mathcal{M}_D$ and the suspect model $\mathcal{M}_{suspect}$. Any model with a matching rate greater than the threshold T will be considered a stolen model.

et al., 2022b], a frequency-aware model for the generation of few-shot images, which employed low-frequency skip connections to preserve outline and structural information, and high-frequency skip connections to enhance fine detail synthesis. The success of these methods lies in the fact that images consist of frequency components with different physical meanings and information. The presented methods demonstrate that frequency domain-based adversarial generation can achieve favorable generation results in a variety of tasks. However, related exploration in the domain of model protection is limited. Only Liu *et al.*'s approach [Liu *et al.*, 2024] has discussed the impact of different frequency bands on model protection, arguing that embedding information in mid-low frequency bands strikes a balance between watermark robustness and imperceptibility. However, the potential of frequency-aware techniques for modifying the image representation to obtain fingerprints for the model protection task has remained underexplored. In this paper, we propose a frequency-aware GAN to generate adversarially perturbed fingerprint samples that induce misclassifications, which exhibit strong robustness for model protection and high imperceptibility.

3 Proposed Method

3.1 Problem Definition

This study considers the scenario in which a legitimate developer (defender) invests considerable resources in training a high-performance DNN model, referred to as the source model $\mathcal{M}_D$. The model is typically deployed as a paid service through cloud platforms or client software for user access. However, malicious attackers can use various methods to steal $\mathcal{M}_D$ and create stolen versions $\mathcal{M}_A$ for unauthorized use or resale. To protect the IP of their model, defenders often use fingerprinting or watermarking techniques to verify if a suspect model $\mathcal{M}_{suspect}$ is stolen.

The general pipeline of the proposed Deep IP Fingerprinting Protection model, as illustrated in Figure 1, consists of two stages: fingerprint construction and fingerprint verification. We will describe these two stages as follows.

3.2 Fingerprint Construction

Existing research emphasizes the role of misclassified samples in improving adversarial robustness of a model [Wang *et al.*, 2020]. Inspired by this, we propose using misclassified samples and their predicted labels from the source model as a unique fingerprint for model IP protection. For high-performance classifiers, the limited number of misclassified samples is insufficient for IP verification. To address this limitation, we propose a GAN-based method to produce the misclassified samples. Such an approach allows us to generate an unconstrained number of misclassified samples and is more flexible for model protection.

Misclassified Samples

Given the source model $\mathcal{M}_D$ and a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, we divide $\mathcal{D}$ according to the predictions of h_θ into two subsets: one containing correctly classified samples and the other containing misclassified samples, as shown in Eq. (1) and (2):

$$S^+_{\mathcal{M}_D} = \{x_i : x_i \in \mathcal{D}, \mathcal{M}_D(x_i) = y_i\} \quad (1)$$

$$S^-_{\mathcal{M}_D} = \{x_i : x_i \in \mathcal{D} : \mathcal{M}_D(x_i) \neq y_i\} \quad (2)$$

For each misclassified sample $x_i \in S^-_{\mathcal{M}_D}$, x_i and its prediction $\mathcal{M}_D(x_i)$ can be recorded as the unique fingerprint of the model for subsequent verification. However, for high-performance classifiers, the limited number of misclassified samples is insufficient to enable effective IP verification. To overcome this limitation, we adopt a GAN-based approach

to augment the fingerprint set $\mathcal{D}_f$ by generating misclassified samples.

We design a GAN framework where the generator G generates adversarial perturbations ϵ, which, when added to the original training sample x_{clean} (clean samples), form x_{mis} and induce misclassifications in the source model. Meanwhile, the discriminator D is designed to distinguish between x_{clean} and x_{mis}. Notably, for a given input sample, the generator is capable of generating diverse perturbations that result in misclassification, thereby enabling the creation of a virtually unlimited $\mathcal{D}_f$.

Frequency-Aware Generative Adversarial Network
According to research [Liu *et al.*, 2024], as a watermarking method, perturbations in the frequency domain of clean samples are effective in generating trigger samples. But the watermarking method embeds unique identifiers in the model which may affect the performance or fidelity of the model. In contrast, model fingerprinting is a non-invasive technique. By avoiding model modifications, fingerprinting maintains optimal fidelity and reliability. However, there is no fingerprinting model protection method based on the frequency domain.

The idea of our method comes from the observation that introducing perturbations in the frequency domain can change the representation of an image, thereby affecting the classifier's decision, but without producing a noticeable change in the image that can be observed. Building on these insights, we propose a Frequency-Aware GAN to generate frequency-domain perturbations that induce misclassifications in the source model.

Wavelet Transform Generator: Inspired by [Fu *et al.*, 2021], we use a Discrete Wavelet Transform (DWT) branch as the foundation of our frequency-aware generator and modify its loss function, where the generator loss L_G consists of adversarial loss $\mathcal{L}_{adv}$ and misclassification loss $\mathcal{L}_{mis}$:

$$\mathcal{L}_G = \mathcal{L}_{adv} + \mathcal{L}_{mis} \tag{3}$$

The adversarial loss ensures that the perturbed sample $x_{pert} = x_{clean} + \epsilon$ deceives the discriminator D, thus producing realistic adversarial samples.

$$\mathcal{L}_{adv} = \mathbb{E}_{x_{clean} \sim p_{data}(x)}[\log(1 - D(x_{clean} + \epsilon))] \tag{4}$$

The misclassification loss maximizes the predictive divergence between the clean images from the real data distribution p_{data} and their perturbed counterparts on the source model.

$$\mathcal{L}_{mis} = \mathbb{E}_{x_{clean} \sim p_{data}(x)} CE(\phi(x_{clean}), \phi(x_{clean} + \epsilon)) \tag{5}$$

Here, ϕ represents the soft output of $\mathcal{M}_D$, and CE denotes the cross-entropy loss function.

Fourier Spectrum Discriminator: To enhance the discriminator's ability to distinguish perturbed images, we leverage the discriminator architecture proposed in [Chen *et al.*, 2021]. The Fourier spectrum discriminator focuses on analyzing frequency-domain features, enabling more robust differentiation by capturing fine-grained spectral variations. The discriminator's loss function is formulated as:

$$\begin{aligned}\mathcal{L}_D =& \mathbb{E}_{x_{clean} \sim p_{data}(x)} - [\log D(x_{clean})] \\ &+ \mathbb{E}_{x_{clean} \sim p_{data}(x)} - [\log(1 - D(x_{clean} + \epsilon))]\end{aligned} \tag{6}$$

D is a discriminator consisting of a spatial discriminator D_{sp}, which measures spatial realness, and a spectral discriminator D_{spec}, which measures spectral realness. The overall realness of a sample x is represented as:

$$D(x) = \lambda D_{spec}(x) + (1 - \lambda) D_{sp}(x), \tag{7}$$

where λ is a hyperparameter that controls the relative importance of spatial realness and spectral realness.

3.3 Fingerprint Verification

For a suspect model $\mathcal{M}_{suspect}$, (referred to as query samples) in $\mathcal{D}_f$. Let M be the number of query samples with predicted labels that match those of $\mathcal{M}_D$. The matching rate is M/N. If the matching rate exceeds or equals a predefined threshold $T \in [0, 1]$, the model is considered stolen.

4 Experiments and Results

In this section, we first introduce the various IP threats that our approach aims to address. Next, we provide a detailed description of the experiments. Finally, we compare the experimental results of our method with existing baselines to demonstrate its effectiveness in defending against the aforementioned attacks.

4.1 IP Threats

IP Removal Attack
The attacker removes or forges IP identifiers in the source model to hinder ownership verification while minimizing performance degradation.

- Probability-based Model Extraction (MEP) [Jagielski *et al.*, 2020]: A black-box attack where the attacker queries the source model for probabilistic outputs to train a copy version.
- Label-based Model Extraction (MEL) [Jagielski *et al.*, 2020]: A black-box attack where the attacker queries the source model for hard labels to train a copy version.
- Adversarial Model Extraction (MEA) [Guan *et al.*, 2022]: A black-box attack where the attacker applies adversarial training to the label-based extraction model.
- Weight Pruning (WP) [Blalock *et al.*, 2020]: A white-box attack where the attacker prunes neurons based on their weight to simplify the model.
- Model Fine-tuning (FT) [Uchida *et al.*, 2017]: The attacker fine-tunes (FT) or retrains (RT) the final layer or the entire model to create a stolen version.
- Transfer Learning (TL) [Zhuang *et al.*, 2021]: The attacker transfers knowledge from source model to a related task by retraining.

IP Detection and Erasure Attack
The defender uploads query samples to the attacker's model prediction API for verification, while the attacker actively disrupts the verification process.

- Query Modification [Namba and Sakuma, 2019]: The attacker trains an autoencoder to reconstruct query samples, erasing embedded information.

- Input Smoothing [Xu *et al.*, 2018]: The attacker applies smoothing techniques to reduce adversarial perturbations.
- Feature Squeezing [Xu *et al.*, 2018]: Reduces the pixel color bit depth to minimize the detection of adversarial perturbations.

4.2 Experiment Setup

Datasets and Model Architectures

For better comparison with other works, we adopt the data-split method and model architectures used in [Guan *et al.*, 2022]. To validate the effectiveness and robustness of our method, we conduct experiments on CIFAR-10, GTSRB, and Tiny-ImageNet. The training dataset is split into two parts: $D_{Defender}$ for the defender and $D_{Attacker}$ for the attacker, simulating model IP attack defense scenarios.

- CIFAR-10: consists of 60K 32 $\times$ 32 color images in 10 distinct classes, with a training set of 50K images and a test set of 10K images.
- GTSRB: contains over 50K images of German traffic signs in 43 classes, with 39K images for training and 12K images for test. The images are resized to 32 $\times$ 32 pixels for evaluation.
- Tiny-ImageNet: contains 110K 64 $\times$ 64 images of 200 different object classes, with 100K images for training and 10K images for test.

Similar to the settings in [Guan *et al.*, 2022; Lukas *et al.*, 2019; Cao *et al.*, 2021], we select the commonly used model VGG16 as the source model; VGG13, ResNet18, DenseNet121, MobileNetV2 as irrelevant models. For each setting, we train five models under each stealing attack and average the results across these models to mitigate the impact of randomness.

Comparison Approaches and Evaluation Metrics

We compare our proposed method with nine state-of-the-art fingerprint protection schemes: IPGuard [Cao *et al.*, 2021], ModelDiff [Li *et al.*, 2021], DI [Maini *et al.*, 2021], SAC-w [Guan *et al.*, 2022], SAC-m [Guan *et al.*, 2022], Zest [Jia *et al.*, 2022], IBSF [Bai *et al.*, 2024], MarginFinger [Liu and Zhong, 2024], and MetaFinger [Yang *et al.*, 2022a].

Based on previous research, we choose AUC (Area Under the Curve) as the metric to evaluate model protection methods. The AUC value ranges from 0 to 1, with higher values indicating a stronger ability to distinguish between stolen models and irrelevant models.

4.3 Robustness Against Model IP Removal Attack

As illustrated in Table 1, we evaluate our method on CIFAR10, GTSRB, and Tiny-ImageNet. The experimental results demonstrate that our method is highly robust against the aforementioned stealing techniques, achieving the highest or second highest AUC values. In particular, in the complex classification task on the Tiny-ImageNet dataset, the average AUC of our method reaches 0.98, significantly outperforming other compared methods. This can be attributed to the fact that in complex datasets, the classification boundaries between models vary greatly. The stolen model's classification boundary is typically closer to the source model than to irrelevant models, leading to similar classification results on misclassified samples. In contrast, for simple datasets, such as GTSRB, the classification boundaries of the source and irrelevant models are also similar, thereby limiting the influence of misclassified samples.

Dataset	Methods	MEP	MEL	MEA	WP	FT	TL	AVG
CIFAR-10	IPGuard	0.60	0.55	0.52	**1.00**	**1.00**	**1.00**	0.78
	SAC-w	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	SAC-m	**1.00**	0.81	0.76	**1.00**	**1.00**	**1.00**	0.93
	MarginFinger	**1.00**	**1.00**	**1.00**	0.87	**1.00**	**1.00**	0.98
	zest	**1.00**	0.81	0.80	**1.00**	**1.00**	**1.00**	0.94
	ModelDiff	0.44	0.55	0.38	**1.00**	**1.00**	**1.00**	0.73
	DI	0.84	0.64	0.54	**1.00**	**1.00**	**1.00**	0.84
	IBSF	**1.00**	0.46	0.10	**1.00**	**1.00**	0.62	0.70
	MetaFinger	0.84	**1.00**	0.99	**1.00**	**1.00**	0.48	0.89
	Proposed	0.98	0.89	0.68	**1.00**	**1.00**	**1.00**	0.93
GTSRB	IPGuard	0.61	0.41	0.31	**1.00**	0.68	0.12	0.52
	SAC-w	0.40	0.56	0.30	0.85	0.40	0.13	0.44
	SAC-m	**1.00**	0.29	0.02	**1.00**	0.41	**1.00**	0.62
	MarginFinger	0.92	0.49	0.16	**1.00**	0.79	0.88	**0.71**
	zest	**1.00**	0.32	0.18	**1.00**	0.87	0.01	0.56
	ModelDiff	0.69	0.48	0.25	**1.00**	0.56	0.34	0.55
	DI	0.27	**0.74**	0.34	**1.00**	**1.00**	0.00	0.55
	IBSF	0.01	0.68	**0.54**	**1.00**	0.37	0.16	0.46
	MetaFinger	0.92	0.31	0.31	**1.00**	**1.00**	0.43	0.66
	Proposed	0.68	0.52	0.12	**1.00**	0.94	**1.00**	**0.71**
Tiny-ImageNet	IPGuard	0.45	0.45	0.45	**1.00**	**1.00**	0.35	0.62
	SAC-w	0.18	0.53	0.21	**1.00**	**1.00**	0.04	0.50
	SAC-m	0.54	0.94	0.37	**1.00**	**1.00**	0.00	0.64
	MarginFinger	**1.00**	0.98	0.70	0.80	**1.00**	**1.00**	0.91
	zest	0.93	0.96	0.81	**1.00**	**1.00**	0.51	0.87
	ModelDiff	0.09	0.16	0.40	**1.00**	**1.00**	0.99	0.61
	DI	0.61	0.75	0.78	**1.00**	**1.00**	0.00	0.69
	IBSF	0.25	0.50	0.25	**1.00**	**1.00**	0.24	0.54
	MetaFinger	0.84	**1.00**	0.99	**1.00**	**1.00**	0.48	0.89
	Proposed	**1.00**	**1.00**	**0.89**	**1.00**	**1.00**	0.99	**0.98**

Table 1: The AUC of different IP protection methods when facing six IP removal attacks on three datasets. (BOLD IS THE BEST)

MarginFinger exhibits the greatest performance in CIFAR-10 and ranks second on GTSRB and Tiny-ImageNet. This is because MarginFinger generates fingerprints by controlling the distance between the fingerprint and the classification boundary, effectively differentiating the classification results of the source and stolen models from those of irrelevant models. However, its performance declines for more complex tasks, as the GAN in MarginFinger is trained on the defender's dataset. Complex and small-scale datasets lead to insufficient GAN training, preventing it from effectively controlling the distance of the generated fingerprint relative to the classification boundary. In contrast, although our method is also trained on the defender dataset, it only requires GAN to generate perturbations on samples that can trigger misclassifications of the classifier. This simplicity allows us to maintain a strong fingerprint protection ability even in complex tasks. SAC-m performs well overall; however, its reliance on the correlation matrix of predicted labels between fingerprint samples makes it vulnerable to model extraction attacks that affect classification boundaries. Even a single alteration of the label of a fingerprint sample can significantly disrupt the correlation matrix, weakening its ability to resist model extraction attacks.

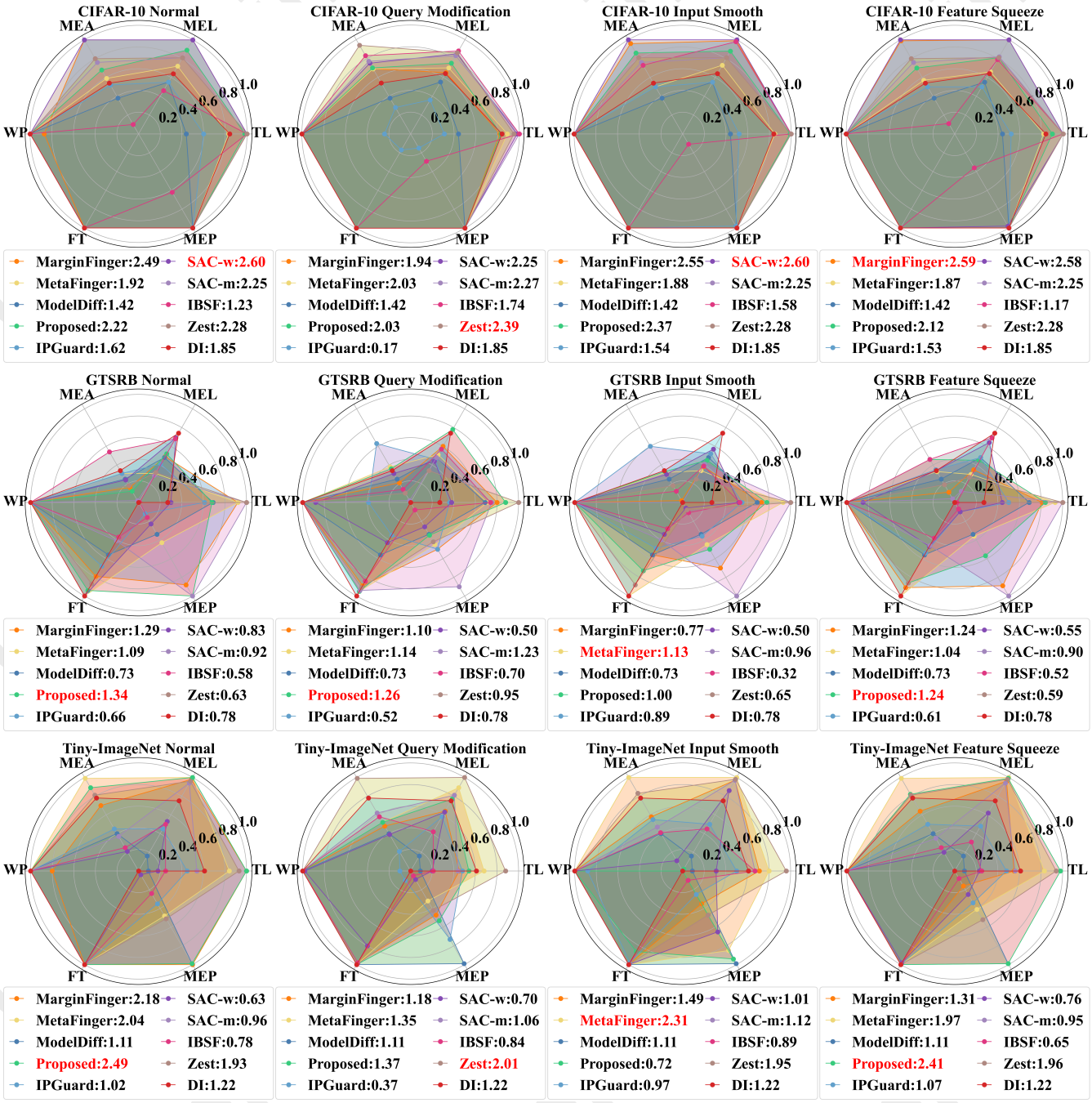

Figure 2: Evaluation of different methods before (Normal) and after three types of IP detection and erasure attacks on CIFAR-10. We compare the performance of different methods based on ABP (RED IS THE BEST).

4.4 Robustness Against Model IP Detection and Evasion Attack

As shown in Figure 2, we discuss the robustness of different methods on three datasets (CIFAR-10, GTSRB, and Tiny-ImageNet) under three IP detection and erasure attacks. We assume that the attacker, being aware of the process described in Section 3.3, employs IP detection and erasure attacks to corrupt query samples, preventing the model owner from collecting evidence. Therefore, it is essential to ensure the high quality and imperceptibility of the query samples. For each method, we first evaluate the AUC values under six IP removal attacks using unmodified query samples (normal). These six AUC values serve as the vertices of a polygon to construct a radar chart. Subsequently, we apply three types of IP detection and erasure attacks on the query samples and recalculated the AUC values under the six IP removal attacks. The area bounded by the polygon (ABP) is used to compare the performance of different methods against IP detection and erasure attacks. A higher ABP value indicates stronger robustness.

In the CIFAR-10 dataset, SAC-w, Zest, and MarginFinger exhibit strong robustness, while our method performs moderately with an ABP of approximately 2. In contrast, IPGuard and ModelDiff show weaker performance, scoring below 1.5. In the GTSRB dataset, where all methods generally score below 1.3, our method consistently outperforms others in ABP and achieves SOTA performance. On the more complex Tiny-ImageNet dataset, our method and Zest obtain the best results, demonstrating strong adaptability. Across the three detection and erasure attack types, SAC series and Zest remain stable under query modification, while our method shows slight declines. Under input smoothing and feature squeezing, both Zest and our method maintain high ABP scores, outperforming the SAC series. Notably, Zest's query samples use a masking technique, which causes parts of the main content to be lost. After applying query modification and input smoothing attacks, the image content is restored, leading to an increased ABP value compared to the original query samples.

5 Conclusion

In this paper, we propose a robust and imperceptible fingerprint generation method. Inspired by the adversarial robustness of misclassified samples, we utilize them as fingerprints and address their scarcity by introducing a frequency-aware GAN to generate frequency-domain perturbations for normal samples, creating misclassified fingerprints. Our method offers the following advantages: 1) **Efficiency**: Training a single GAN enables the efficient generation of numerous fingerprints. 2) **Imperceptibility**: The fingerprints are high-quality images and achieve the best ABP across three IP detection and erasure attacks on the GTSRB dataset. 3) **Robustness**: Achieves a SOTA AUC of 0.98 on Tiny-ImageNet under IP removal attacks, outperforming existing methods by 8%. Therefore, our method provides an effective solution to protect model IP.

In future work, we will apply frequency-domain perturbation fingerprinting techniques to other fields, such as natural language processing, brain-computer interfaces, etc.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (62472291), Guangdong Basic and Applied Basic Research Foundation (2025A1515012154, 2023A1515012685, 2023A1515011296), Open Fund of National Engineering Laboratory for Big Data System Computing Technology (Grant No. SZU-BDSC-OF2024-14).

References

[Adi *et al.*, 2018] Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th USENIX security symposium (USENIX Security 18)*, pages 1615–1631, 2018.

[Bai *et al.*, 2024] Xiaofan Bai, Chaoxiang He, Xiaojing Ma, Bin B. Zhu, and Hai Jin. Intersecting-boundary-sensitive fingerprinting for tampering detection of dnn models. In *International Conference on Machine Learning*, 2024.

[Blalock *et al.*, 2020] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? *Proceedings of machine learning and systems*, 2:129–146, 2020.

[Cao *et al.*, 2021] Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Ipguard: 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*, pages 14–25, 2021.

[Chen *et al.*, 2021] Yuanqi Chen, Ge Li, Cece Jin, Shan Liu, and Thomas Li. Ssd-gan: measuring the realness in the spatial and spectral domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1105–1112, 2021.

[Fu *et al.*, 2021] Minghan Fu, Huan Liu, Yankun Yu, Jun Chen, and Keyan Wang. Dw-gan: A discrete wavelet transform gan for nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–212, 2021.

[Guan *et al.*, 2022] Jiyang Guan, Jian Liang, and Ran He. Are you stealing my model? sample correlation for fingerprinting deep neural networks. *Advances in Neural Information Processing Systems*, 35:36571–36584, 2022.

[Guo and Potkonjak, 2018] Jia Guo and Miodrag Potkonjak. Watermarking deep neural networks for embedded systems. In *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 1–8. IEEE, 2018.

[Huang *et al.*, 2019] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet domain generative adversarial network for multi-scale face hallucination. *Int. J. Comput. Vision*, 127(6–7):763–784, June 2019.

[Huang *et al.*, 2024] Muqi Huang, Chaoyue Wang, Yong Luo, and Lefei Zhang. Eliminating the cross-domain misalignment in text-guided image inpainting. In Kate

Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 875–883. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track.

[Jagielski *et al.*, 2020] 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 (USENIX Security 20)*, pages 1345–1362, 2020.

[Jia *et al.*, 2021] Hengrui Jia, Christopher A Choquette-Choo, Varun Chandrasekaran, and Nicolas Papernot. Entangled watermarks as a defense against model extraction. In *30th USENIX security symposium (USENIX Security 21)*, pages 1937–1954, 2021.

[Jia *et al.*, 2022] 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*, 2022.

[Jiang *et al.*, 2021] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13899–13909, 2021.

[Lao *et al.*, 2022] Yingjie Lao, Peng Yang, Weijie Zhao, and Ping Li. Identification for deep neural network: Simply adjusting few weights! In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 1328–1341. IEEE, 2022.

[Le Merrer *et al.*, 2020] Erwan Le Merrer, Patrick Perez, and Gilles Trédan. Adversarial frontier stitching for remote neural network watermarking. *Neural Computing and Applications*, 32(13):9233–9244, 2020.

[Li *et al.*, 2019] Zheng Li, Chengyu Hu, Yang Zhang, and Shanqing Guo. How to prove your model belongs to you: A blind-watermark based framework to protect intellectual property of dnn. In *Proceedings of the 35th annual computer security applications conference*, pages 126–137, 2019.

[Li *et al.*, 2021] Yuanchun Li, Ziqi Zhang, Bingyan Liu, Ziyue Yang, and Yunxin Liu. Modeldiff: Testing-based dnn similarity comparison for model reuse detection. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 139–151, 2021.

[Li *et al.*, 2022] Yiming Li, Yang Bai, Yong Jiang, Yong Yang, Shu-Tao Xia, and Bo Li. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. *Advances in Neural Information Processing Systems*, 35:13238–13250, 2022.

[Liu and Zhong, 2024] 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*, pages 129–136, 2024.

[Liu *et al.*, 2024] Yong Liu, Hanzhou Wu, and Xinpeng Zhang. Robust and imperceptible black-box dnn watermarking based on fourier perturbation analysis and frequency sensitivity clustering. *IEEE Transactions on Dependable and Secure Computing*, 21(6):5766–5780, 2024.

[Lukas *et al.*, 2019] Nils Lukas, Yuxuan Zhang, and Florian Kerschbaum. Deep neural network fingerprinting by conferrable adversarial examples. *arXiv preprint arXiv:1912.00888*, 2019.

[Maini *et al.*, 2021] Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. Dataset inference: Ownership resolution in machine learning. *arXiv preprint arXiv:2104.10706*, 2021.

[Namba and Sakuma, 2019] Ryota Namba and Jun Sakuma. Robust watermarking of neural network with exponential weighting. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, Asia CCS '19, page 228–240, New York, NY, USA, 2019. Association for Computing Machinery.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[Schwarz *et al.*, 2021] Katja Schwarz, Yiyi Liao, and Andreas Geiger. On the frequency bias of generative models. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc.

[Sun *et al.*, 2021] Zhichuang Sun, Ruimin Sun, Long Lu, and Alan Mislove. Mind your weight (s): A large-scale study on insufficient machine learning model protection in mobile apps. In *30th USENIX security symposium (USENIX security 21)*, pages 1955–1972, 2021.

[Sun *et al.*, 2023] Yuchen Sun, Tianpeng Liu, Panhe Hu, Qing Liao, Shaojing Fu, Nenghai Yu, Deke Guo, Yongxiang Liu, and Li Liu. Deep intellectual property protection: A survey. *arXiv preprint arXiv:2304.14613*, 2023.

[Uchida *et al.*, 2017] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, pages 269–277, 2017.

[Wang and Chang, 2021] Si Wang and Chip-Hong Chang. Fingerprinting deep neural networks-a deepfool approach. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2021.

[Wang *et al.*, 2020] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020.

[Xu *et al.*, 2018] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep

neural networks. In *Proceedings 2018 Network and Distributed System Security Symposium*, NDSS 2018. Internet Society, 2018.

[Yang and Lai, 2023] Kang Yang and Kunhao Lai. Naturalfinger: Generating natural fingerprint with generative adversarial networks. *arXiv preprint arXiv:2305.17868*, 2023.

[Yang *et al.*, 2022a] Kang Yang, Run Wang, and Lina Wang. Metafinger: Fingerprinting the deep neural networks with meta-training. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 776–782. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[Yang *et al.*, 2022b] Mengping Yang, Zhe Wang, Ziqiu Chi, and Wenyi Feng. Wavegan: Frequency-aware gan for high-fidelity few-shot image generation. *ArXiv*, abs/2207.07288, 2022.

[Yu *et al.*, 2021] Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, and Chunyan Miao. Wavefill: A wavelet-based generation network for image inpainting. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14094–14103, 2021.

[Yuan *et al.*, 2024] Yao Yuan, Wutao Liu, Pan Gao, Qun Dai, and Jie Qin. Unified unsupervised salient object detection via knowledge transfer. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024.

[Zhang *et al.*, 2018] Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph. Stoecklin, Heqing Huang, and Ian Molloy. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, ASIACCS '18, page 159–172, New York, NY, USA, 2018. Association for Computing Machinery.

[Zhang *et al.*, 2019] Qi Zhang, Huafeng Wang, Tao Du, Sichen Yang, Yuehai Wang, Zhiqiang Xing, Wenle Bai, and Yang Yi. Super-resolution reconstruction algorithms based on fusion of deep learning mechanism and wavelet. In *Proceedings of the 2nd International Conference on Artificial Intelligence and Pattern Recognition*, AIPR '19, page 102–107, New York, NY, USA, 2019. Association for Computing Machinery.

[Zhu *et al.*, 2024] Qikui Zhu, Chuan Fu, and Shuo Li. Class-consistent contrastive learning driven cross-dimensional transformer for 3d medical image classification. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024.

[Zhuang *et al.*, 2021] 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, 2021.