

# Supplementary Material of Reliable Model Watermarking: Defending Against Theft without Compromising on Evasion

## 1 OVERVIEW

The code and detailed hyperparameter settings for this paper are accessible at the following anonymous repository <https://anonymous.4open.science/r/anonymousMM2024-C828>. In the subsequent sections, we offer an extensive supplement to our paper, concentrating on three crucial aspects:

- Section 2 details the specific process of UAE Selection as outlined in the main text.
- Section 3 examines the ablation studies, which are designed to assess the effectiveness of individual components and their contributions to the overall performance of friendly teacher learning procedure.
- Section 4 offers a discussion on the practicality and applicability of UAE watermarking in real-world scenarios and over the long term.

## 2 DETAILED PROCEDURE OF UAE SELECTION

Conditional diffusion models can efficiently generate UAEs. However, due to imperfect density estimation of the main task distribution, it is still possible to generate noise samples that do not semantically belong to the source class. Constructing trigger set with these samples may still embed harmful shortcuts in the model or make it susceptible to anomaly detection. Therefore, we propose utilizing sample quality assessment [5] to filter out low-quality UAEs. Specifically, we first project the samples  $x$  into a low-dimensional feature space with a pre-trained feature extractor  $f_e$ , such as CLIP [11]. The feature  $\gamma$  is then obtained as  $\gamma = f_e(x)$ . Subsequently, we train a Gaussian Mixture Model (GMM) to estimate the density of features in the training set as shown in Equation 1:

$$p(x) \approx \sum_{i=1}^N \pi_i \cdot \mathcal{N}(\gamma | \mu_i, \Sigma_i), \quad (1)$$

where  $N$  is the number of Gaussian components,  $\pi_i$  represents the mixture coefficients, and  $\mu_i$  and  $\Sigma_i$  denote the mean and covariance, respectively. In the feature space, we utilize the GMM to compare the proximity of candidate UAEs to the original distribution and discard the samples with the lowest scores  $p(x)$ .

Additionally, in functionality stealing, attackers may design training procedures and model architectures differ from the protected model. Therefore, it is crucial to ensure that the UAEs generated on surrogate models are sufficiently challenging for a wide range of model families, highlighting the importance of the transferability of UAEs in trigger set construction. Although diffusion models have been shown to synthesize UAEs with high transferability [2, 14], there is a risk that aggressive UAEs may be excluded due to quality-based filtering. Inspired by ghost networks [7], we design two randomization strategies to efficiently select highly transferable UAEs from the remaining candidate set.

The first strategy involves inserting dropout after each parameterized layer [12], while the second one introduces small Gaussian noise into the parameters of the model [1].

In the dropout strategy, for layer  $l$  and its input  $x_l$  with function  $f_l$ , we redefine the function of layer  $l$  as follows [7]:

$$g_l(x) = f_l \left( \frac{r * x_l}{p} \right), \quad (2)$$

$$r \sim \text{Bernoulli}(p), \quad (3)$$

where  $*$  denotes an element-wise product and each element in  $r$  is independently set to 1 with probability  $p$ .

In the random noise strategy, the corresponding new function  $k_l$  is defined as:

$$k_l(x_l) = f_l(x_l; \theta_l + \epsilon), \quad (4)$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2 I), \quad (5)$$

where  $\theta_l$  represents the parameters of layer  $l$  and  $\epsilon$  represents Gaussian noise.  $\mathcal{N}(0, \sigma^2 I)$  denotes a Gaussian distribution with a mean of 0 and a covariance matrix  $\sigma^2 I$ , where  $I$  is the identity matrix.

These two randomization methods simulate minor variations in both model architecture and model parameters. We utilize these two strategies to generate  $M$  shadow models  $S$  (the  $i$ -th model denoted as  $S_i$ ) from a single model and average their outputs on the candidate UAE set by the following Equation 6:

$$\bar{y}(x) = \frac{1}{M} \sum_{i=1}^M S_i(x). \quad (6)$$

We then select the UAEs with the lowest average confidence in the correct class as the final trigger set. In the complete trigger set construction process, we first call the UAE generation module to generate predefined amount of samples that successfully deceive the surrogate model, then evaluate their quality, remove ineligible samples, and finally compare the transferability of the remaining samples, retaining only the UAEs that meet the trigger set size requirements.

## 3 ABLATION STUDY

In this section, we examine the impact of various components of the "Friendly Teacher" on the watermark success rate. Table 1 illustrates the resistance of watermarks to removal attacks across different training procedures on the CIFAR-10 dataset, ranging from vanilla training method (denoted as Normal Teacher) to optimized function mapping properties, optimized output distribution properties, and finally the complete "Friendly Teacher" training procedure with sharpness-aware minimization. It is evident that even on the simple CIFAR-10 dataset, vanilla training procedure do not offer adequate watermark unremovability. Optimizing function mapping and output distribution properties significantly enhances the transferability of UAE watermark behavior from protected models to extraction surrogates and simultaneously increases resistance to model modification (fine-pruning). This enhancement results from improved function mapping properties, which enable the model to more effectively acquire new knowledge represented by

**Table 1: Comparative Analysis of Friendly Teacher Training Components Against Model Extraction and Fine-Pruning Attacks on CIFAR 10. Normal Teacher means standard training without any special properties.**

Dataset	Victim		Fine-Pruning		Extraction	
	Main Task Acc	Trigger Set Acc.	Main Task Acc	Trigger Set Acc.	Main Task Acc	Trigger Set Acc.
CIFAR 10						
Normal Teacher	93.15±0.08	100.00±0.00	90.51±0.19	62.33±1.53	94.09±0.05	54.67±3.51
+ FM Properties	93.43±0.08	100.00±0.00	90.33±0.11	82.67±1.53	94.20±0.04	79.33±3.06
+ OD Properties	93.43±0.08	100.00±0.00	90.33±0.11	82.67±1.53	94.54±0.09	89.00±1.00
+ SAM (Fully UAE Watermarking)	94.04±0.11	100.00±0.00	90.13±0.38	87.40±4.28	94.59±0.15	88.00±3.16
Spectral Normalization	94.06±0.06	100.00±0.00	91.29±0.17	72.33±2.52	94.65±0.07	91.33±0.58

UAE, rather than simply overfitting to the corresponding samples. Sharpness-aware minimization further strengthens the watermarks under worst-case parameter perturbations, thereby enhancing resistance to model modification. Lastly, while in the main text we solely applied heuristic optimization of Lipschitz continuity inspired by adversarial robustness, we subsequently experimented with spectral normalization [10], explicitly constraining the Lipschitz constant to 1. This slightly improved the watermark effectiveness on extraction surrogates, yet significantly reduced the resistance of watermarks in model modifications. Notable shifts in the decision boundaries following fine-tuning disrupted the strict constraint on the Lipschitz constant, leading to this degradation. Consequently, we conclude that heuristic optimization of Lipschitz continuity is sufficiently effective, and further enhancement is left as future work.

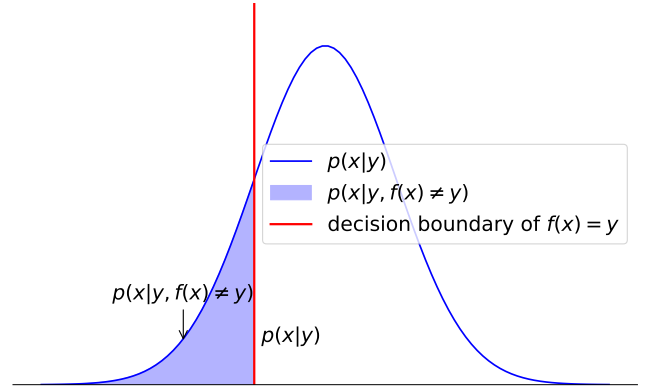
In the context of extraction attacks, adversaries are agnostic to the output distribution adjusting temperature,  $\Gamma_o$ , selected by the model owners. However, they may still employ the distillation temperature  $\Gamma$  in extraction. Thus, in Table 2, we explore the impact of the attacker's temperature,  $\Gamma$ , on the transferability of watermark behaviors. It is observed that a  $\Gamma$  greater than 1 even enhances the mimicry of the surrogate to the watermarked model, and thus facilitating the transmission of watermark behaviors. Conversely, a  $\Gamma$  less than 1 slightly reduces the watermark effectiveness on the extraction surrogate, but it presents risks of numerical instability and results in an overconfident surrogate model, which is seldom applied in distillation or extraction [13]. Therefore, the use of the attacker's temperature  $\Gamma$  in model extraction does not significantly affect the transfer of watermark behaviors. This also indicates the robustness of the choice of temperature  $\Gamma_o$  when optimizing output distribution properties.

**Table 2: Accuracy on Main Task and Trigger set with varying Extraction Temperature.**

Extraction Temperature	0.5	1	2	5	10
Main Task Accuracy (%)	94.46	94.59	94.50	94.32	94.42
Trigger Set Accuracy (%)	85.00	88.00	91.00	92.00	92.00

## 4 DISCUSSION ON THE APPLICABILITY OF UAE WATERMARKING

In this section, we discuss the practical applicability and sustained usability of our algorithm in the evolving landscape of deep learning.

**Figure 1: Probability density illustration of the UAEs synthesized by diffusion model.**

The primary concerns are encapsulated in two pivotal questions: First, as deep models continuously improve in terms of generalization performance and adversarial robustness, will the Unrestricted Adversarial Examples (UAEs) remain challenging enough to serve as a unique identifier? Second, generating UAEs necessitates the construction of task-specific diffusion models, which incurs significant overhead, can UAE watermarking remain practical? We provide an in-depth discussion of these two issues.

### 4.1 Sustainable Unique Identification Capabilities of the UAEs

The identifiability of adversarial examples as model watermarks does not conflict with the overarching goal of enhancing adversarial robustness. Theoretical findings assert that adversarial vulnerability is unlikely to be eradicated [4]. Under certain moderate conditions on data distribution, any classifier can be adversarially deceived with high probability when perturbations slightly exceed the natural noise level inherent in the problem [4]. Even robust models and biological vision systems are susceptible to adversarial example [6]. UAEs generated by diffusion models can be viewed as samples conforming to the underlying distribution while deviating from the tail boundary conditions fitted by the classifier [3], as illustrated in Figure 1. For any non-oracle classifier, such regions inevitably exist. Thus, the gap between adversarial accuracy and clean accuracy will persist, which is sufficient for intellectual property verification. Moreover, model thieves typically seek to

acquire knockoffs at lower costs and monetize through pirated API interfaces, while adversarial defenses often require substantial computational resources, degrade main task performance, and necessitate specialized pipeline designs, far exceeding the costs associated with direct model training. Consequently, the challenge presented by UAEs is non-trivial and suggests long-term scalability as a intellectual property identifier.

## 4.2 Cost of Constructing Diffusion Models to synthesize UAEs

Constructing diffusion models entails substantial training costs, which can sometimes exceed those associated with building main task models. Consequently, creating a diffusion model for each dataset poses significant computational challenges. However, we have discovered that fine-tuning pre-trained diffusion models allows for the straightforward construction of distribution priors on new datasets. Figure 2 illustrates the effects of fine-tuning a pre-trained diffusion model on CIFAR-100 after its initial training on CIFAR-10, comparing the synthesized results after one epoch and 1200 epochs. It is evident that brief fine-tuning can already yield viable results in some classes. Further, leveraging the world knowledge embedded in pretrained multimodal foundational models, such as GPT-4 combined with stable diffusion and latent diffusion, it is feasible to construct UAEs on any dataset [2, 8, 9]. Therefore, generating UAE watermark samples does not require building new diffusion models from scratch for each dataset; the exploration of specific pipelines is left for future work.



**Figure 2: Synthesized images from different finetuning epochs, (a) denotes epoch 1 and (b) denotes epoch 1200.**

## REFERENCES

- [1] Arpit Bansal, Ping-yeh Chiang, Michael J Curry, Rajiv Jain, Curtis Wigington, Varun Manjunatha, John P Dickerson, and Tom Goldstein. 2022. Certified neural network watermarks with randomized smoothing. In *International Conference on Machine Learning*. PMLR, 1450–1465.
- [2] Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. 2024. Content-based unrestricted adversarial attack. *Advances in Neural Information Processing Systems* 36 (2024).
- [3] Xuelong Dai, Kaisheng Liang, and Bin Xiao. 2023. Advdiff: Generating unrestricted adversarial examples using diffusion models. *arXiv preprint arXiv:2307.12499* (2023).
- [4] Elvis Dohmatob. 2019. Generalized no free lunch theorem for adversarial robustness. In *International Conference on Machine Learning*. PMLR, 1646–1654.
- [5] Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. 2020. Giga: Generated image quality assessment. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 369–385.
- [6] Chong Guo, Michael Lee, Guillaume Leclerc, Joel Dapello, Yug Rao, Aleksander Madry, and James Dicarolo. 2022. Adversarially trained neural representations are already as robust as biological neural representations. In *International Conference on Machine Learning*. PMLR, 8072–8081.
- [7] Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. 2020. Learning transferable adversarial examples via ghost networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 11458–11465.
- [8] Yueqian Lin, Jingyang Zhang, Yiran Chen, and Hai Li. 2024. SD-NAE: Generating Natural Adversarial Examples with Stable Diffusion. In *The Second Tiny Papers Track at ICLR 2024*. <https://openreview.net/forum?id=D87rimdkGd>
- [9] Jiang Liu, Chen Wei, Yuxiang Guo, Heng Yu, Alan Yuille, Soheil Feizi, Chun Pong Lau, and Rama Chellappa. 2023. Instruct2Attack: Language-Guided Semantic Adversarial Attacks. *arXiv preprint arXiv:2311.15551* (2023).
- [10] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral Normalization for Generative Adversarial Networks. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=B1QRgzIT->
- [11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [12] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.
- [13] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. 2021. Does knowledge distillation really work? *Advances in Neural Information Processing Systems* 34 (2021), 6906–6919.
- [14] Haotian Xue, Alexandre Araujo, Bin Hu, and Yongxin Chen. 2024. Diffusion-based adversarial sample generation for improved stealthiness and controllability. *Advances in Neural Information Processing Systems* 36 (2024).